A Machine Learning Approach to predict Young Voter Enthusiasm based on Non-Political Factors

by

Md. Nowroz Junaed Rahman
16101158
Md. Humaun Kabir Pantho
16101156
Nafis Fuad
16101267

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
April 2020

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Nafis Fuad

Nafis Fuad
16101267

Pantho

Md. Humaun Kabir Pantho
16101156

Md. Nowroz Junaed Rahman

Md. Nowroz Junaed Rahman
16101158

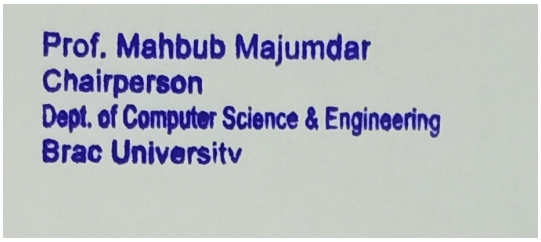
Approval

The thesis/project titled "A Machine Learning Approach to predict Young Voter Enthusiasm based on Non-Political Factors" submitted by

1. Md. Nowroz Junaed Rahman (16101158)
2. Md. Humaun Kabir Pantho (16101156)
3. Nafis Fuad (16101267)

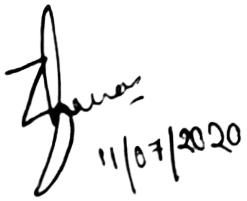
Examining Committee:

Supervisor:
(Member)



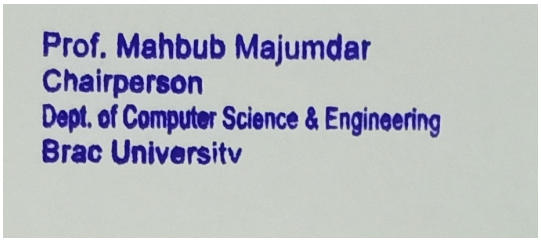
Mahbubul Alam Majumdar, PhD
Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Co-supervisor:
(Member)



Syed Zamil Hasan Shoumo
Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)



Mahbubul Alam Majumdar, PhD
Professor and Chairperso
Department of Computer Science and Engineering
Brac University

Abstract

The right to vote is considered to be the backbone of democracy. As we are entering the third decade of 21st century more and more countries around the world are adopting the democratic government system. One of most important element of democratic country is the power vested in the common people and one of the way the people are expected to exercise this power is to elect a qualified candidate to lead their country. The only way to make this election process effective is to make sure everybody participates in the process. The people who are eligible to participate in this election process to elect a candidate are called "voters". A substantial amount of these voters are young voter or voters who have newly been registered. It has been noticed that young voters in most of the countries are reluctant to participate in the voting process. There are many social, psychological and other non-political factors behind this reluctance. This research seeks to find those factors that motivates or repels a young voter to participate in the voting process. Besides finding the factors this research will also try to determine whether a young voter is likely to vote in an election or not based on those factors mentioned before. The data set of this research was prepared by surveying via Google Forms. Later the data set was analyzed to find out the reasons behind their participation. RFECV was used to select the optimum features and later Support Vector Machine, Random Forest, Extreme Gradient Boosting and Naive Bayes were used to predict voter participation based on the set of optimum features. In such an experimental setup Extreme Gradient Boosting and Support Vector Machine with a Gaussian kernel has shown more promising results than the other aforementioned models. The aim of this research is to predict whether a young voter will participate in the voting process or not and find the reasons behind it so, that maximum voter turnout can be ensured and perfect democracy can be achieved.

Keywords: Machine Learning; Voting; Young Voter Turnout;Random Forest; Support Vector Machine; XGBoost; Naive Bayes; Non-political;

Dedication

We would love to dedicate this research to our loving parents, families and friends.

Acknowledgement

First of all, Thanks to the Almighty, most Gracious, most merciful and the creator of everything who blessed us to pursue and complete our degree in BRAC University and provided us guidance to complete this research.

The people behind this work are Md. Nowroz Junaed Rahman, Md. Humaun Kabir Pantho and Nafis Fuad, students of the CSE department of BRAC University. This research is the outcome of these 4 years of education and the knowledge we gained.

Our journey toward this thesis would not have been possible without the help of many people. We would especially like to thank Dr. Mahbubul Alam Majumdar, PhD, our thesis supervisor, for his immense help and support through the course of this work. Without his continuous guidance this work would have been impossible to complete. We would also like to thank Syed Zamil Hasan Shoumo, the Co-supervisor of our thesis, for his expert advice and extraordinary support throughout this thesis process. We are truly grateful to Mr. Samiul Islam for being so inspiring and helpful through this journey. We are also grateful to Dr. Wasiqur Rahman Khan for helping us designing the survey which we used to collect data and build our data set. We are thankful to the lecturers and counselors of our university for their valuable comments and feedbacks while designing our survey and data collection. We would also like to thank everyone including our friends and faculty members, juniors and all the respondents who were involved and participated in the data collection process and completed our survey by taking time out of their busy schedule. This research has been possible because of their help.

Finally we would like to thank and express our gratefulness to our beloved parents and siblings. They are the reason of our existence. Their love and support are truly indescribable and without them this research wouldn't have been possible.

Table of Contents

List of Figures

List of Tables

Chapter 1

Overview

In this chapter we will give a short overview of the topic and we will address the problem statement, our objective and how we will solve it will be discussed in this chapter.

1.1 Introduction

The world is becoming more populated day by day and to lead them and for the prosperity of every nation a government in each nation and region is necessary. Democratic system is a type of government in which people elect a leader for them. This is considered the best form of government. In the words of Abraham Lincoln," Government of the people, by the people for the people, shall not perish from the earth" [46]. And the backbone of every democratic government is voting. It is the fundamental right of a democratic Society. According to Macmillan dictionary, voting is a process to formally express an opinion by choosing between two or more issues, people etc [50]. It is the privilege of being a citizen of a country or part of a nation that gives them the power to choose the representative for their government. Every government has a purpose to establish and flourish policies that will be beneficial to its citizens. So, voting is an important factor that will decide the growth of one's society. In most countries and regions the minimum age to vote to elect a leader for the government is 18. In every country the number of young adult voters precisely aged between 18 to 26 are pretty large in number. In the USA the percentage of this range of people is more than 13%. In other countries it is also huge in population. In Bangladesh the young voters are more in number than many other countries which is more than 19% [45] so their participation can have a significant impact on the election process. So, these young voters have a big impact on the democracy and electing their leader. Lately the young voters are not participating in voting and are not showing much interest in this sort of activity. There is almost no existence of any research regarding finding out the non-political factor behind the enthusiasm on young voters at least in Bangladesh.

1.2 Problem statement

Even though most of the world follows democratic government system, the amount of voter turnout in an election is not what we might expect. In 2012, the voter turnout in the US Presidential election was 55 percent of the total voters and in 2014, only 36 percent of total voters voted in the midterm election [22]. This was the election situation in the USA but in most of the countries we see similar type of situation. Enough people do not show interest and do not attend the election to choose amongst the candidates. Even from the amount of voters who attend the election the number of young and newly registered voters are very low. There can be several reasons behind this and because, of the lack of voter participation and not voting several things can happen. First of all, if enough people do not vote then it is difficult for a qualified candidate to win. In an election voters often get to choose between candidates representing different parties. Having fewer overall voters means all votes must align in order for any given candidate to win. If a person do not vote, it means he or she has to rely on other people to go to the polls and vote to elect a qualified candidate. That is not exactly a good way to ensure a win for the qualified candidate because, while someone has relatively easy access to vote, others may face obstacles like lack of transportation or mobility. Also if a person does not vote he gives up his constitutional right. It is a constitutional right just like freedom of speech in countries like the USA and most of the democratic countries which makes their voice heard. A single vote can change the world for the better and the course of a country if people choose the right candidate. Finally, the voting system or the election system faces many problems and sometimes the selection of the winner from the candidates is not properly justified. So, by not giving vote not only one gives away a constitutional right but also it can cause a major change to the election result.

We mainly focused on the students of BRAC University who became newly eligible to vote in recent years. While conducting the research we found out that most of students who are eligible to vote and new voters are not properly aware of the voting system of our country. Also while collecting our survey from them many students did not show any interest and was not serious when completing the survey. Many of them did not took enough time to fill up the survey and that is why we got a lot of noisy data in our data set.Another thing we found out is that even though our survey was completely anonymous and was based on non-political factors, many of the students did not feel safe completing the survey and asked about the anonymity of the survey before filling it out even though we informed them clearly multiple times before the survey started. It is also true that the real reason for someone to vote or not is very difficult to know. This may happen because, students are sometimes not comfortable sharing their thoughts or they may feel uncomfortable in answering some questions while filling out the survey. So, to find out the reasons why new voters want to vote or not there is no effective platform or strategy. We choose to survey online because it was easy to reach the people and it was easy for people to fill out the information on the survey.

The basic formula for verifying if someone will cast vote or not, assuming that people will act completely rationally is as follows:

$$PB + D > C$$

Where,

- P is the probability of an election to be affected by an individual's vote
- B is the perceived benefit that an individual will receive if his/her favoured candidate/ political party were elected
- D stands for Democracy or civic duty but nowadays it represents that by voting the social or personal gratification an individual gets
- C is the time, effort, financial cost involved in voting

P and PB is virtually zero and D is the most important element to motivate people in voting and for someone to vote or to show interest in participation the factors should exceed C [2].

1.3 Aim of Study

The main aim of this research is to find out the non-political factors that influences a young voter to participate in the voting process or not and use those factors to predict whether someone will participate in the voting process or not. There were no pre-established data set so, we had to resort to surveying to collect data. The questionnaires are divided into several parts. Firstly, we focused on their usage of news resources such as: newspaper, social media etc. Then we asked them about their interest about next elections and their willingness to vote in it. We also gathered information about their educational background, family's educational background, financial health and their general knowledge on politics and voting system. Based on these queries we wanted to find out the root reasons behind their interest and their probability to give vote in the future elections. Also, if they don't want to vote in the future election in that case we also tried to find out the reasons behind that too. We have used 5 algorithms to build our system which would predict whether a person is going to give vote or not and find out the reasons behind their decision. The algorithms are Random Forest, Extreme Gradient Boosting, Linear SVM, Kernel SVM and Naive Bayes.

1.4 Research Methodology

This research has been conducted on a span of 14 months and 15 days. However we are still working on this topic and we intend to expand our work in near future to explore more possibilities and try to refine the results we have got so far. Our work started on approximately 15th of January, 2019. We spent better half of January trying to pin-point the topic. It was not until middle of February we were satisfied with our topic. After that we started looking for articles and past works, research papers on anything that is related to our work. It took another month and we are already in March of 2019. We did not find enough previous work but, we kept on searching. Back then the nature of our work was somewhat

different. We intended to scrap popular social medias like: facebook, twitter etc. for status given by young voters about their reflection regarding voting and then apply sentiment analysis on them. Due to some difficulties we had abandon that approach. During April we decided to opt for survey rather than scrapping data from social media. We contacted several faculties from ESS department of BRAC University to consult about how we should set questions for our survey. We also conducted some psychological counselors from counseling service of BRAC University about how to make the survey more efficient and capture the psychological factors correctly in the survey. Then we took April 15th to May 15th to construct the survey and test them among the people we knew before releasing it to the actual target audience. During this time we made some minor changes as per the feedback we got. We started doing the survey from May 16th. After collecting almost 557 records which took till June 10th we realized there were some major issues in our survey. We resolved those issues but we had to start from scratch. We deployed the survey on July 5th and collected data until October 15th. After that we started to pre-process the data so, we could apply various learning algorithms. It took us until December 10th to clean and pre-process the data as it had a lot of noise in it. We tried our best to remove noise as much as possible and create reduced feature for each algorithm so, we can get maximum accuracy. We spent rest of the December and better half of January experimenting with different algorithms and feature set combination to find out the best algorithms to use. Then we used rest of January and February to tune those algorithms and apply them and record the result. We spent the month March, 2020 to write the paper and review it for mistakes. It took us from March 24th, 2020 to March 31th, 2020 to convert our work into latex format as per the university's requirement. Even though we have finished writing the paper that does not mean our work here is done. It is far from done. We faced numerous difficulties in our work and some them we could solve or avoid but, some of them we could not so, we have plans to keep on working on these and due to lack of scope and man power we could not achieve some of the goals we had on our mind which we will try to achieve in our future work. Those will be discussed later on the future work section. Figure 1.1 shows the timeline of the survey in a gantt chart.

1.5 Thesis Outline

- Chapter 2 focuses on background study about voting, some important statistics about voting, previous work related to our research and background study on the algorithms we have used in our work
- Chapter 3 discusses data set collection, description of the survey and the data set, idea behind the reasons of voter participation and the how the data set was pre-processed
- Chapter 4 talks about the process to find the features behind voter enthusiasm and the process to predict voter participation, algorithms used to achieve in the process and results of these algorithms based on different metrics
- Chapter 5 contains conclusion, findings, limitations we faced and our future plan

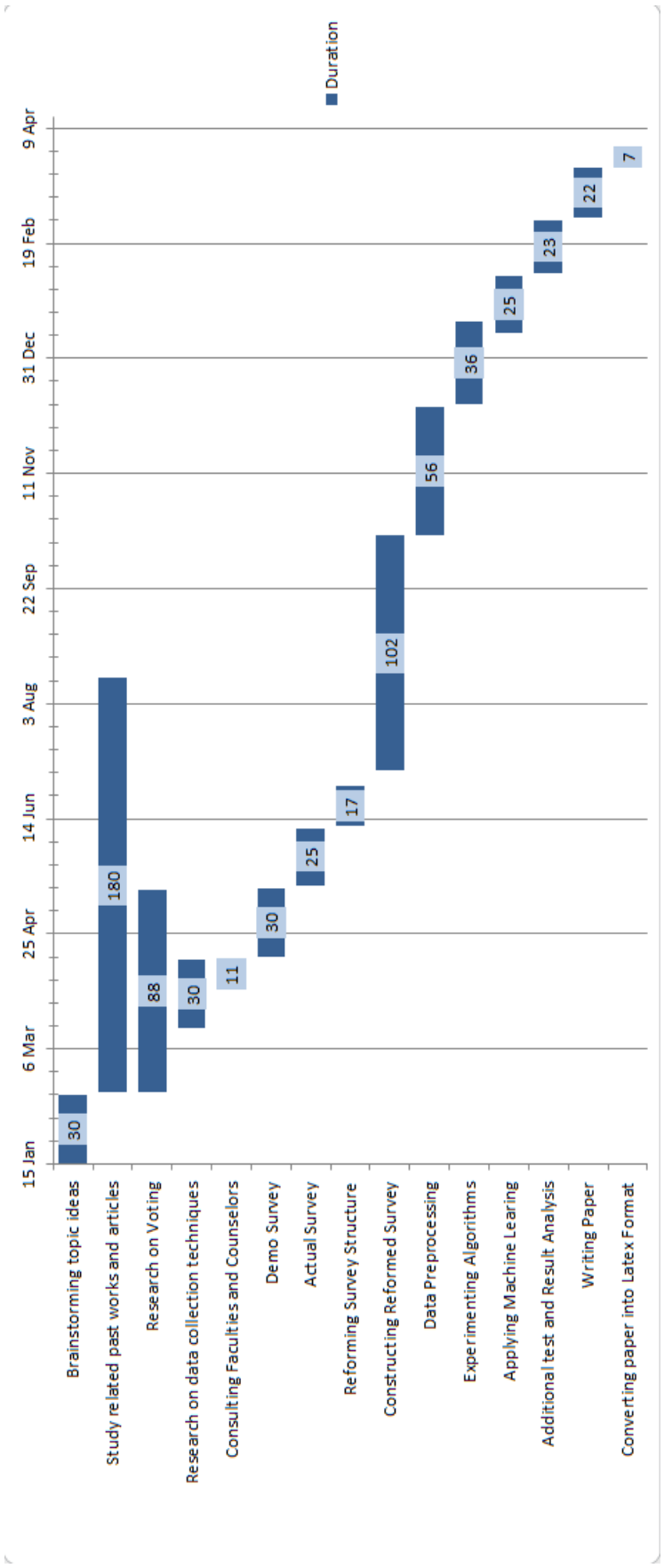


Figure 1.1: Research timeline

Chapter 2

Background Study

This chapter is going to further describe the problem with the overall statistics of voting in the world and Bangladesh, and some related works on voting and machine learning.

2.1 Electoral systems around the world

Voting is a simple term but the importance of voting in a democratic country and the effect of it is without any doubt very important. Voting means a formal expression of opinion or choice, either positive or negative, made by an individual or a body of individuals[52]. In electoral term it indicates the activity where people vote or give opinion by choosing someone or something for the government [53]. In most democratic countries, the minimum age to give vote is 18. Also the person has to be a citizen of that country and maintain certain protocols to give vote. There are different forms of government or type of administration prevalent around the world. Most of them support voting either by public or by the officials. The major government systems and leadership styles in the world right now are autocracy and democracy. Autocracy which is also known as authoritarian leadership is basically a system of government in which a single person or party process the power of the whole country or region. The government system characterized by individual control over all the decision of a country and little input from group members. In this government system the people of the country don't get to choose the leader or they don't get to talk about or give their opinion on the matters of their country. To sum up, the people do not exercise the freedom of speech. Autocratic leadership involves absolute, authoritarian control over a group. Dictatorship is not exactly as autocracy because, dictatorship is a form of autocracy where authority is wielded by either one parson or a group of people but in autocracy the power is not wielded. On the other hand the system where the choice of people or the voting of people matters is the Democratic system. Democracy is the form of a democratic system. It is a form of government in which the power is vested in all the people eligible for the power or in the representatives elected by the people. To elect the representative a voting process usually takes place and by voting the people actually forms a government. So, it is a huge responsibility to vote and not giving vote may sometimes lead to an unexpected and unwanted situation.

In 1977, only 35 out of 143 countries (24%) were qualified as democracies, while 89

of them (62%) were mentioned as autocratic government country. But at the end of 2017 there were 167 countries that has at least 500,000 in population and 96 out of them (57%) were following some of the aspects of democracy or had fully democratic government, where only 21 (13%) were autocracies. Other 46 (28%) had the element of both democratic and autocratic elements [32][12]. So, over the years the amount of democratic country increased and the election system became an important part almost in the whole world.

The Economist Intelligence Unit (EIU) is a UK- based company that compiled an index of the countries and overall statistics currently running government systems which is titled "The Democracy Index". The purpose of the index is to measure the state of democracy in 167 countries. It was first made in 2006 but by 2019 it has been updated many times with latest statistics. From those 167 countries, 113 countries are run by full or flawed Democracy. And 54 of them are maintaining Autocracy or Full Dictatorship [40]. Most of the major countries in the world supports democracy and holds a full-fledged election to elect their governing body in which votes are given by the people of the country. Some of the example of these countries are: Canada, Australia, United State of America etc. The population of Canada in 2016 was 35.15 million [31] and in 2015, in the general election of Canada, the voter turnout or the number of people gave vote was 66% of total eligible voters[51]. The young voter turnout specially voters between 18 to 24 was 57% that means almost half of the total young voters did not attend the election [51]. Similarly in USA the total estimated population was 3.24 million in 2016 [24]. Total voter turnout of 2016 presidential election was 61.4% [27]. The people aged between 18-29 who are considered as young voter in USA were 46.1% to be included into voter turnout in the same election [25]. Here we can see that, more than half of the young voters did not participated in the biggest election of USA. So, by looking at these statistics it can be said that the participation of voters especially young voters in biggest democratic countries are not that good and it affects the election systems overall.

2.2 Electoral system in Bangladesh

Our main focus is new voters especially who are aged between 18-26 years old. Usually they are students of Universities or other equivalent institutions so, our priority was to perform the survey on these university students. But, due to lack of manpower and unavailability of resources we could not perform the survey outside of BRAC university but we have plans to expand our work in the future.

Like most other democratic countries in Bangladesh people need to have some qualifications to be able to vote in the elections. By constitution a person can participate in an election of the parliament if he/she is a citizen of Bangladesh, is not less than eighteen years old of age, does not stand declared by a court to be of unsound mind, considered by law to be a resident of the constituency and has not been convicted of any offence under the Bangladesh Collaborators (Special Tribunal) Order, 1972 [42]. According to the 2011 census of Bangladesh the total population were 144.04 million where the young people who are aged specifically between 15-24 years were 26.16 million which was 18.16% of total population in Bangladesh and almost 70% of those young aged people were eligible as voters [49][17]. In July, 2018 the esti-

mated population were 159.45 million and among them the number of young age people were 19.14% [43], The increase of young age group ratio can be seen in those 8 years. The recent parliamentary election of Bangladesh was held in 2019 and like other countries the voter turnout was not good enough. Even in 2020, City based elections like in Dhaka North and Dhaka South city corporation elections the voter turnout was very poor in number. The Chief Election Commissioner of Bangladesh, KM Nurul Huda estimated that it was not over 30% [41] and the number of young voters was even less. The major reason for this is basically people's reluctance to participate and due to misunderstandings. There are more reasons such as family background, parents' representation of voting, parent's educational qualification, participant's knowledge to current affairs etc. which will be discussed later.

2.3 Literature Review

When we started working on this research our goal was somewhat different. We intended to perform sentiment analysis on the statuses of young voters regarding their thoughts and outlook on elections to ultimately determine whether they will participate in the voting or not but later on due to unavoidable circumstances we had to take on a different approach. Later on we collected data through our personalized survey and used machine learning algorithm on them to find out whether someone will participate in the voting process or not. We noticed that there is almost nonexistent research related to the approach we took. Most of the research was related to sentiment analysis. That is why we had we studied those research to gather ideas and counter any issues or problems we faced. Some of the researches that helped us are as follows:
Some of the works related to sentiment have helped us in our research. Sentiment analysis has become one of the most explored areas of research in Natural language Processing. In the very first paper [26] opinion lexicon was used to label the words as positive and negative. Then it was fed into an instance of Naive Bayes algorithm which used condition probability of label and features (the features are the words present in the tweets). The paper mentioned above aimed to find out the correlation between the sentiment of the tweets and the polling data of Donald Trump and Hillary Clinton's presidential campaign.

As for the next paper, which aimed to predict the Indian Election result of 2016 via sentiment analysis of "Hindu Twitter" [23]. In this work the author's much like the previous work collection tweets over a span of a month and performed data mining on them. They collected almost 452,235 tweets. The tweets were basically about what people's thoughts regarding 5 major parties of India. They applied Naive Bayes, Dictionary based and SVM on these tweets to determine who would win the election. SVM predicted that with a 78.4% probability "BJP" would win and reality that is what happened.

This work primarily focused on extracting multi-label sentiment from Bangla and Romanized Bangla comments from "YouTube" videos. This study proposed a three label sentiment analysis and a five lable emotion classification. Additionally, the next paper [30] used word2vectorizer to vector representation of the text. Continuous Bag of Word (CBOG) and Skip Gram (SG) are also implemented to find frequency of

words in the text and probability of context for those word. The LSTM algorithm is used to classify and making prediction of time series data and also CNN is used here. On the other hand, this paper also mentioned baseline method SVM like previous paper.

In this work the authors tried to find correlation between events and sentiment related to that event using Tweets related to that event [20]. FIFA World Cup 2014 was used as a case study in this work. Bayesian Regression classification method was used in this research. Along with Bayesian Regression Classification Lexicon was used to identify between objective and subjective tweets. Also TF-IDF was used to filter feature variables. Twitter's official API was used to scrap tweets and hastags were used to classify tweets. Stemming was performed to pre-process data. WEKA was used as a machine learning platform in this work.

In this work the authors tried to extract sentiments about entities from Newspaper and blog articles [9]. They used lexicon in their work. They recursively queried for synonyms in WordNet. They found 4 separate ways to go from good to bad in their work.

In another paper [34] focused Twitter as a data source to collect data about political events based on user tweets. The main purpose of this research was observing that when a candidate advanced in the election based on the amount of reaction he/she got from the people on twitter. They used features such as: number of replies, retweets and likes. Their target variable of label was reaction which was divided into three levels such as: high, medium and low. This is somewhat related to our work because, although our work is binary classification and their work is multi-class classification but, the process is somewhat same. They used Naive Bayes classifier to predict the output variable.

As mentioned before our research has changed a lot from what it was originally at first. We originally wanted to scrap popular social medias like: Facebook and twitter to collect statuses and then perform sentiment analysis on them. Due to some unavoidable circumstances we had change our strategy and pursue different tactic. We later decided that we would used personalized survey to collect data and build a classification model. We also noticed that most of the work done related to voter turnout are related to sentiment analysis and there almost non-existent work related to how we are trying to pursue the problem that is why we had to refer to the works related to sentiment analysis to design our research and solve any problem we faced.

2.4 Proposed Model

In our proposed model, the first step is to collect the data from survey forms. We had to conduct survey as there was no pre-existing data set. Due to the major data privacy concern at that time all major organization like Facebook, Twitter, Instagram put restrictions on their api for collecting data.

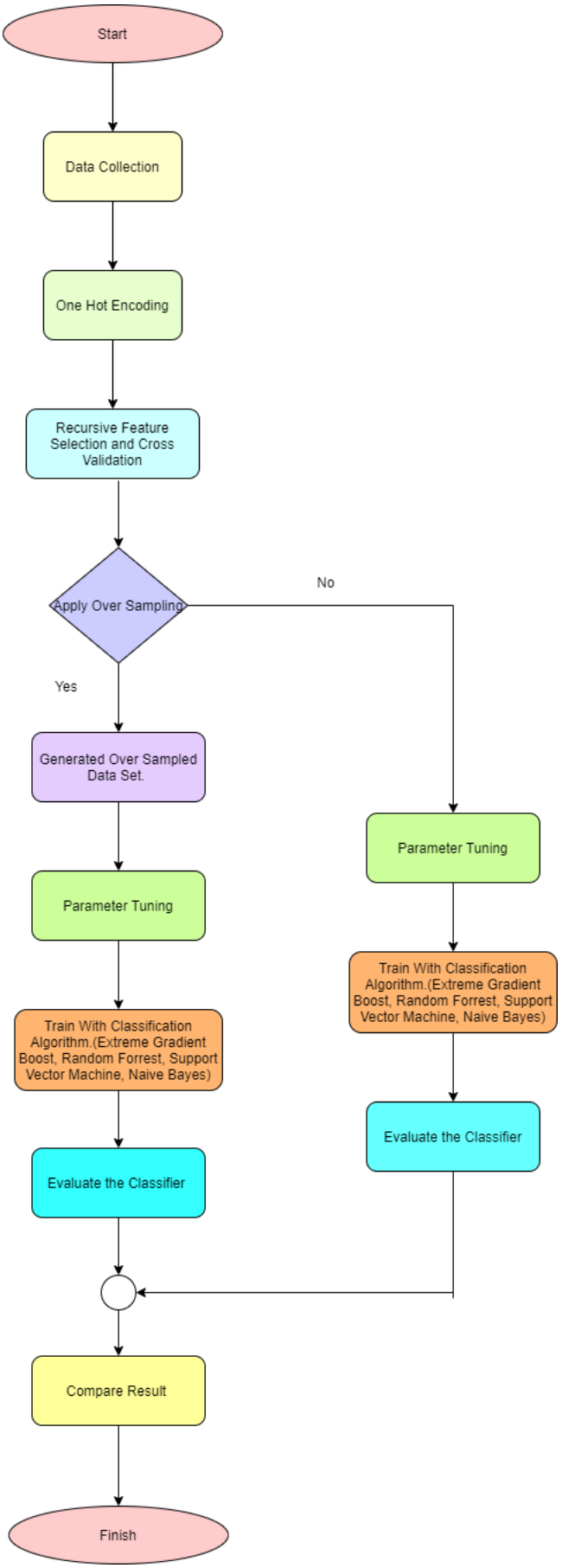


Figure 2.1: Illustration of our proposed model

Secondly, we applied one hot encoding. As our data set have all categorical data, we can perform either label encoder or one hot encoding. In label encoding, every category will be given a number. On the contrary, one hot encoding will create new column for each category and put 1's and 0's. One means that category exist and zero means does not exist. However, for label encoder as we assign number's and then feed that data to machine learning, it assumes order in those category even if there is not. As a result we have only used one hot encoding.

Additionally, recursive feature selection and cross validation technique was used to find out the optimal features. The feature which are mostly related to the students decision of voting or not voting is extracted using REFCV. The whole process of feature extraction will be described later in this chapter and chapter 4.

On top of that, we have over-sampled the data set using Synthetic Minority Over-sampling Technique. The data set we collected was imbalanced.There was a presence of a minority class. There were very few students without willingness for voting. So, we over simplify this data points using SMOTE. Which will be described in details in chapter 3 and chapter 4.

Moreover we have used Grid search algorithm to fine tune our model to get the optimal predictor. Grid search is a popular way of tuning hyper-parameter in data science world. Parameter tuning is more elaborately described for our model later in this report.

After parameter tuning we have the model or the predictor which can predict if any individual will vote or not given a data point. The machine learning algorithm we used are Extreme Gradient Boost, Support Vector Machine, Random Forrest and Naive Bayes.

lastly, we compare the result of different algorithm with the model before over sampling and under sampling using accuracy, f1 score, precision recall and AUC values.

2.5 Classifier

A classifier is a predictor or a hypothesis that can predict the class of a given data points. It can work on both structured and unstructured data. In our work, we used classifier to predict if any young voter would vote or not. This is a binary classification problem. The classes are often referred to as target, label or categories. Machine learning algorithm are of two types Supervised and Unsupervised. We are using supervised learning technique to train classifier or predictor. In supervised learning the training data contains labels or target class. In our data set the data points have labels which are: whether a someone will vote in the election or not. We used this data to train different algorithms like Random Forest, Extreme Gradient Boosting, Support Vector Machine and Naive Bayes. A brief description about those algorithms are given below and their performance measurement is discussed later in this report.

2.5.1 Random Forest

Random Forest is an extended version of decision tree algorithm. Decision trees are great for only the data set that was used to build it. So, decision tree is not flexible and accurate. Decision trees also tend to over fit. To overcome this problem random forest was introduced [5]. Random forest uses the power of crowd or collective effort. Which is a simple but a very powerful technique. Random forest creates multiple decision trees according to the data. Here is how it works, firstly it makes a bootstrap data set in random. Which means it takes data points in a data set randomly to create a new data set which is known as bootstrap data set. Then it creates decision trees for each bootstrap data sets. For prediction random forest feeds the given input and performs voting on all the outputs of the trees. The output or classification that have the highest vote will be the prediction of random forest. The working process of Random Forest is shown in figure 2.2.

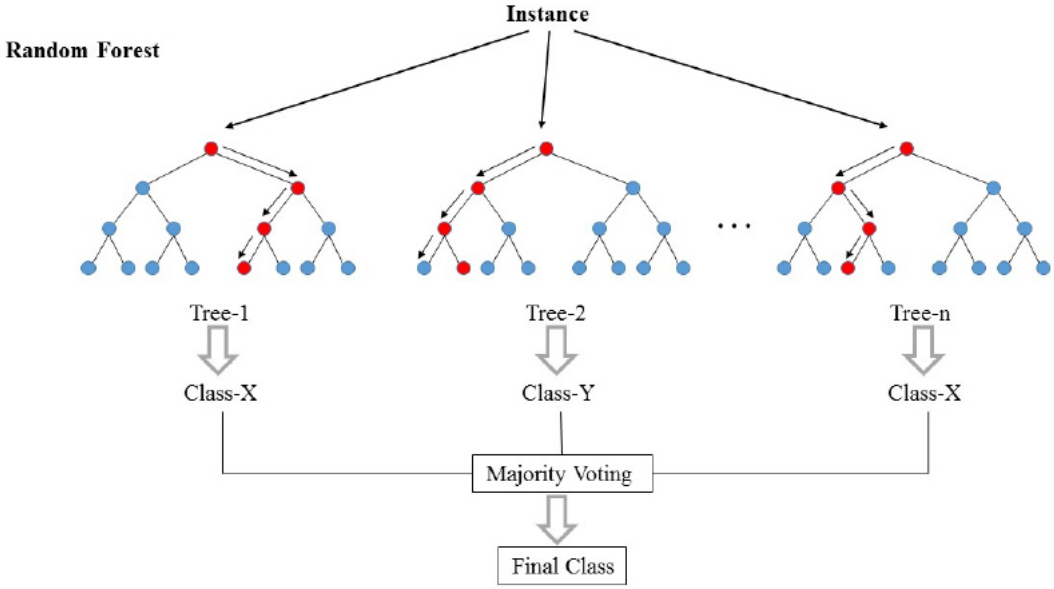


Figure 2.2: Illustration of Random Forest Algorithm

2.5.2 Extreme Gradient Boosting

Extreme Gradient Boosting is a boosting based decision tree inspired algorithm. Boosting is one of the most popular technique in the data science world that can turn a week learner in to a strong learner. Extreme Gradient boosting also known as XGBoost was a research project at University of Washington. The paper was presented by Tianqi Chen and Carlos Guestrin in a conference in 2016 [5]. Right after introduction of it XGBoost has become the recipe for winning several completion. Most of the winner teams in kaggle used XGBoost. XGBoost use the gradient boosting algorithm as a building block and use different system optimization and algorithm optimization technique to boost it's performance. XGBoost improves performance by using parallelization and hardware optimization. This algorithm takes the nested loop and interchange the loop sequence and creates parallel threads to execute. XGBoost also takes in to account the availability of space in ram and cache memory. On top of that, XGBoost evaluates the space on the memory while

loading large Data frame. XGBoost uses weighted Quantile Sketch algorithm and built in cross validation technique to increase the efficiency of the model and validate the data set. The working process of Extreme Gradient Boosting is shown in figure 2.3

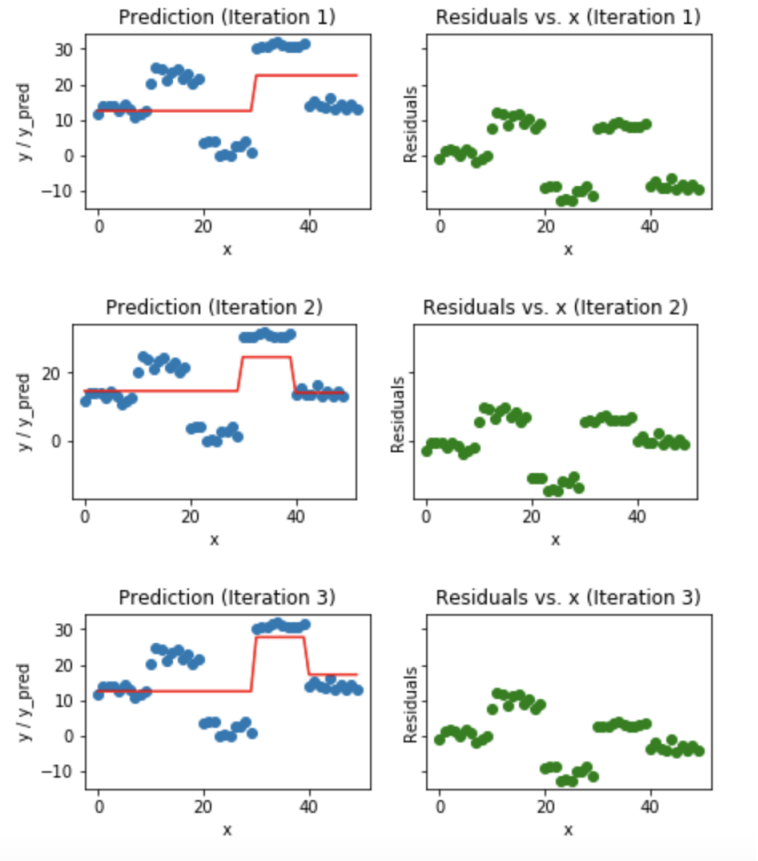


Figure 2.3: Extreme Gradient Boosting method

2.5.3 Support Vector Machine

Support Vector Machine or SVM is on of the most powerful algorithm used in classification problem. SVM creates a separation line between the data points to predict an outcome. It first plots all the data points in the data set. So, the graph will be n minus one dimensional. Here n is the number of feature in a data set. Then in this hyper-plane SVM will draw a constrain or boundary and separate different classes and then it would classify each class. For example, in our work one side of the boundary will represent that the young voters willing to vote. Whereas, other side will represent who are not willing not vote. This is how simple linear SVM works [6]. On the contrary, if the data points are placed in a way that they cannot be linearly separable then we will need soft margin or kernel SVM. Soft margin will try to separate and keep some misleading points inside the margin. It will calculate to minimize the misleading points. Soft margin would somewhat solve the problem but is not accurate enough in nature. Kernel SVM makes brings some new feature to the table alongside the existing features and some transformation technique. This transformation technique is Radial Basis Function (RBF). In this function gamma

value is the controlling parameter for any new feature. The greater the gamma value the more twisted the boundary will be. In our work we have used this gamma parameter to fine tune our model. Here in figure 2.4 we can see the working process of Liner SVM and in figure 2.5 we can see the working process of kernel SVM.

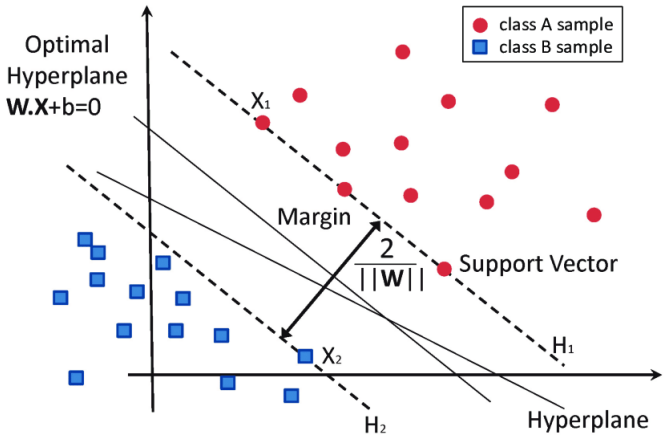


Figure 2.4: Illustration of Linear SVM Algorithm

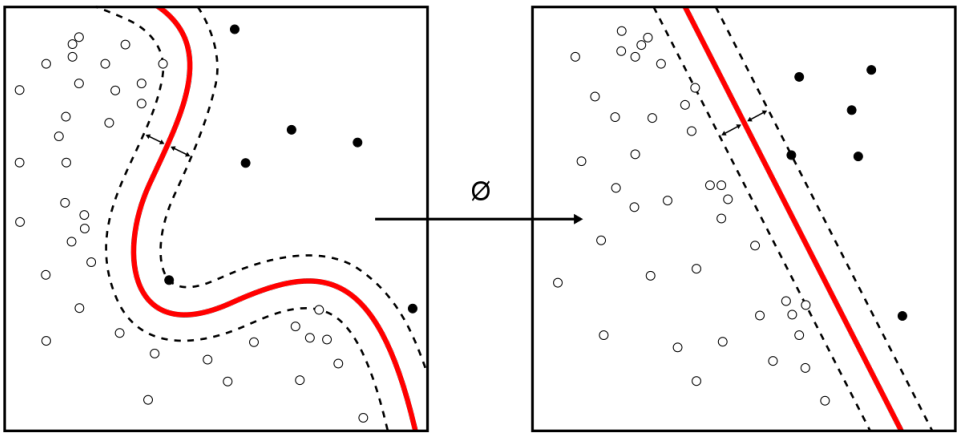


Figure 2.5: Illustration of Kernel SVM Algorithm

2.5.4 Naive Bayes

Naive Bayes is a machine learning algorithm based on probabilistic model. Naive Bayes is based on and named after the Bayes theorem. Naive Bayes is considered less complex and faster algorithm. This algorithm works reliably, most accurately and quickly for a very lengthy data set [14]. Naive Bayes assumes something that other algorithm do not, which is a feature in a data set does not influence other features in data set. For categorical data points as our work Naive Bayes calculate finds out the probability of something happening or not happening given all the attribute of the categorical feature. For example, the outcome or the label of our data set is willing to vote and or not. Additionally, one of the feature of our data set is if any participant's family lives in their own house or rented house. So, this algorithm will find the probability of participant's willing to vote given whether the participant owns the house, probability of not willing to vote given that the participant owns the

house, probability of willing to vote given that the participant rented the house and probability of not willing to vote given that the participant rented the house. It will follow the same step for all other 22 features and store the probability score. When we give any new data point which is not labeled Naive Bayes algorithm can give prediction by using those probability score in the Bayes Theorem. Bayes theorem for our predictor is given in equation 2.1. One of the limitation of Naive Bayes is that due to how it works, it fails to capture any complex underlying relations between the features.

$$P(\textbf{Voting}|\textbf{Data}) = \frac{P(\textbf{Voting})P(\textbf{Data}|\textbf{Voting})}{P(\textbf{Data})} \tag{2.1}$$

2.6 Parameter Tuning

We used scikit-learn's GridSearchCV library to tune the hyper-parameters of the classifiers we used [15].

2.6.1 GridSearchCV

Grid search algorithm is used to tune hyper-parameters for machine learning algorithms [44]. Hyper parameter are the parameter that are external to the machine learning algorithm which can be tweaked according to personal need. Parameters are the internal values of algorithm. Parameters shape the algorithm to increase efficiency and accuracy of the algorithm. For example, SVM has parameters called C and gamma to tune and increase it's efficiency. Grid search takes a set of value as grid and then applies all possible combination of values in grid to determine which combination gives the best performance then it returns the best combination of parameter values. It increases an algorithms ability to generalize on a data set.

2.7 Feature Selection

We used scikit-learn's RFECV library to perform feature selection and find reduced feature set [15].

2.7.1 Recursive Feature Elimination with Cross Validation

Recursive Feature Elimination with Cross Validation or REFCV is another advanced technique to find out the best sets of features for a machine learning algorithm. REFCV brings out the feature that has the highest impact on the data set. Firstly, REFCV takes the data set and construct models recursively. After that it determines the best and the worst features. It then discard these feature and follows the previous steps. In this way REFCV algorithm finds out the features that affects the prediction the most and the feature that affects the prediction the least. It also uses cross validation technique to evaluate the classifier. There are a lot of technique to validate but for our works we used K fold cross validation technique. K fold technique divides the data set in to specified sections then it uses one of the sections for testing and rest of the sections for training. Again, it repeat the same process with this time

different section for testing and training. Then average of all errors across all k fold is taken. Usually 5 or 10 is considered as value for k. For our case we took 5 as k.

Chapter 3

Data collection and Pre-processing

3.1 Recruitment and Procedure

The survey was approved and verified by the Department of Computer Science and Engineering, Counselling unit and Department of Economics and Social Science. We designed the survey with the help of Samiul Islam sir of CSE department and Dr. Wasiqur Rahman Khan sir of ESS department. The main objective of the survey was to find out the factors behind voter participation and also keeping the participants information anonymous. The answer of the participant in the survey was completely anonymous and cannot be backtracked to them. Additionally, during all the procedure the identity and privacy of the participants was protected. This data set will only be used for this thesis and will be handed over to the department of Computer Science and Engineering for any future research. Data set will not be under any circumstances distributed online or to any third party outside of university and the Department. Our survey questionnaires were revised and improved in a few versions. The versions are described bellow.

Version 1

In version 1, we used printed version of our survey forms to some selected participant to taste the feasibility of the survey and perform a test run. They filled up the form and gave their feedback, comments and complaints. We showed the feedback to our thesis supervisor and co-supervisor and modified the questions and formats according to the suggestions.

Version 2

In version 2, we used a google form rather than a printed form. Additionally, this time the participants were the first and second year students of Computer Science and Engineering department. Moreover, we gave a brief instruction about the survey before the participant filled out the form.

Version 3

In version 3, we changed the question that recorded our label or target variable. Previously there were some issue with this question and it resulted in a very bad

distribution of target variable. We added another question which recorded the probability of the participants voting in the next election on a scale of 1 to 5 to elimination possible bad distribution of data. Now the participant could easily expresses their willingness to vote in the next election.

Version 4

In version 4, we added two new questions. Firstly, now they can mention any non-political factor that may influence their participation in their own opinion. Additionally, in the another newly added question they could write any comments or their thoughts regarding voting in general.

3.2 Data set Description

There was no pre-existing data set for our research. That is why we had to collect the data and create out own data set. At first, we wanted to take data from Facebook and Twitter using web scrapper. Although, we faced some difficulties on this front. During that time all the major social network site like Facebook, Twitter or Instagram restricted the access of data collection due to increase concern of data privacy. As a result, we decided to make a survey form and collect data set on our own.

3.2.1 Determining the factors that Stimulates Voter Participation Visualizing the Data

Subscription of Newspaper and It's impact on Voting Participation

One of the questions that was asked to the participants was if they have any newspaper subscription and if they read it regularly. Newspaper is an authentic and reliable source of information. However, we have seen in recent times that there is a surge of misinformation being spread and targeted political advertisements. Newspapers play an important role of fighting back misinformation. Additionally, newspaper subscription also indicates socio-economic situation and educational qualification of any individual. According to "The Age and the effects of news media attention and social media use on political interest and participation" journal by Holt et al[18] showed that the young people are seeking political information on the social media and online version of newspapers. However, older people like to get informed by reading physical version of actual newspaper in their old fashioned way. Additionally, this papers confirmed and showed the fact that newspaper and other media influences the participation of political activity. On top of that, newspapers also determine the educational qualification of our survey group. The wealthy families in our country tend to subscribe to English newspaper. In contrast, the people who are not financially strong does not have access to newspaper subscription. The figure 3.1 illustrate that among the people who do not have any newspaper subscription, 127 will not vote for the next election.On the other hand 139 of them want to vote and will vote in the next election. Similarly, among the participants who have newspaper subscription but do not read it daily 261 have willingness to vote and 156 do not want to vote. Lastly, among the young people who have newspaper subscription

and read them daily 79 of them want to vote and 37 of them do not want to vote. In conclusion, according to our survey very few people read newspaper daily and among them voting participation is the lowest.

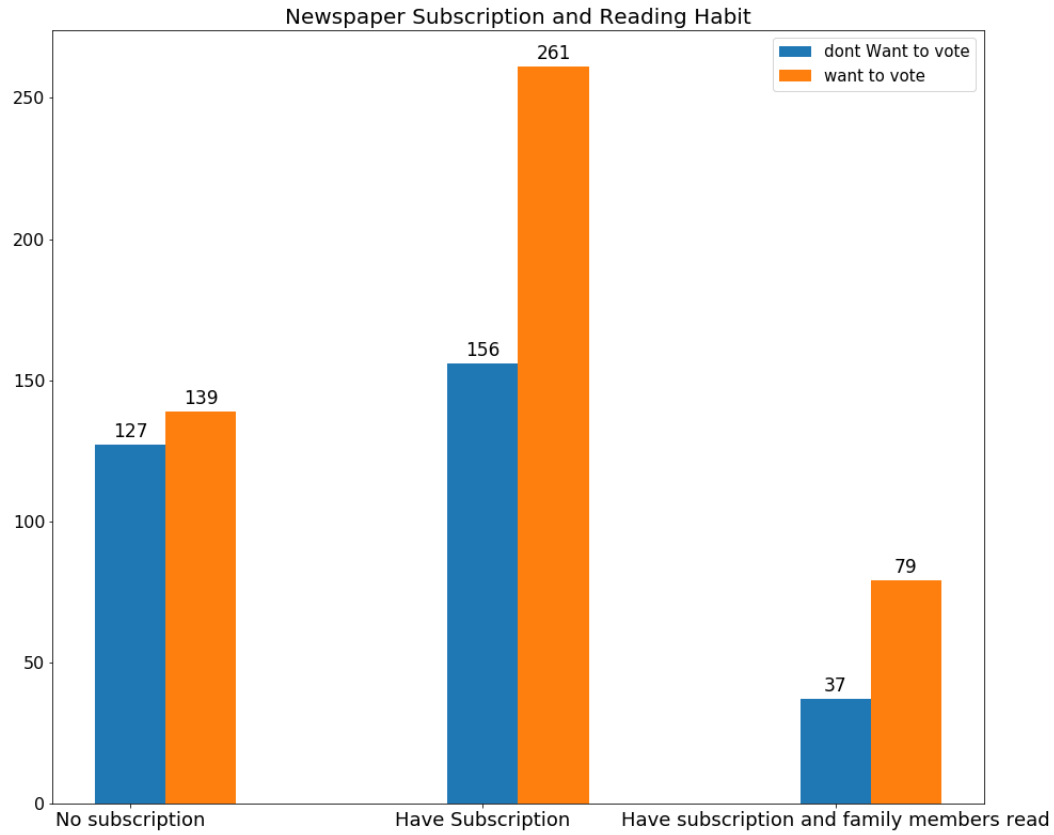


Figure 3.1: Illustration of Newspaper Subscription and Reading Habit

Ramification of Family Political Involvement on Voting Participation

The knowledge of politics among the parents gradually affect the political perspectives of their children. The discussion between parents about government party or the opposite party describes a lot about the perspective of the parents. Children tend to try to follow their parents in every aspect of their life. As a result they would generally try to follow their parent's political perspective. There is also a chance of conflict between the perspectives of two generations. A child might get his own perspective of politics where in it's subconscious mind it has the political preference of their parents. It's not unusual that a family has been supporting a political party for years from generations to generations. Torney Purta and Andolina[11] suggested that they develop higher levels of political knowledge, show greater intention to vote in the future, and do better on a range of civic outcomes from petitioning and boycotting to raising money for charities and participating in community meetings. Additionally, The New York Times stated that lifelong voting habits are formed in childhood and adolescences and that those issues of routine and habit may be important in determining voter behavior and therefore election results. The young voters usually do not hold a university degree at the age when they become voters. As a result, the political environment of their education and

also the political environment of their home put them on a conflict about which political party they should support or not. The young voters receive slightly different and advanced education that their parents had received. Due to the development of technology and communication the political perspective of parents might not be the same as their kids are used to due to the modern technology and communication system. The young voters have clearer information and knowledge than their parents.According to the survey we conducted, 124 participants have at least had one family member who is involved in politics and did not want to vote. Whereas, 209 participants also have a political person in their family and they are willing to vote in the next election. Secondly, there are 196 participants who does not have any family member involved with political party and not willing to vote. Although, there are 270 participants who also do not have any political history in family and wiling to vote in the next election. As for influence of participant's voter participation due to presence of political family member 28 said that presence of political member in their family influences their voting but they are willing to vote in the next election and 76 said that presence of political member in their family influences their voting and they would vote in the next election. On the other hand 292 admitted that presence of political member in their family does not influences voting and they are not willing to vote in the next election and 403 admitted that presence of political member in their family does not influences their voting and they are willing to vote in the next election.

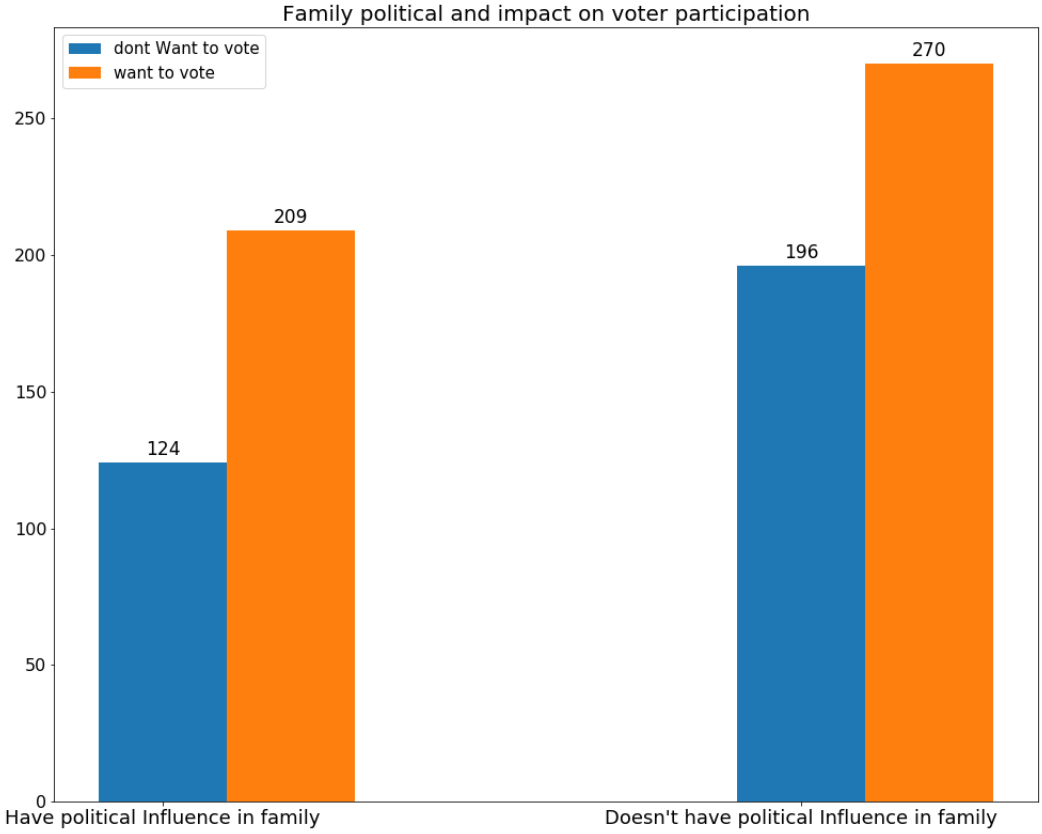


Figure 3.2: Illustration of Family Political Involvement and Voting Participation

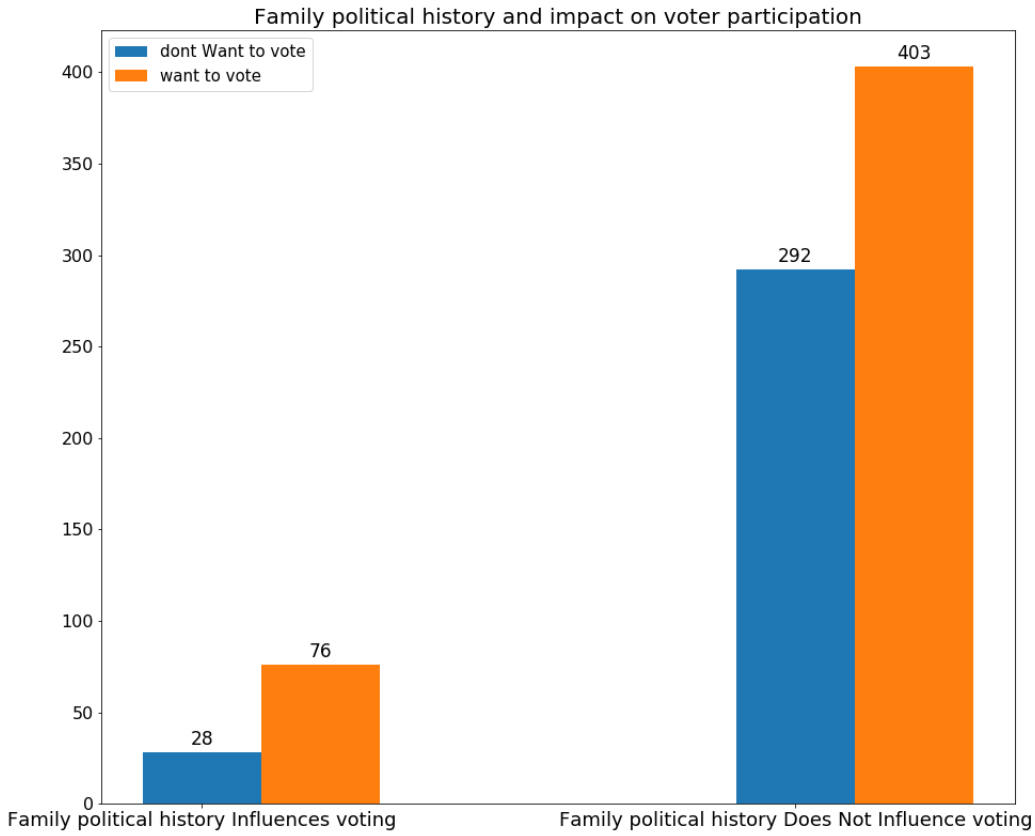


Figure 3.3: Illustration of Family Political Involvement and It's Direct Impact on Voting Participation

Repercussion of Family Education Credentials on Voting Participation

For many years scholars believe the fact that educational qualification of parents plays an important role in a child's wllingness to vote. The children tend to follow everything their parents do from very young age. As a result children of an educated family have a higher possibility to gain better educational qualification and political awareness. Education is one of the many effective ways to make people aware about their rights and their responsibilities towards the country. Additionally, educated people have a better understanding and decision making capability when choosing candidates. They would analyze the policies of different candidates and then select the best one who they think will preserve their best interest. Berinsky and Lenz[13] concluded in their paper that education and political participation does not have direct co-relation. However, college educated people who grew in a household where parents are college educated have higher participation rate in political activities. Additionally, the paper also concluded that upper class and middle class people in America have higher participation rate in politics because, of their higher education and awareness of civic rights and responsibilities. Similarly, in our country's capital Dhaka is the center of all political activity. Dhaka is the habitation for most of the highly educated people in our country. Moreover, most of the finest and prestigious educational institutions are situated in Dhaka. As a result Dhaka has become the political hub of Bangladesh with all the major parties head offices. As a result, we can say children of higher educated family background who also have pursued or pursuing higher education have higher possibilities of being active in voting and

other political activities. Among the 800 participants who took the survey 283 participant' father have an Honours or higher degree and the participants do not want to vote, although 413 participants' father have the same qualification and they want to vote. 23 of the participants' father have a HSC or equivalent degree and the participants do not want to vote. Although, 50 of the participants fathers' have HSC or equivalent degree and they still want to vote. Additionally, among the participants who's father have a SSC or equivalent degree 14 do not want to vote and 16 wants to vote. figure 3.4 illustrate the description we just explained now. Lastly, the participants who's mother have a Honours or higher degree 219 of them do not want to vote and 334 want to vote. On top of that, among all the participant who's mother have a HSC or equivalent degree, 70 participants do not want to vote and 91 want to vote. Again, 31 participants' mother have a SSC equivalent degree and they do not want to vote. whereas, 54 participants' mother have a SSC or equivalent degree and they want to vote. Figure 3.5 illustrate the scenarios that we explained above.

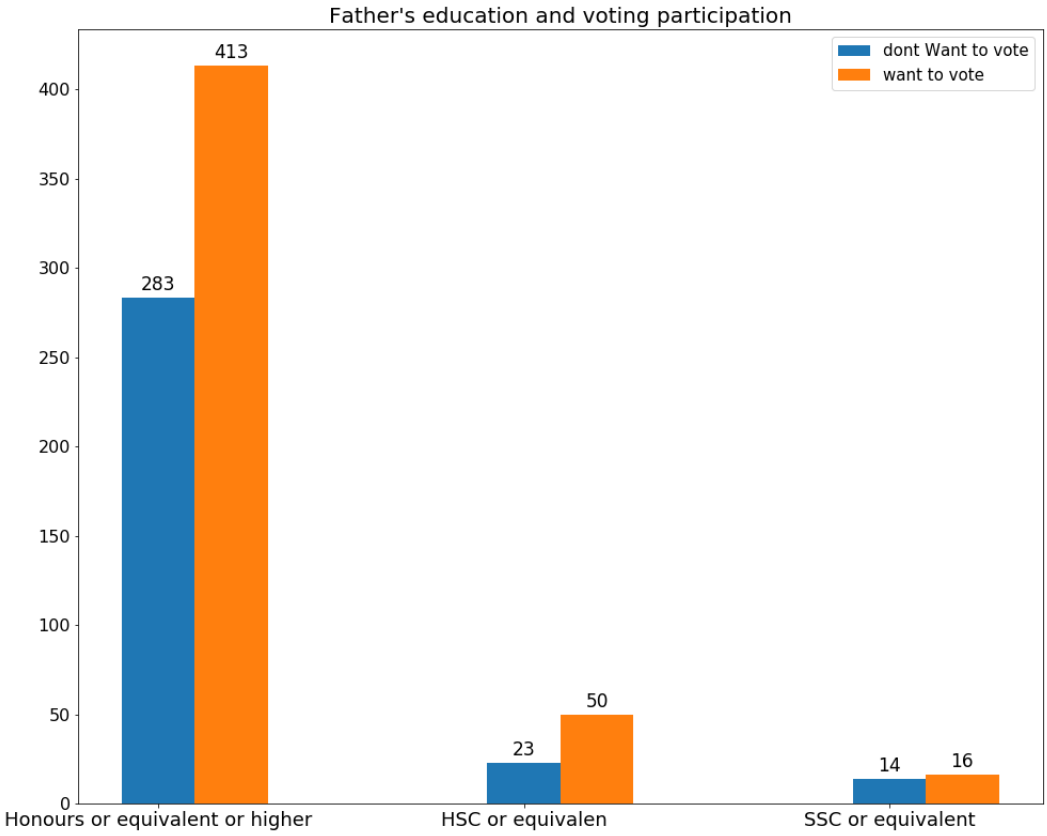


Figure 3.4: Illustration og Father's Educational Qualification and It's Impact on Voting Participation

Dissimilitude of Metropolitan Area and Countryside

The major cities are the epicenter of political activity. As a consequence, the young people who grew up in the major cities have a better understanding of political activities and also it's importance. According to Ward, R. E. (1960)[1], rural people do not participate in voting due to lack of strong sense of civic duty. Moreover, rural

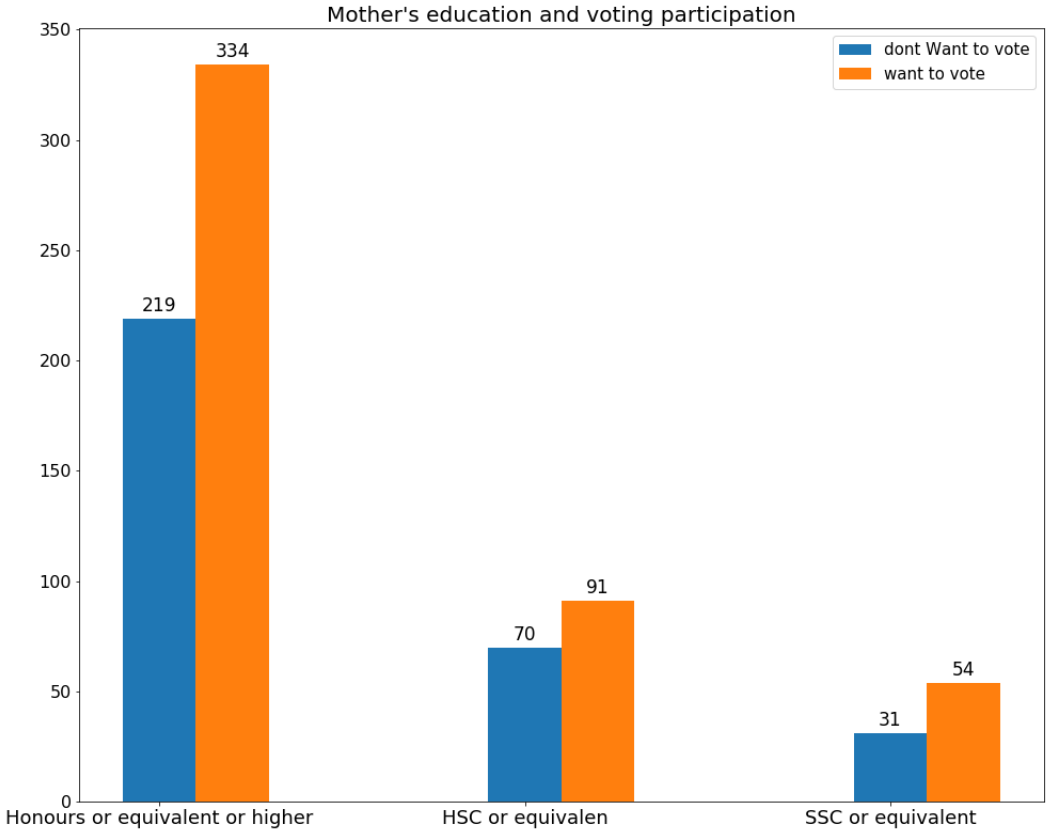


Figure 3.5: Illustration of Mother's Educational Qualification and It's Impact on Voting Participation

people do not analyze all the possible options and plans handed out by candidate. On the other hand, people in urban area involvement is motivated by strong sense of civic duty and responsibilities. Additionally, if we look at the history of our country, during independence higher educational institution of Dhaka, University of Dhaka mobilized and lead the movement. Additionally, after independence this university organized many other protest along with other educational institutions in Dhaka. As a result, educational institutions in Dhaka has always been the first one to start the political activity or a demand for change. In recent year most successful protest was on canceling the vat on education on private university students. Although, it started in Dhaka and all the private university around Bangladesh participated actively in this protest. This movement gained momentum very quickly and spread all over Bangladesh and government eventually have to remove their vat policy on private university. That is why, people of Dhaka city and those younger ones who grew up here is more politically aware. That is why we asked the participants about where they completed their higher education. Participants outside Dhaka move out of their home when they attend university. Another thing to point is that if someone completed their high school outside Dhaka they have a higher possibility of being exposed to participants politics. Also nowadays political parties focus their campaign more on other districts than Dhaka. As a result, we tried to investigate their home town.In our survey data we found that 236 participants completed their high school inside Dhaka and do not want to vote. Although, 317 participants completed their higher school inside Dhaka and want to vote. On top of that,

among people who completed high school outside Dhaka 84 participants do not want to vote and 162 participants want to vote.

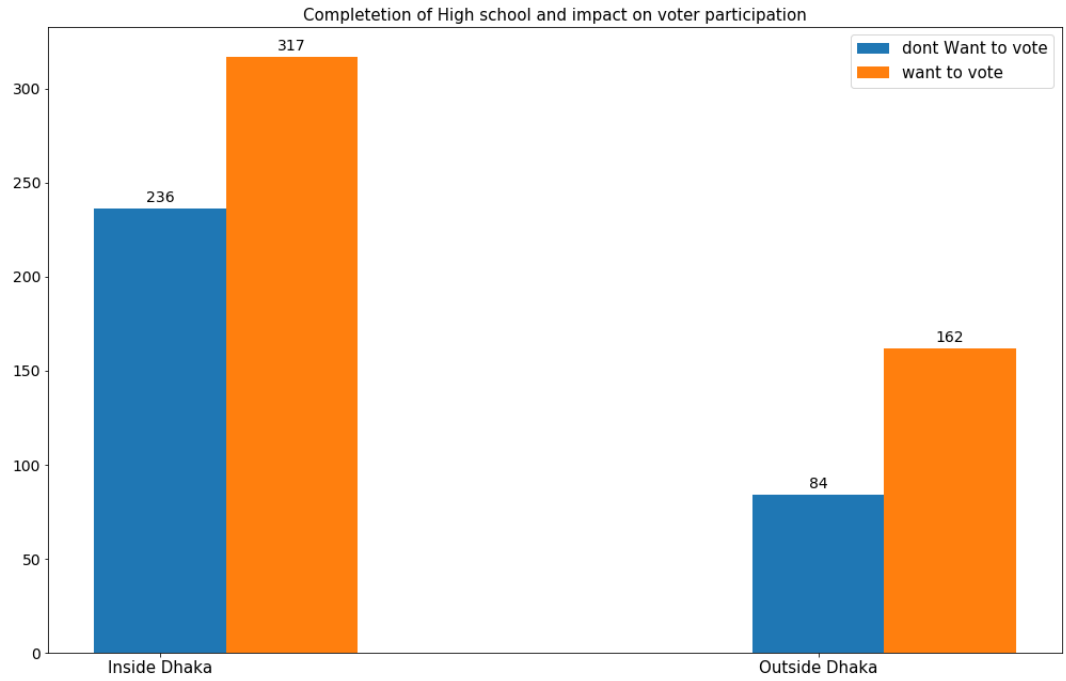


Figure 3.6: Illustration of High School and It's Impact on Voting Participation

Knowledge of Parliamentary System and Impact of Voting Participation

The access to the information about the process of electing representatives determines the participation of voters specially young voters.The exact process or the number of votes one person need to cast or whether the government is formed with people's direct votes or indirect votes is considered knowledge of parliamentary system. For example Bangladesh has a direct voting system to select the head of the acting government where as U.S.A has a system of electing their president by indirect votes. Proper knowledge of the process of election effects the young voters as many of them can be affected by the patriarchal nature of the household. From various research we can notice that many of the young voters follow passive politics which means they follow debates or go to meetings. But this doesn't mean that the young voters are willing to engage in the participation in politics. If the young voters are provided the clearer concept of the system they will be more encouraged to participate in the system.The young voters can also be influenced by not only the knowledge of parliamentary system but also the political opportunities that takes place before election like political debate and political discussions or campaigns[10]. The young voters need to broaden their knowledge about the MP's policy work. The parliamentarian need to expose global and national issues to the young voters. Among the participants who claim their knowledge to be 1 in scale of 5, 108 do not want to vote and 94 want to vote. Additionally, the participants who claimed 2 in scale of 5. 96 of them want to vote and 150 do not want to vote. When knowledge level is 3 out of 5 in parliamentary system, 63 participants do not want to vote and 133 of them want to vote. participants who give themselves 4 out of 5 in parliamentary system knowledge, 35 do not want to vote and 75 want to vote. Lastly,

the participants who think have full knowledge in parliamentary system, 18 of them want to vote while 27 of them do not want to vote.

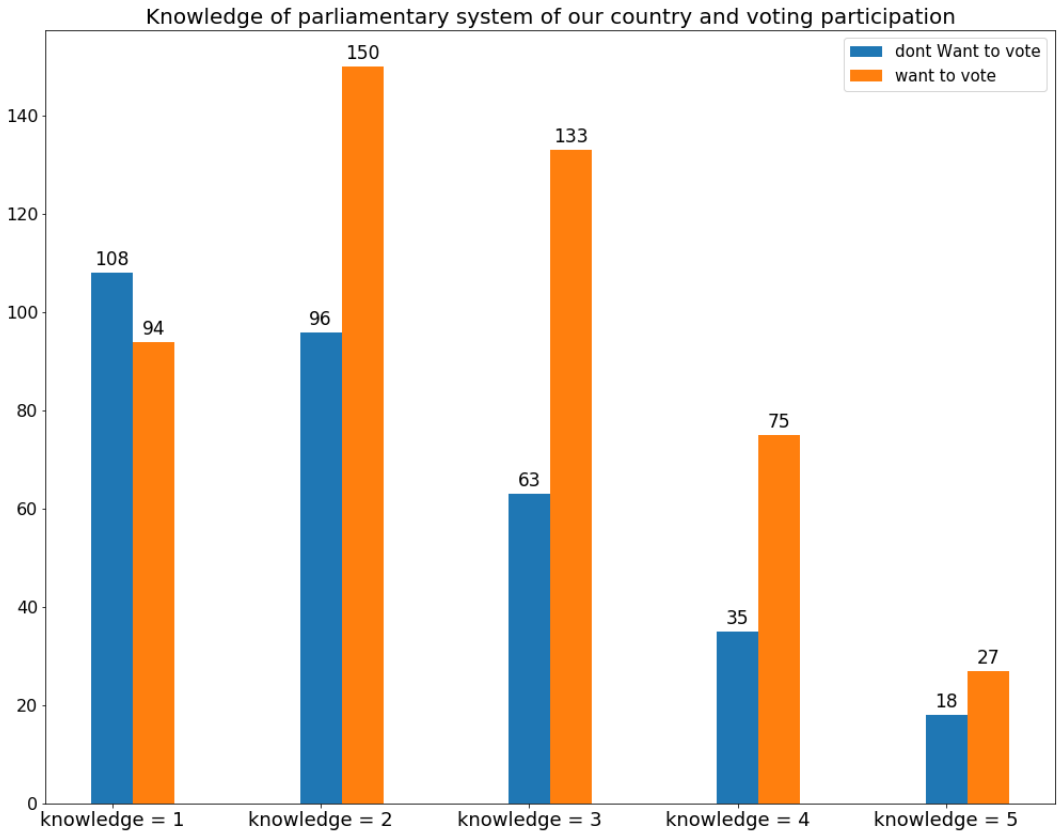


Figure 3.7: Illustration of Knowledge of Parliamentary System and Impact of Voting Participation

Impact of Financial Condition on Voting Participation

The experts believed for years that financial condition and voting participation have a strong co-relation. In American politics we can see that the wealthy people and big corporation can lobby to political office to gain their interest. The politicians tend to listen to wealthy people more than other people because they donate a lot of money in their campaign. The less wealthy people find it more difficult to get proper representation in politics as they do not have enough money to spend on their political campaign [3]. Large business sometimes influences the politics so much that they find their own benefit against the best interest of a lot of the people. Even in Bangladesh, wealthy and big corporation fund political campaign and influence political scenario. Moreover, wealthy people have more access to education then less wealthy people. Additionally, wealthy individual have the opportunity to run for election and office. Moreover, they can spend more resources on publicity and branding their image to the public. On the contrary, less wealthy people and lower income people do not have the privilege of that. Even though most of the people in our country our not wealthy but they do not have proper representation in local and national government. That is why we see a general apathy for politics in general. To understand financial situation of individual we asked two question, one of them

was about house ownership and other was car ownership. According to our finding participants who live in a rented house 117 participants do not wanted to vote and 162 participants will vote in the next election. On the other hand, the participants who live in a house of their own 203 do not have willingness to vote in the next election but 317 have willingness to vote. Moreover, the participants who do not have a car 108 of them do not want to vote and 275 of them want to vote. On the contrary, participants who do have a car 140 of them do not have willingness to vote and 240 of them are want to vote

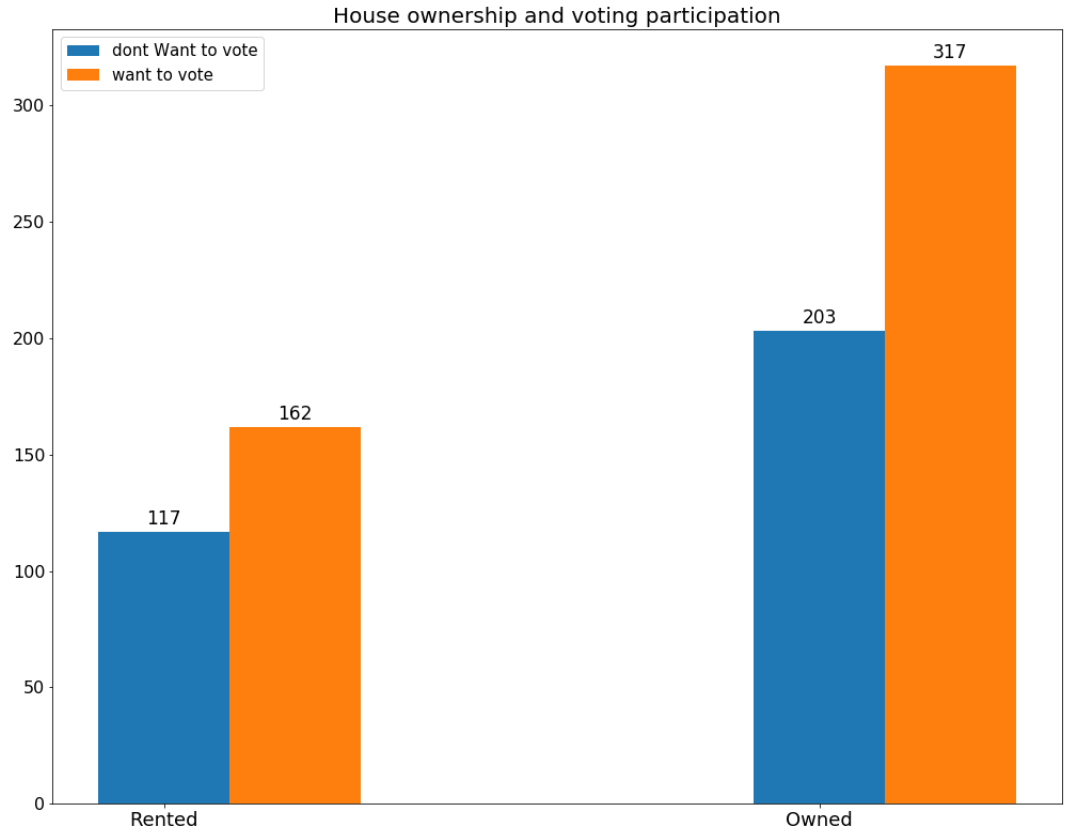


Figure 3.8: Illustration of House Ownership and Impact on Voting Participation

Impact of judicial Knowledge on Voting Participation

The knowledge of a state's general rules and it's enforcement has an important effect on the young voters. Even the initial qualification of judges should be known to the young voters. For example a judge can not have any relation to the defendant or his family members or anyone related to him. This assures the neutral perspective of a judge regarding the judgement. The right of a civilian according to the constitution is also an important factor for the young voters to know. Every citizen of a country should know their rights so that they can take proper measurement whenever there is a violation of rights. Voting is a basic right of any citizen of a state. The citizen needs to utilize this right to choose the proper government for everyone's well being. There are also specific rules regarding the process of judicial instruction. For example First, under Rule 201(e) a party who requests it must be given an opportunity to be heard with respect to the propriety of judicial notice, if not before then after

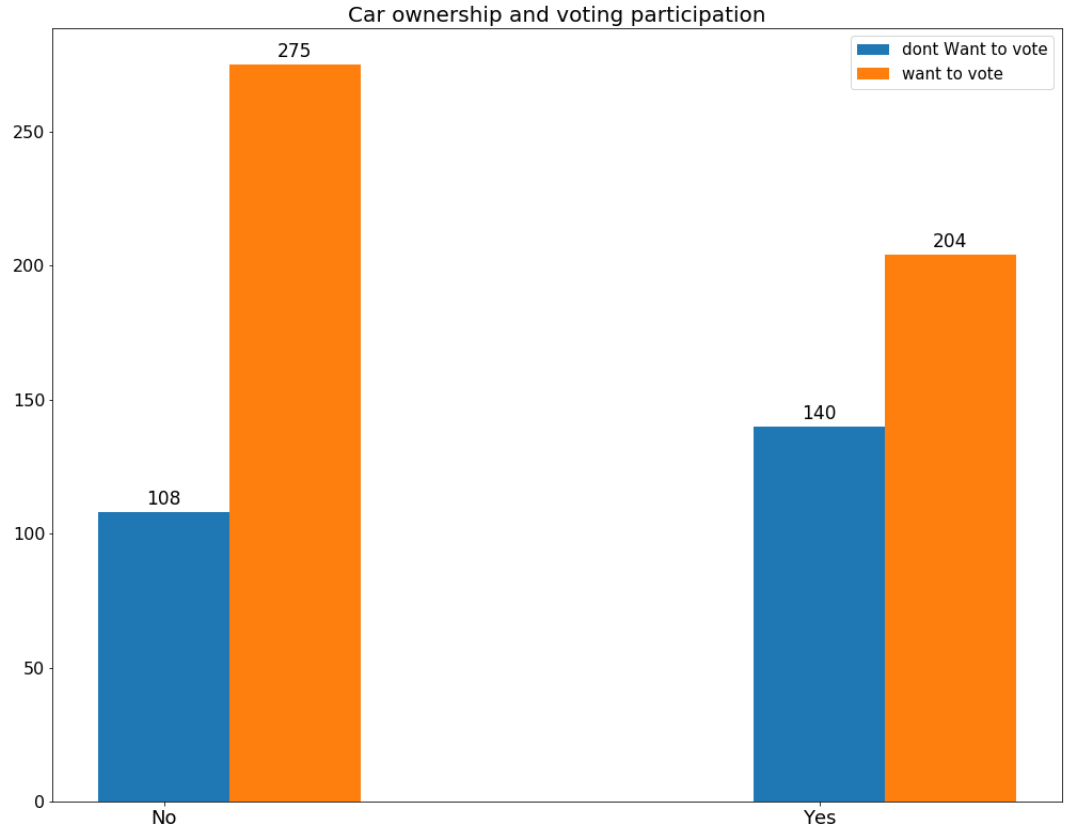


Figure 3.9: Illustration of Car Ownership and Impact on Voting Participation

notice is taken; in making the determination of noticeably, the court may utilize any appropriate source of information without regard to formal rules of evidence. The judge and the Jury board has to be neutral, they can not choose side and make any judgement based on their emotion. That's why both the judge and the jury board can not be related to the defendant in any aspect of their life. The transparency of the judicial institutions about their work and process should be known to the young votes so that they can have proper faith in the judicial system and therefore make their decision about participating in the election process. Among the participants who think themselves least knowledgeable about our judiciary, 134 participants do not want to vote and 112 participants want to vote in next election. Again, the participants who ranked themselves 2 on scale of 5, 90 of them do not want to vote and 155 of them want to vote. Additionally, the participants who think themselves at 3 on a scale of 5, 59 of them do not want to vote whereas 146 want to vote. On top of that, participants who evaluated themselves in 4 out of 5 for judiciary knowledge, 28 of them do not want to vote and 47 of them want to vote. Lastly, the participants who are most knowledgeable about their judiciary, 9 of them do not want to vote and 19 of them want to vote Figure 3.10 illustrates the findings.

Previous Voting Experience and Its Effect on Political Participation

Voting is the highest right to exercise in democracy. Voting is the process of selecting representative. As past voting experience can create a huge impact in the mind of the voters, everyone should be careful about this. Moreover, if a voter has a bad experience when voting it is most certainly possible that it would impact

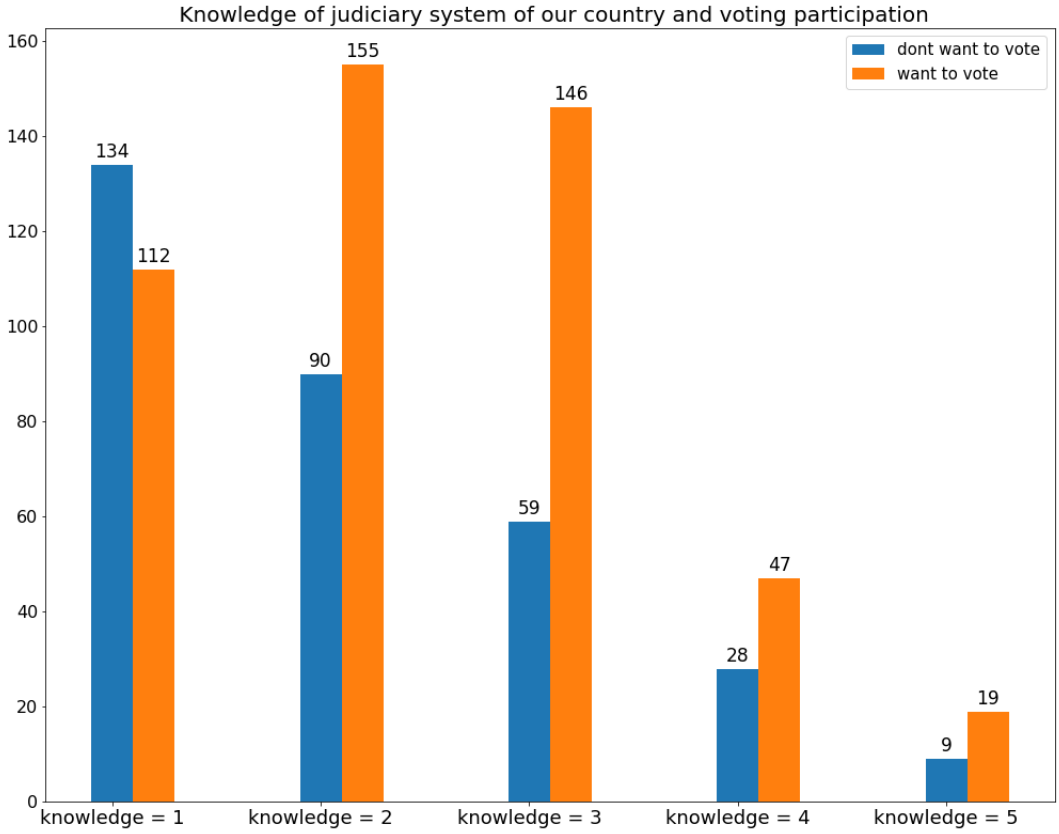


Figure 3.10: Illustration of Knowledge of judiciary System and Impact on Voting Participation

their participation in voting in the future. Most of the counties around the world have independent institution that conduct national and local election. Government cannot influence election in any form. Bangladesh have Election Commission which ensures free and fair election. Additionally, this commission creates a safe place for the voters to exercise their rights. Although, sometime technical difficulties may arise i.e. in last election some of them people's fingerprint were not matching with the database. Election Commission is working hard to solve these kinds of issues and ensure voter participation. Among all the topics we discussed in this chapter probably this topic is more important and directly related to voting participation. The young voters are probably the one who might vote for the first time or voted for the first time. As a result, any negative experience related to their voting will impact their voting participation and they will become more apathetic to election. That is why we tried to find out their previous voting experience. Out of the 800 participants 254 participants had no prior experience of voting and don't want to vote. 320 participants had no prior experience of voting and want to vote. 38 participants rated their experience 1 out of 5 and didn't want to vote, 62 participants rated their experience 1 out of 5 and wanted to vote. 11 participants rated their experience 2 out of 5 and didn't want to vote, 22 participants rated their experience 2 out of 5 and wanted to vote. 9 participants rated their experience 3 out of 5 and didn't want to vote, 32 participants rated their experience 3 out of 5 and wanted to vote. 2 participants rated their experience 4 out of 5 and didn't want to vote, 30 participants rated their experience 4 out of 5 and wanted to vote. 6 participants

rated their experience 5 out of 5 and didn't want to vote, 13 participants rated their experience 5 out of 5 and wanted to vote. Figure 3.11 shows the illustration of the findings.

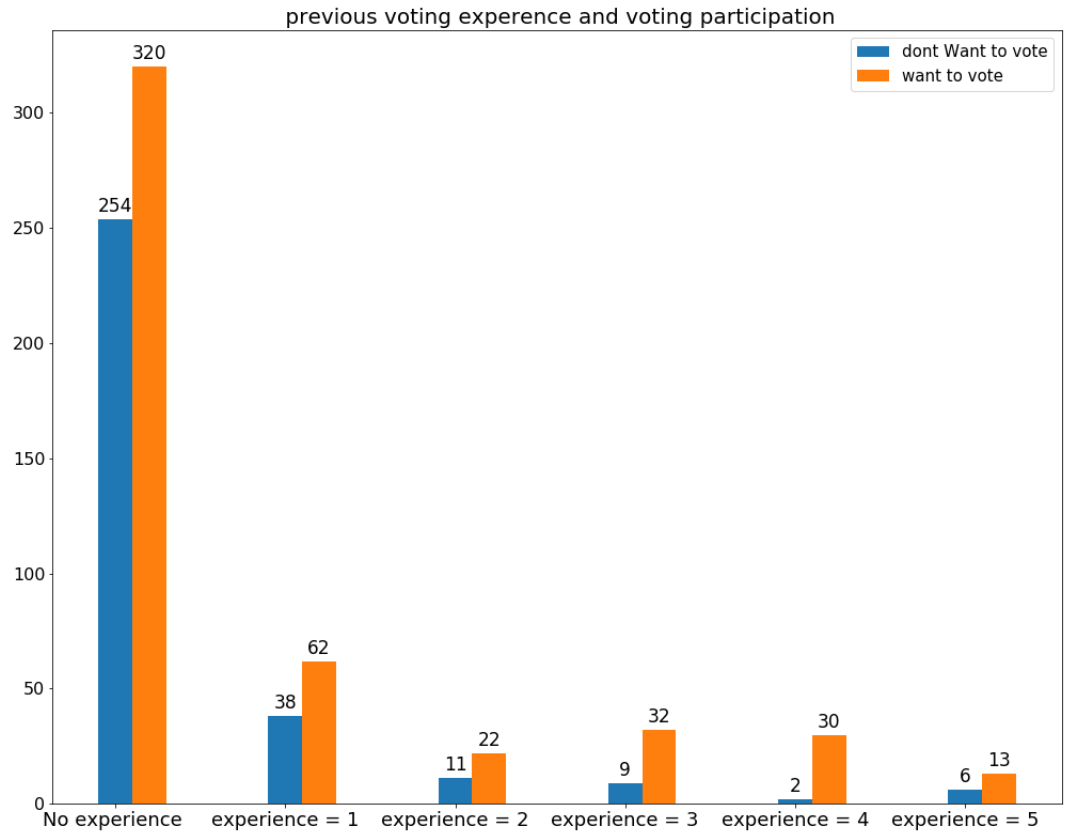


Figure 3.11: Illustration previous voting experience and Impact on Voting Participation

Parents Political Awareness and Its Impact On Political Participation of Children

Parent's political awareness plays a vital role in the young voters participation in election. Parents who are politically aware tend to discuss about the political incident to among each other and young voters usually live with their parents. So, they willingly or unwillingly get to know these political facts or news happening around them. For example: when the whole family takes their lunch or supper the parents usually discuss about various things and if they are politically aware then there is a possibility that they would discuss about political incidence too. This is an important source of political awareness for the young voters. Figure 3.12 shows the effect of political awareness on voter participation. 279 participants claimed that their parents are aware and they would not vote, 461 participants claimed that their parents are aware and they will vote, 41 participants claimed that their parents are not aware and they would not vote, 63 participants claimed that their parents are not aware and they still will vote.

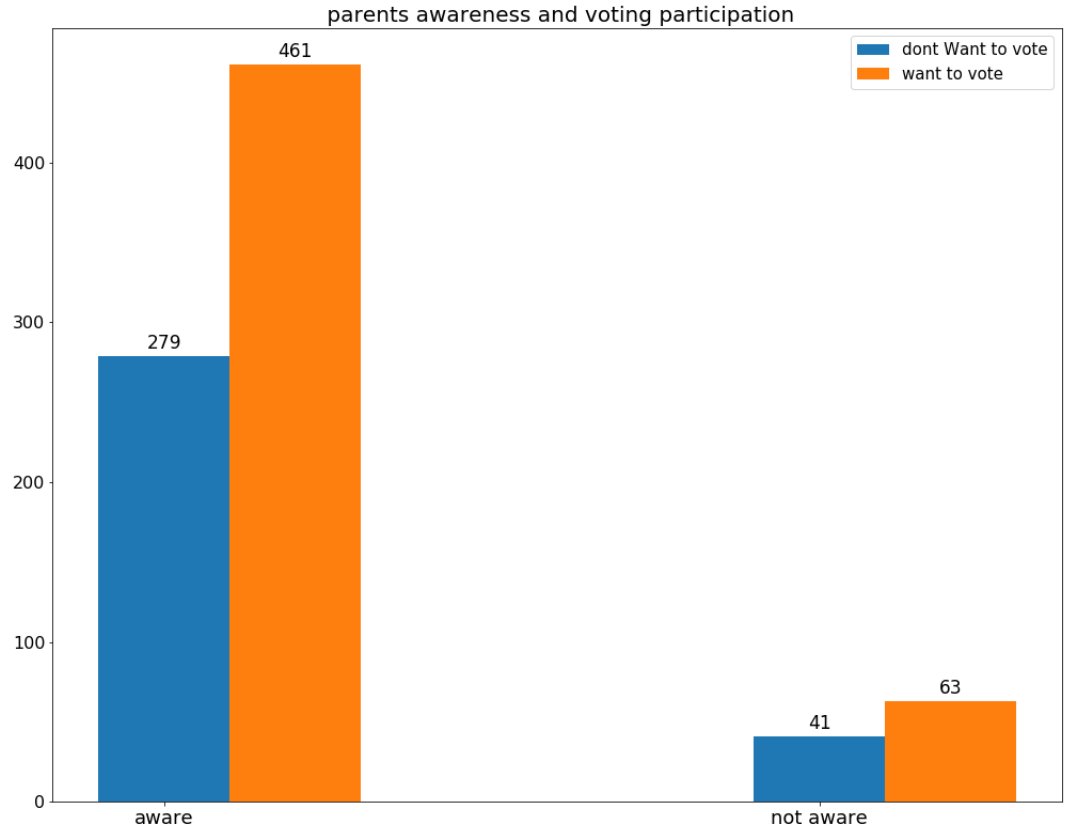


Figure 3.12: Illustration of Parents Awareness and Impact on Voting Participation

Parents Voting Representation

The way parents teach their kids about their political rights has an important effect on the young voters.The parent can represent the whole voting procedure as a crucial right and important responsibility to their kids which will make the kids eager to participate in the election system.They will be encouraged to give some thoughts about who to vote and who to not or about policies of different political parties. The parents can also represent the voting procedure as a negative system which will obviously bring negative impact on the kids.They will not take the system seriously and also will not participate in the election system let alone casting their valuable vote. The parents need to introduce the voting system as an very important civil right because this is the only direct way to choose a proper government for the country and ensure perfect democracy. The necessity of casting vote properly to choose a right candidate is a very important thing to teach to the next generations. Because the next generations are the ones who will lead the country and will pass on the wisdom and knowledge they have learnt over the years from their previous generations to their next generation. So, representing the whole voting procedure to the kids as an important civil right is something mandatory thing to do for the parents. To ensure more participation from the young voters the parents must carefully play this role. In our survey 161 participants reported that their father does not talk about voting and they don't want to vote, 132 reported their father does not talk about voting yet they still want to vote. 29 participants reported their father represents voting as trend and they do not want to vote and 49 participants reported their father represents voting as trend yet they still want to vote. 130 participants

reported their father represents voting as an important duty but they don't want to vote and 298 participants reported their father represents voting as an important duty and they to vote.on mother's side 204 participants reported that their mother does not talk about voting and they don't want to vote, 201 reported their mother does not talk about voting yet still want to vote. 41 participants reported their mother represents voting as trend and they do not want to vote and 79 participants reported their mother represents voting as trend yet they still want to vote. 75 participants reported their mother represents voting as an important duty but they don't want to vote and 299 participants reported their mother represents voting as an important duty and they want to vote. Figure 3.13 and 3.14 represents the information.

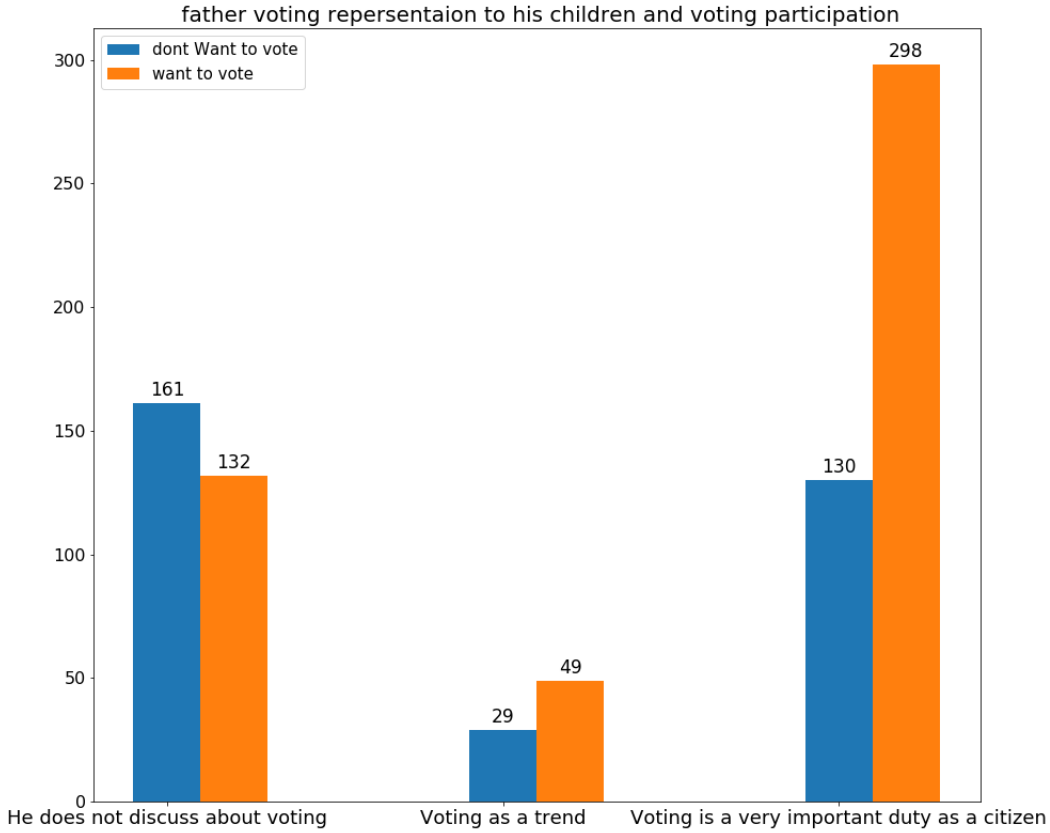


Figure 3.13: Illustration of fathers representation of voting to his children and Impact on Voting Participation

Parents Voting Enthusiasm and Its Impact on Young voter turnout

Family is the smallest form of social units. Family is where the children learn their first lesson as a sociological human beings. Parents are the first teacher of any children. The younger ones try to follow their parents in every aspect of their life. As a result, if they see their parent's enthusiasm from very early stage of their lives they also learn and act from it. In our country we see a trend that the children of political leaders later join the political party of their parents. On the contrary when parents misinterprets politics to their children and only focus on the dirty aspects of politics, as a result, the kids grow up and tend to hate any form of political activity.

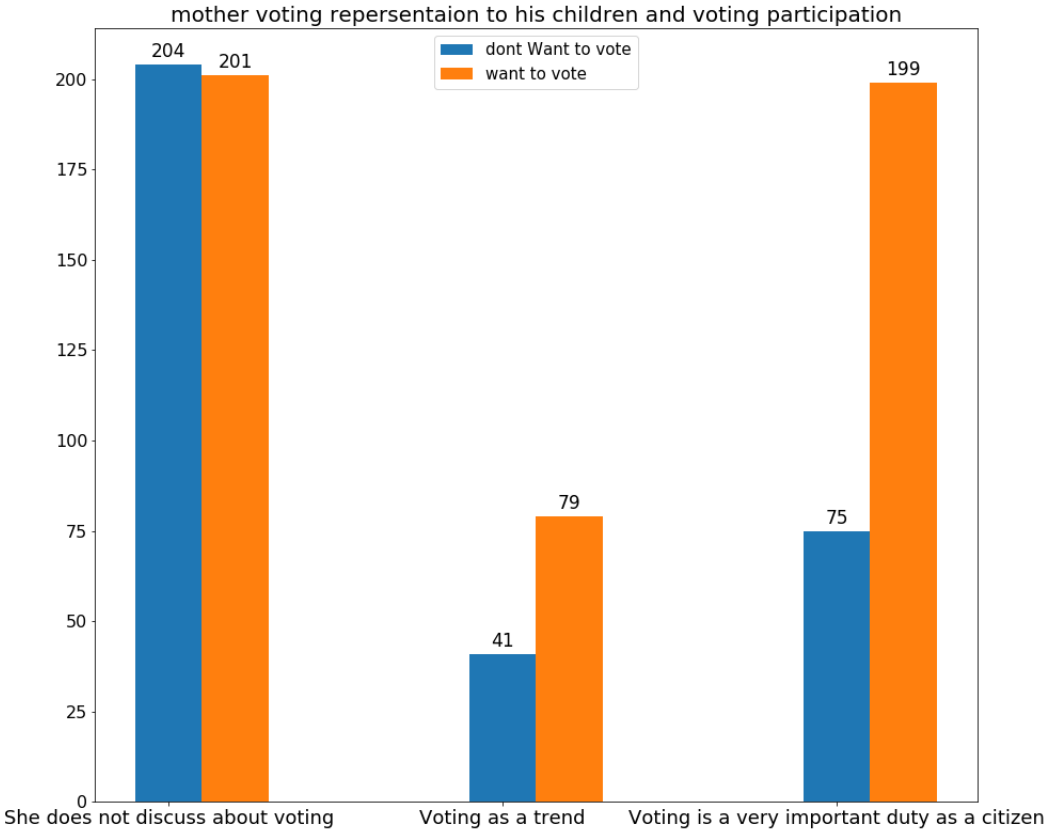


Figure 3.14: Illustration of mothers representation of voting to her children and Impact on Voting Participation

They do not even perform their civic duty such as: voting seriously. Moreover, the parent's sentiment about our country's history and how they celebrate the historical days touch the children's mind. Because of that, when they grow up they learn to value independence and voting right. Also from good parenting we can get our future responsible adult with leadership quality. For this reason, we asked the participants about the enthusiasm of their parents. On top of that, we also tried to find its impact on their willingness to vote. Out of 800 participant 356 were willing to vote and their parents were enthusiastic about voting and 176 were not willing to vote even though their parents were enthusiastic to vote. 144 participants were not willing to vote and their parents were not enthusiastic to vote and 123 participants were willing to vote even though their parents were not enthusiastic to vote. in figure 3.15 parent's voting and it's impact on young voter is illustrated.

Age Group and its impact on voting Enthusiasm

Age is one of the most important factors in terms of determination of voter turnout or willingness to vote. People in youth mostly aging from 18 to 20 deem voting as a non-essential duty. In the process of getting educated and determining job prospects, the youth is too busy to go to vote and determine the future of the nation. Most of the youth population are aware of their voting rights. However, due to social and economic factors they are sometimes unable to exercise their rights. People in the age group of 20 to 35 are a minor significant portion of people to actively exercise their voting rights. Among the factors that they feel enthusiastic about voting involve

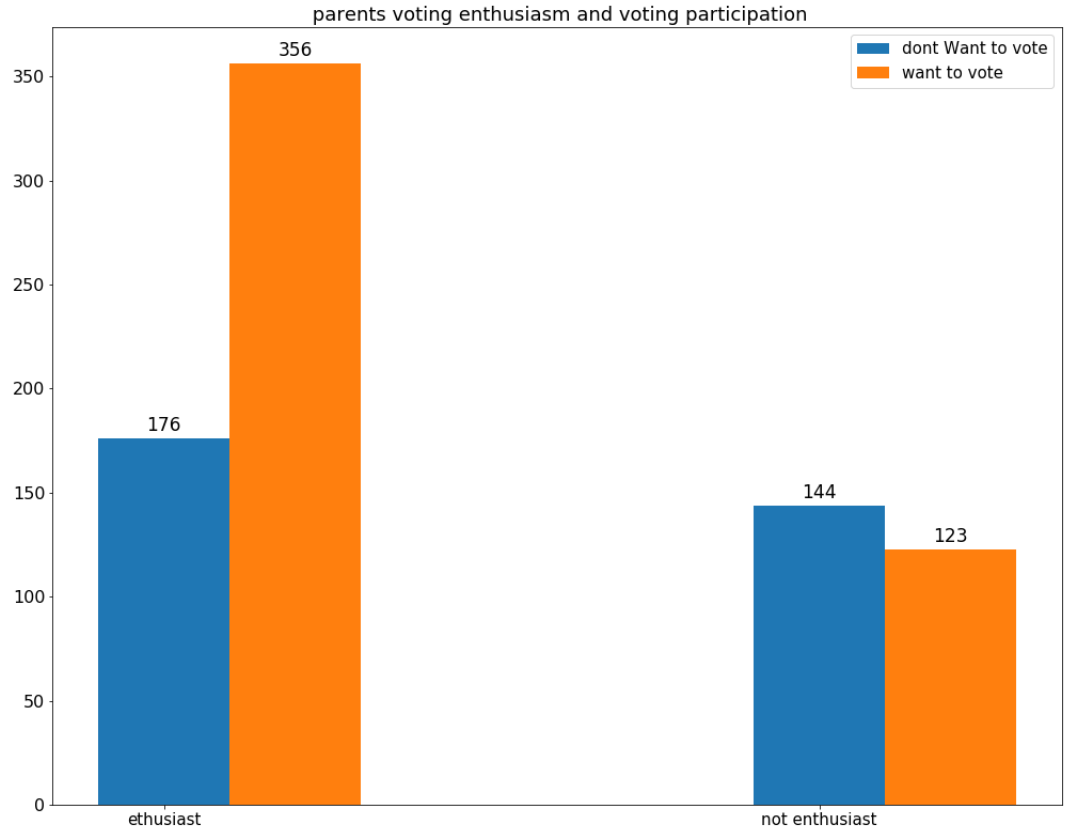


Figure 3.15: Illustration of Parents enthusiasm and Impact on Voting Participation

hope of improvement in economy which in turn can make their life better, a better way of living and a better future for their family. Whereas, people in the age group above 35 are most active and accumulate a larger portion of a countries voting poll. People in this age group are mostly concerned about the well-being and development of their country. Moreover, with evolving maturity over the span of their life this age group of people has the right mentality and decision taking capability to choose the right candidate to lead their country. Moreover, young people of this decade may experience situation of political struggle. As a result they are more apathetic then their previous generation. Among the participants whose age is between 18 to 20 years old, 113 participant do not want to vote and 172 want to vote. Moreover, by all of the participants from age 21 to 23 years old, 186 participants do want to vote and 291 participant want to vote. Additionally, between the participants from age 24 to 26, 20 participants do not want to vote and 16 participant want to vote. Lastly, only one participant was above the age of 26 years and he/she does not want to vote. In figure 3.16 age group and it's impact on voting enthusiasm is illustrated.

3.3 Data Pre-processing

As we have collected our own data set it had some drawbacks due to which we could not feed it to any machine learning algorithm. We had to do some pre-processing before using any classification algorithm. We had to check if there were any null values or if someone made any blunder while filling out the forms. Pre-processing will ensure better prediction and accurate result of our sample space.

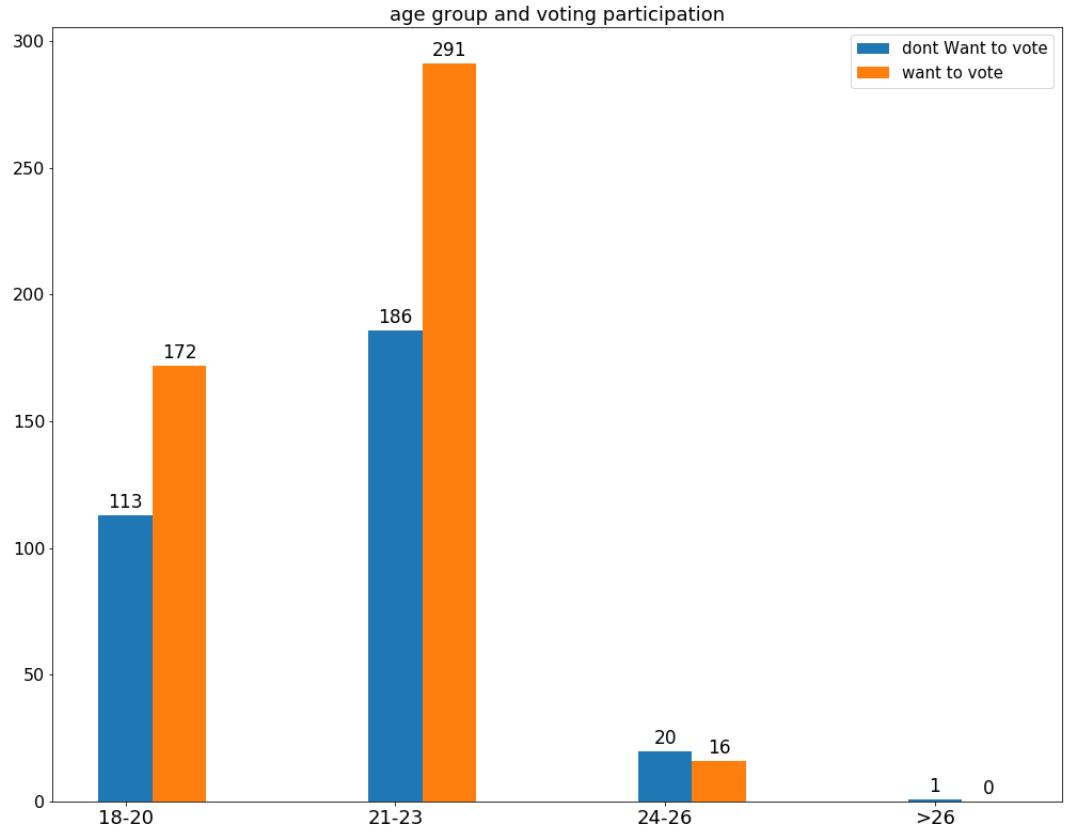


Figure 3.16: Age Group and Impact on Voting Participation

3.3.1 Handling categorical variables

Categorical data holds category of values rather than direct numerical values. Categorical data can not be directly used to fit in any machine learning algorithm. It will give prediction to some unreasonable outcome or predict misleading output. One of the ways to deal with categorical data is to use label encoder. Although label encoder have some limitation. Label Encoder thinks there is a natural relationship in the categorical data. In our data set we have columns that consist of text data. For example, there is a column in which father's and mother's educational qualification is stored. The values of this columns are "SSC or equivalent", "HSC or equivalent" and "Honours or higher". Label encoder will label them as "SSC or equivalent" as 1, "HSC or equivalent" as 2 and "Honours or higher" as 3. It will assume that there is some order as they are labeled 1,2 and 3. In reality there is not any order. Moreover it can also have misleading results or unreasonable prediction i.e halfway between the categories. Which is not acceptable. For this reason we are using One Hot Encoding method [48]. It will create new column for every category. For example for our data set it will create new column for "SSC or equivalent", "HSC or equivalent" and "Honours or higher" each. Then it will store 0 or 1 in the rows. 0 (zero) means not in this category and 1 (one) means exist in this category. This way we prepared our data set to be fit in classification based machine learning algorithm.

3.3.2 Handling Imbalanced Data set

Sometimes when working with real life data set we faced data sets with imbalanced classes. As it is known that machine learning algorithm is bias to the sample that are majority in the data set. For example, we are trying to find out if a young adult wants to vote or not. This is a classification problem. In our findings we got that, very few people are not willing to vote. So, we have to over sample or under sample our data in such a way that it does not effect the representation of the data set.In imbalanced data set there will be presence of minority class. If any machine learning algorithm is applied at this stage, misleading accuracy will be generated. We can deal with imbalanced data set by random under sampling or random over sampling. Although, these method will discard some of the valuable information of the existing data set and can also lead to bias. Synthetic Minority Over-sampling or SMOTE works in these following steps [7]. SMOTE is an state of the art over sampling technique. At first, it plots all the data points in the data set. Then identifies the feature vector and its nearest neighbour. After that, it finds out the difference between the data points. In other words, it finds out the linear distance between any two data points and multiply it by a random number from 0 (zero) to 1 (one). After that it plots the new point in the line with the result it got from the previous step. As a result we get new synthetically generated data points. SMOTE repeats the process to get more synthetically generated data points. In this way SMOTE removes the bias of the majority class. We have used SMOTE to handle the class imbalance by oversampling our minority class.

Chapter 4

Analysis and Interpretation of Experimental Result

In the previous chapters we have discussed how we have collected the data and formed our data set and description of each features in our data set. In this chapter we will use various machine learning algorithm to find out the most influencing features from all the features we have selected. After that we will apply some learning algorithms on those features to predict whether anyone is likely to vote or not. As we have mentioned before the objective of our research is two-fold. Firstly, finding the factors that motivate a young voter to participate in the voting process or repel them and secondly, we will try to predict whether someone will participate in voting based on those factors. This chapter will be divided into two parts with first part focusing on finding the most influential factors and create reduced feature set and second part to use the reduced feature set to try to predict whether someone will participate in the voting process or not.

4.1 Prominent Feature Selection Methods

There are many available algorithms or procedures to find out feature importance. Most of them fall into three category and they are: filter method, embedded method and wrapper method. We will apply RFECV which stands for Recursive Feature Elimination with Cross Validation and it falls under wrapper method. [44] [15]. We are using RFECV instead of RFE because, it has cross validation capability. We tried other feature selection methods and got best result with RFECV that is why we will discuss about RFECV only. We will discuss the process in the following section in details.

4.2 Applying Recursive Feature Elimination with Cross-validation

Recursive Feature Elimination with Cross-validation also known as RFECV as from it's name suggests recursively removes features then, builds a model and evaluates the newly created model with the newly created feature set. RFECV gradually finds the best subset of features and then ranks them. RFECV keeps creating

models until all the features are exhausted. We have to pass a classifier based on which it evaluates the feature importance. We primarily applied five models in our work and they are: Random Forest, XGBoost, Linear SVM, Naive Bayes and Kernel SVM with Gaussian Kernel. As mentioned before RFECV requires a classifier to be passed as it's argument so, we used RFECV with Random Forest, XGBoost and Linear SVM. We had to use Linear SVM instead of Kernel SVM as, Kernel SVM does not have the "coef_" attribute which is used by RFECV to perform the feature selection [44]. In the following section we will briefly discuss how we implemented RFECV using each algorithm, the selected features we got and the parameters we used.

4.2.1 Applying RFECV with Random Forest

As mentioned before we used RFECV with Random Forest. The input parameters used were: we passed a Random Forest Classifier with default parameters as "estimator", we used "step" value 1 which signifies how many feature(s) will be deducted in each iteration, we passed a cross-validation ("cv") value of 5 as we intend to perform a 5 fold cross validation to properly evaluate the model, we passed accuracy for "scoring" parameter and we passed -1 for "njobs" parameter. In figure 4.1 we plotted the accuracy obtained vs. the number of features selected.

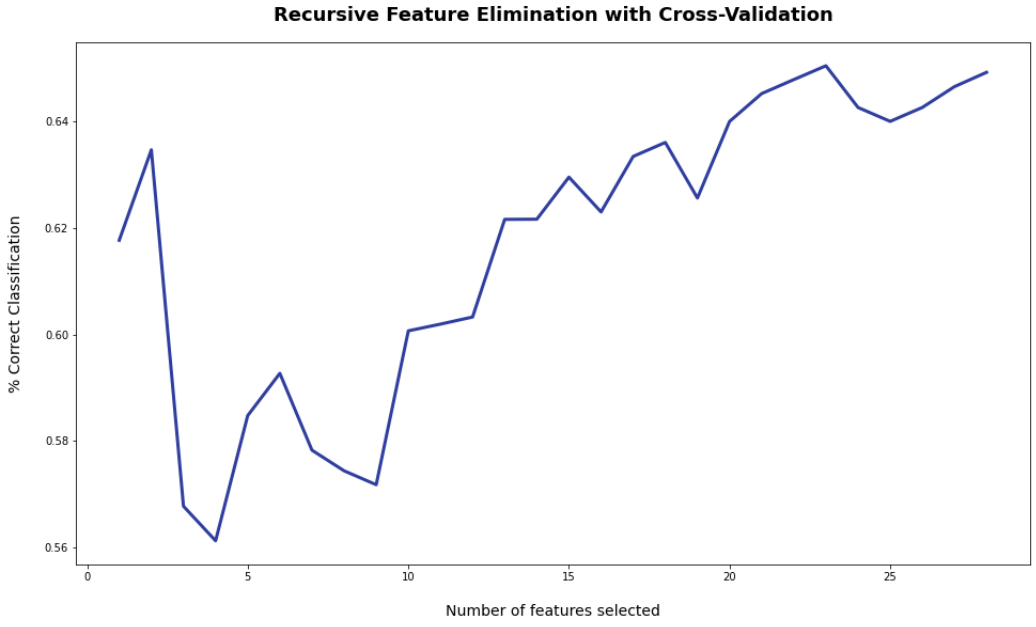


Figure 4.1: Accuracy vs. Number of features selected by RFECV using Random Forest

From the figure 4.1 we can see we get highest accuracy score is achieved when the number of features selected were 23.

The reduced feature set contains: age, presence of political person in family, influence of presence of political person in family in survey participants voting, father's

education, father's representation of voting, financial condition, where the participant completed high school,mother's education, mother's representation of voting, whether the participant reads newspaper or not, parents' political awareness, parents' enthusiasm to vote, participant's social media usage, participant's political awareness, participant's knowledge of functionality of parliamentary and judicial systems, support for democratic system and lastly participant's previous experience of voting. RFECV pointed out that this 23 features gave the best accuracy so, we will use this reduced feature set to predict voter participation when using Random Forest which we will discuss later.

4.2.2 Applying RFECV with XGBoost

We also applied RFECV with XGBoost. XGBoost is a fairly new optimized, distributed gradient boosting library. The input parameters used were: we passed a XGBoost Classifier with default parameters as "estimator", we used "step" value 1 which signifies how many feature(s) will be eliminated in each iteration, we passed a cross-validation ("cv") value of 5 so, we intend to perform a 5 fold cross validation to properly evaluate the model, we passed accuracy for "scoring" parameter and we passed -1 for "njobs" parameter. In figure 4.2 we plotted the accuracy obtained vs. the number of features selected.

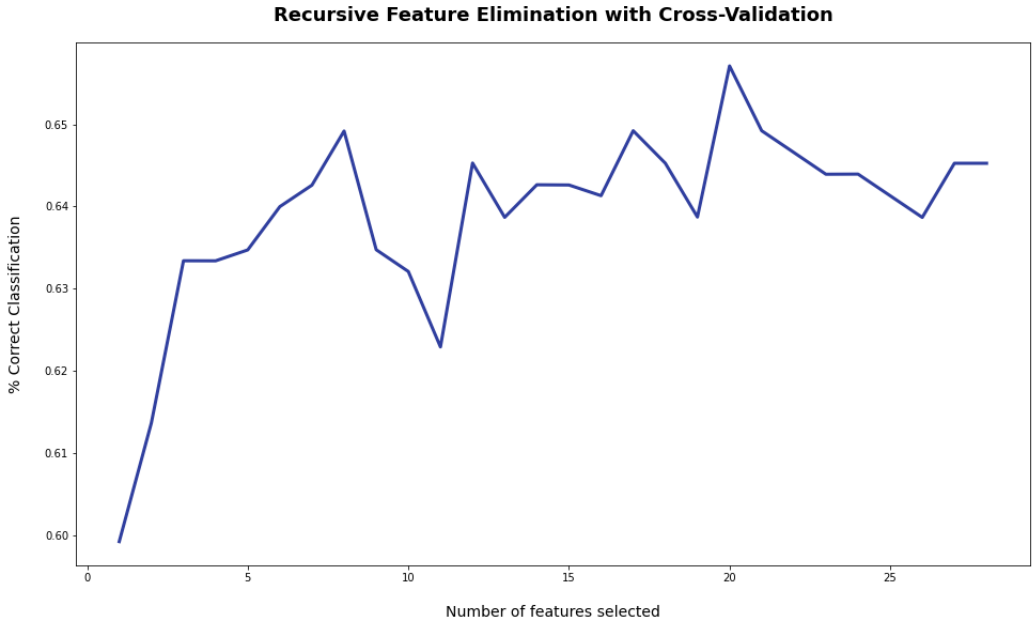


Figure 4.2: Accuracy vs. Number of features selected by RFECV using XGBoost

From the figure 4.2 we can see we get highest accuracy score is achieved when the number of features selected were 20.

The reduced feature set contains: age, influence of presence of political person in family in survey participants voting, father's education, father's representation of voting, financial condition, where the participant completed high school, mother's representation of voting, whether the participant reads newspaper or not, parents'

political awareness, parent's enthusiasm to vote, participant's social media usage, participant's political awareness, participant's knowledge of functionality of judicial system, support for democratic system and lastly participant's previous experience of voting. RFECV pointed out that this 23 features gives the best accuracy so, we will use this reduced feature set to predict voter participation when using XGBoost which we will discuss later.

4.2.3 Applying RFECV with Linear SVM

We used Gaussian or RBF kernel SVM and Linear SVM for prediction so, naturally we had plan to perform RFECV with both Kernel SVM and Linear SVM but, we only used Linear SVM for feature selection because, Kernel SVM or in our case Gaussian SVM lacks the "coef_" attribute which is used by RFECV to perform feature selection [44]. We will use the reduce feature set found from applying Linear SVM and RFECV for applying both the Linear SVM and Kernel SVM. For the case of linear SVM we passed default linear SVM with "C" value of 1.0 as "estimator" parameter for RFECV and again like before we passed value of "step" parameter as 1, we passed 5 as value for "cv" because, we intended to perform a 5 fold cross-validation so that the classifier is evaluated properly on a portion of data-set that is representative of the whole data-set, we passed accuracy for "scoring" parameter and "njobs" as -1. In figure 4.3 we plotted the accuracy obtained with vs. the number of features selected.

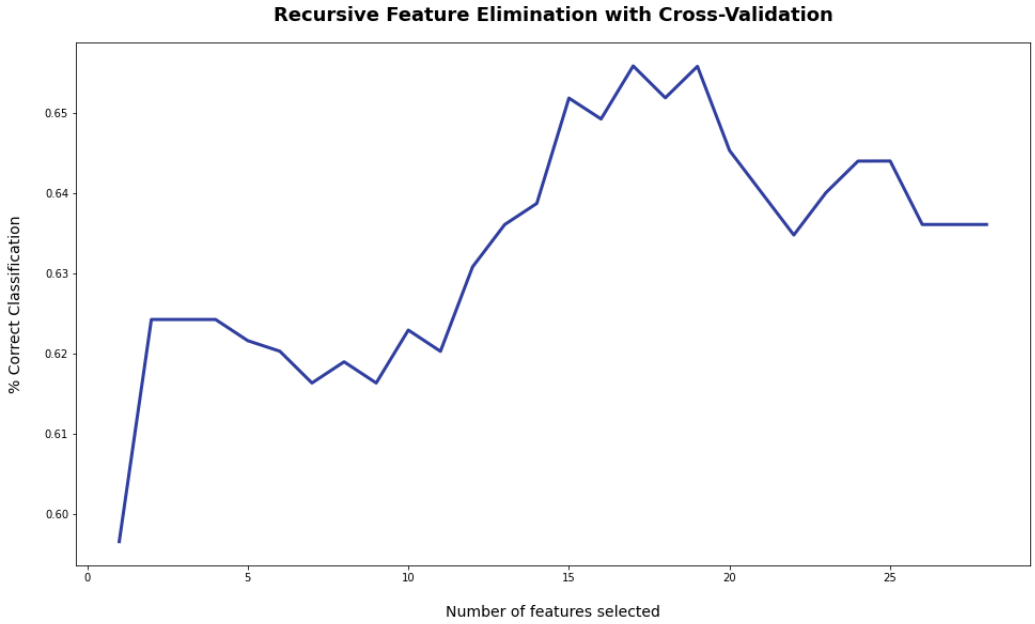


Figure 4.3: Accuracy vs. Number of features selected by RFECV using SVM

From the figure 4.3 we can see we get highest accuracy score is achieved when the number of features are 17.

The reduced feature set contains: age, influence of presence of political person in family in survey participants voting, father's education, father's representation of

Algorithm name	Number of features common with Random Forest	Number of features common with XGBoost	Number of features common with SVM
Random Forest	23	18	16
XGBoost	18	20	16
SVM	14	16	17

Table 4.1: Comparison of common features between algorithms

voting, where the participant completed high school, mother's representation of voting, whether the participant reads newspaper, parents' political awareness, parents' enthusiasm to vote, participant's social media usage, participant's knowledge of functionality of judicial system and participant's previous experience of voting. RFECV pointed out that this 17 features gives the best accuracy so, we will use this reduced feature set to predict voter participation when using bothvRBF SVM and Linear SVM which we will discuss later.

4.3 Finding the co-relation of each features in the data set

We also used co-relation of each feature with the label to try to find which features are linearly co-related with the label. Though the data-set is hardly linear but, as we have used RFECV so far we tried to do something different and see if it backs the already found feature set.The features are: father's representation of voting, mother's representation of voting, parents' enthusiasm to vote, participant's social media usage, whether the participant reads newspaper or not, influence of presence of political person in family in survey participants voting, knowledge of functionality of parliamentary and judicial systems, participant's previous experience of voting and participant's political awareness. The features corresponds to the feature we have found earlier from the output of RFECV applied by different algorithms. So, we can concretely say at this point that these features are in fact responsible behind young voter enthusiasm.

4.4 Final Verdict

As mentioned before finding the most influencing features responsible for influencing a young voter to attend or avoid participating in the voting process are one of our two primary goals of this study. We have extensively discussed about finding these features in this chapter. We have used RFECV using three different classifiers which are: Random Forest, XGBoost and SVM. We have seen that they give the best accuracy when number of features are taken are: 23, 20 and 17 for Random Forest, XGBoost and SVM respectively. Table 4.1 gives a general idea about the overlapping number features among the reduced feature set selected by each algorithm.
So, from table 4.1 we can see that the feature set generated by Random Forest has

18 features common with XGBoost's feature set, 16 features common with SVM's feature set. XGBoost has 18 features common with Random Forest's feature set and 16 features common with SVM's feature set. SVM has 14 features common with Random Forest's feature set and 16 features common with XGBoost's feature set.So, we can say that majority of selected features are common between the three algorithms and also they corresponds with the features carrying high co-relation with label. For further safety we will union each reduced feature set created by each algorithm to find the absolute feature set. The features present in union feature set is showed in figure 4.2.

Most influencing features behind young voter enthusiasm
age_21-23
age_24-26
fam_pol_Yes
fam_pol_2_Yes
father_edu_Honours or equivalent or higher
father_edu_SSC or equivalent
father_rep_vote_Voting as a trend
father_rep_vote_Voting is a very important duty as a citizen
finance_1_Rented
finance_2_Yes
high_school_Outside Dhaka
mother_edu_Honours or equivalent or higher
mother_rep_vote_Voting as a trend
mother_rep_vote_Voting is a very important duty as a citizen
newspaper_There is subscription and me and my family members read them daily
newspaper_There is subscription but I don't read them daily but my family member do
parents_aware_Yes
parents_enthu_Yes
social_No, I don't use social media
social_1 hour
social_3 hours
pol_aware
parliament
judicial
demo_support
vote_xp

Table 4.2: Most influencing features behind young voter enthusiasm

Figure 4.2 contains the union of all the features selected by each iteration of feature selection we performed. As we can see from the figure the selected features

are: participant's age, presence of political person in family, influence of presence of political person in family in survey participant's voting, father's education qualification, father's representation of voting, financial condition, where the participant completed high school, mother's educational qualification, mother's representation of voting, whether the participant reads newspaper or not, parents' political awareness, parents' enthusiasm to vote, participant's social media usage, participant's political awareness, participant's knowledge of functionality of parliamentary and judicial systems, support for democratic system, participants previous experience of voting. To sum up we can say that describing the process of finding these features were the objective of this part.

4.5 Prediction

In the previous part we have completed the feature selection process and we have found out the optimal features needed for the classification. As mentioned before the objective of this study is twofold: first objective is to find out the features which motivates or repels a young voter to vote and the second objective is to use those features to try to predict whether someone will participate in the voting process or not. This chapter will focus on the second objective. In this chapter we will apply multiple algorithms and will discuss and analyze the results of each of the algorithms to find out which one performs the best. We also have class imbalance in our data set as mentioned before so, we will also apply an oversampling method called SMOTE on the algorithms to try to counter the effects of imbalanced classes and again apply classifications algorithms and then we will analyze the results between feature set with imbalanced class and feature set balanced class [8].

4.6 Applying Classification Algorithms

Our study primarily revolves around investigating whether a person will participate in the voting process or not which is basically a binary classification. There are many prevalent algorithms available for binary classification but, there is a catch, The data set we are using as mentioned before are prepared by us and due to lack of resources we could not collect more than 800 samples and the data set has on an average 20 features after feature selection. Furthermore, the survey participants did not fill up the survey form properly thus there are a lot of noise in the data. Another problem that we faced was that the classes of our data set was imbalanced. There was substantial amount of difference between the number of participant who want to vote and the number of participant who do not want to vote. We will counter this class imbalance later in this chapter. For now as we mentioned before due to the scarcity of data we failed to apply any kind of neural network model. We selected four popular models which are: Random Forest, XGBoost, SVM and Naive Bayes. We will generate accuracy score, confusion matrix, precision, recall and F1-score for each algorithm and will use them to compare the performance of each algorithm on our data set and on the overall performance of our work.

To understand precision and recall first we have to understand what is false positive, true positive, false negative and true negative in our work. True positive in our case

is if a participant wants to vote and the system classifies the participant as he/she is willing to vote, true negative is that if the participant is not willing to vote and the system classifies the participant as he/she is not willing to vote, false negative is if a participant is willing to vote but, the system classifies that he/she is not willing to vote and lastly false positive means if a participant is not willing to vote but, the system classifies them as they are willing to vote. As we can understand the last one is the most fatal case as if there are some false negatives then it is no major headache as at the end of the day they will vote but, if the system classifies someone as they will vote but, in reality they do not then it is a big problem if we want to achieve perfect democracy.

Now, as we have gotten a grasp on the concept of what is true positive, true negative, false positive and false negative it would be easy to understand what is precision and what is recall. Precision is the ratio between true positive and the sum of true positive and false positive and on the other hand recall is the ratio between ratio between true positive and sum of true positive and false negative [16].

$$Precision = \frac{True\,Positive}{True\,Positive + False\,Positive} \tag{4.1}$$

$$Recall = \frac{True\,Positive}{True\,Positive + False\,Negative} \tag{4.2}$$

From equation 4.1 and 4.2 it is apparent that in our case it is more important to have a higher value of precision than recall.

Now as for F1-score combines both precision and recall into a single metric. It is the weighted harmonic mean precision and recall. F1-score considers both false positive and false negatives. F1-score is more effective for evaluation model when it is applied on an uneven class. F score is used in information retrieval for measuring document classification, query classification performance etc [33]. The following equation shows how F1-score is measured [16].

$$F1\,score = \frac{2}{Recall^{-1} + Precision^{-1}} \tag{4.3}$$

From equation 4.3 we can see that it is important to have a high value of F1-score for a classifier to perform good.

Another way to evaluate a model is by using AUC(Area under the curve) and ROC(Receiver operating characteristics curve) curve. It is also known as AUROC. ROC is a curve which represents probability and AUC represents how well a model can separate classes [16] [37]. A model can have a AUC value between 0.5 to 1 with 0.5 representing a bad classifier and 1 representing a good classifier [37] [39]. For example: if a breast cancer classifier's AUC value indicates how good it is in identifying whether a patient has breast cancer or not. To find AUC and ROC we need to know TPR and FPR. TPR is the ratio between True positive and the sum of True positive and False negative and FPR is the ratio between False positive and sum of true negative and False Positive [cite]. Equation 4.4 and 4.5 shows the formula of TPR and FPR [37].

$$TPR = \frac{True\,Positive}{True\,Positive + False\,Negative} \tag{4.4}$$

$$TPR = \frac{False\,Positive}{True\,Negative + False\,Positive} \tag{4.5}$$

4.7 Cross Validation

Cross validation or K fold cross validation is an statistical technique which is widely used in machine learning [19] [35] [29]. Cross validation is used to properly evaluate a model's performance. It is also used to prevent over fitting. It is sometimes used as an alternative to the traditional train,test, split because, in train, test, split we always keep substantial amount of data for training the machine and a small amount of data to test. There is a possibility that the test set selected is not representative of the whole data-set and that can cause our model to either over fit or perform poorly on the standards of different metrics.

Now as for functionality how cross validation functions is that it requires the user to set a value for k then what it does is that it divides the data set into k parts and uses one part for testing and k-1 parts for training. For example: in our work we used 5 fold cross validation so, our data would be divided into 5 portions and 4 portions of data would be used to train the algorithms and 1 portion would be used for testing. This way we can ensure that each portion is representative of the whole data set rather than being noise or anomaly.

4.8 Accuracy of the algorithms after applying on their respective reduced feature set

Now we will apply each algorithm on the reduced feature set we prepared in the previous chapter by using those algorithms with RFECV. As we mentioned before we have imbalanced class in our data set so, in the next section we will apply oversampling technique and will again evaluate the algorithms with over sampled data set. One thing to mention here is that, hyper-parameter tuning was performed on Random Forest, XGBoost and both Linear and Kernel SVM using GridSearchCV [47] [15].

4.8.1 Random Forest

From table 4.3 we can see that we got an accuracy of 69.935% and the number of true positive predictions are 82, true negative predictions are 25, false positive predictions are 14 and false negatives are 32. Our system predicted 57 participants would not participate in the voting process and 96 would out of total 153 test samples. We got a class precision of 0.64 for participants who would not vote and a class precision of 0.72 for the opposite class. As for the recall we got 0.44 for negative class and 0.85 for positive class and lastly for F1-score we got 0.52 for negative class and 0.78 for positive class. So, we can see that we have moderate amount of false positives which

is 14 out of 153 training sample. Another observation can be made from the result is that due to the imbalance of classes we can see that the F1-score of negative class is substantially less than positive class which we will try to counter later on in this chapter.

Accuracy: 69.935%			
	Actual No	Actual Yes	Precision
Predicted No	25	32	0.64
Predicted Yes	14	82	0.72
Recall	0.44	0.85	
F1-score	0.52	0.78	

Table 4.3: Accuracy, Precision, Recall and F1-score of Random Forest on it's decreased feature set

4.8.2 Extreme Gradient Boosting

From table 4.4 we can see that we got an accuracy of 67.32% and the number of true positive predictions are 75, true negative predictions are 26, false positive predictions are 12 and number of false negatives are 40. Our system predicted 66 participants would not participate in the voting process and 87 would out of total 153 test samples. We got a class precision of 0.68 for participants who would not vote and a class precision of 0.65 for the opposite class. As for the recall we got 0.39 for negative class and 0.86 for positive class and lastly for F1-score we got 0.50 for negative class and 0.74 for positive class. So, we can see that we have moderate amount of false positives which is 12 out of 153 training sample. Another observation can be made from the result is that due to the imbalance of classes we can see that the F1-score of negative class is substantially less than positive class which we will try to counter later on in this chapter.

Accuracy: 67.32%			
	Actual No	Actual Yes	Precision
Predicted No	26	40	0.68
Predicted Yes	12	75	0.65
Recall	0.39	0.86	
F1-score	0.50	0.74	

Table 4.4: Accuracy, Precision, Recall and F1-score of XGBoost on it's decreased feature set

4.8.3 Linear SVM

From table 4.5 we can see that we got an accuracy of 73.856% and the number of true positive predictions are 89, true negative predictions are 24, false positive predictions are 15 and false negative predictions are 25. Our system predicted 49 participants would not participate in the voting process and 104 would out of total 153 test samples. We got a class precision of 0.62 for the participants who would not vote and a class precision of 0.78 for the opposite class. As for the recall we got 0.49 for negative class and 0.86 for positive class and lastly for F1-score we got 0.55 for negative class and 0.82 for positive class. So, we can see that we have moderate amount of false positives which is 15 out of 153 training sample. Another observation can be made from the result is that due to the imbalance of classes we can see that the F1-score of negative class is substantially less than positive class which we will try to counter later on in this chapter.

Accuracy: 73.856%			
	Actual No	Actual Yes	Precision
Predicted No	24	25	0.62
Predicted Yes	15	89	0.78
Recall	0.49	0.86	
F1-score	0.55	0.82	

Table 4.5: Accuracy, Precision, Recall and F1-score of Linear SVM on it's decreased feature set

4.8.4 Kernel SVM

We also applied Kernel SVM specifically Gaussian Kernel also known as RBF kernel SVM. From table 4.6 we can see that we got an accuracy of 68.627% and the number of true positive predictions are 75, true negative predictions are 30, false positive predictions are 18 and false negative predictions are 30. Our system predicted 60 participants would not participate in the voting process and 93 would out of total 153 test samples. We got a class precision of 0.62 for the participants who would not vote and a class precision of 0.71 for the opposite class. As for the recall we got 0.50 for the negative class and 0.81 for positive class and lastly for F1-score we got 0.56 for negative class and 0.76 for positive class. So, we can see that we have moderate amount of false positives which is 18 out of 153 training sample. Another observation can be made from the result is that due to the imbalance of classes we can see that the F1-score of negative class is substantially less than positive class which we will try to counter later on in this chapter.

4.8.5 Naive Bayes

We also applied Naive Bayes specifically Bernoulli Naive Bayes. From table 4.7 we can see that we got an accuracy of 64.7% and the number of true positive

Accuracy: 68.627%			
	Actual No	Actual Yes	Precision
Predicted No	30	30	0.62
Predicted Yes	18	75	0.71
Recall	0.50	0.81	
F1-score	0.56	0.76	

Table 4.6: Accuracy, Precision, Recall and F1-score of Kernel SVM on it's decreased feature set

predictions are 70, true negative predictions are 29, false positive predictions are 20 and false negative predictions are 34. Our system predicted 63 participants would not participate in the voting process and 90 would out of total 153 test samples. We got a class precision of 0.59 for participants who would not vote and a class precision of 0.67 for the opposite class. As for the recall we got 0.78 for positive class and 0.46 for the positive class and lastly for F1-score we got 0.52 for the negative class and 0.72 for the positive class. So, we can see that we have moderate amount of false positives which is 20 out of 153 training sample. Another observation can be made from the result is that due to the imbalance of classes we can see that the F1-score of negative class is substantially less than positive class which we will try to counter later on in this chapter. Another thing to point is that we applied Naive Bayes on the whole feature set instead of a reduced feature set.

Accuracy: 64.7%			
	Actual No	Actual Yes	Precision
Predicted No	29	34	0.59
Predicted Yes	20	70	0.67
Recall	0.46	0.78	
F1-score	0.52	0.72	

Table 4.7: Accuracy, Precision, Recall and F1-score of Naive Bayes

4.9 Comparing results of the Algorithms

We will now compare accuracy, precision, recall, AUC value and F1-score of the respective algorithms to determine which algorithm generalized the problem better as one metric can be deceiving to compare the algorithms so, we used multiple metrics. Figure 4.4 shows the accuracy, AUC value and F1-score of each algorithm. The F1-score and AUC value has been multiplied by 100 to easily represent alongside accuracy score.

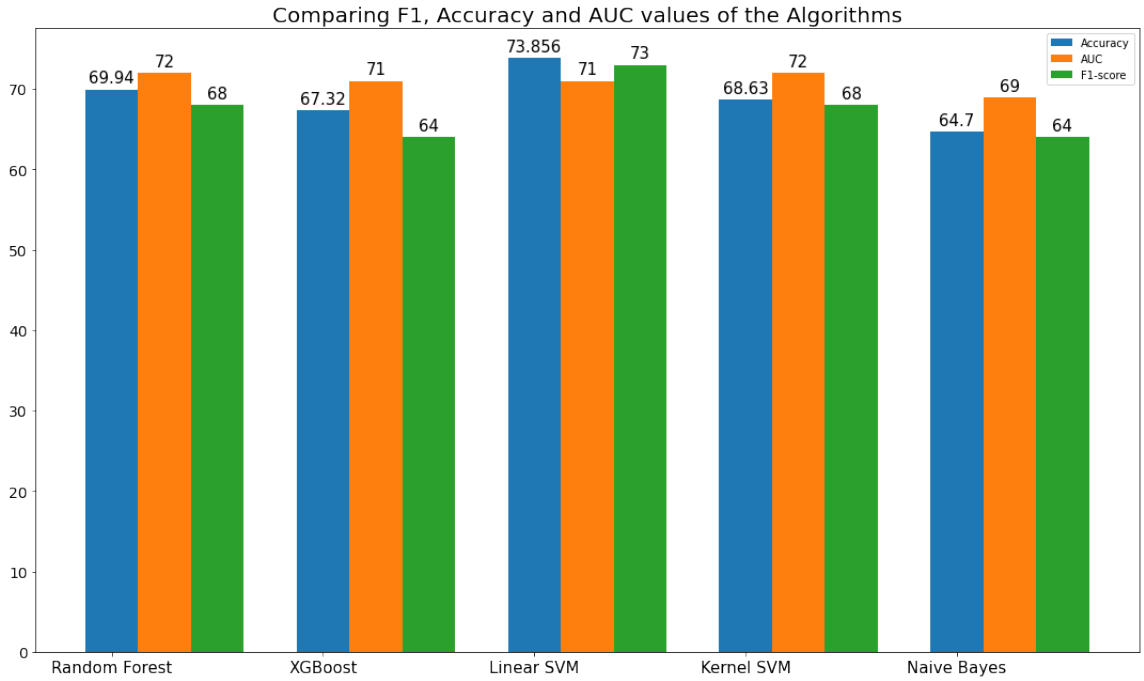


Figure 4.4: Comparing F1, Accuracy and AUC values of the Algorithms

From figure 4.4 we can see that except Linear SVM and Naive Bayes other 3 algorithms have pretty much similar accuracy score. Linear SVM has an accuracy score of 73.856% and on the other hand Naive Bayes has an accuracy score of 64.7%. From this standpoint it might be apparent that Linear SVM is performing better than other algorithms but, we have to remember that this data set contains class imbalance so, it is possible that the accuracy of Linear SVM is affected by the bias of dominating class but, we will analyze it further later when we apply over sampling. On the other hand the lowest accuracy is recorded by Naive Bayes which 64.7% so, it is apparent that Naive Bayes is performing worse compared to other algorithms. We can see that Random Forest has got an accuracy score of 69.935%, XGBoost has got an accuracy score of 67.32% and Kernel SVM has got an accuracy score of 68.627%.

As we are done with accuracy now we will discuss about the AUC scores of the different algorithms. As discussed before AUC score helps determine how well a classifier can differentiate between two classes when performing classification jobs. If we take a look at the AUC scores of the algorithms used we can see that Random Forest has a AUC value of 0.72 and Kernel SVM also has a AUC value of 0.72. On the other hand XGBoost has a AUC value of 0.71. Linaer SVM has a AUC value of 0.71. Naive Bayes has a AUC value of 0.69. From the AUC values we can say that Naive Bayes performed worse than other classifiers. Random Forest and Kernel SVM distinguished between the two output classes better than other algorithms comparatively.

Now we will discuss about the F1-scores of different algorithms. F1-score is a very strong and important alternative to accuracy as accuracy can seldom mislead us when F1-score can give us a clear picture about what is really going on. It will be

more apparent when we will apply over sampling to our data set. As for now Linear SVM has the highest F1-score of 0.73 compared to other algorithms. After Linear SVM highest F1-score is achieved by both Random Forest and Kernel SVM. The value of both their F1-score is 0.68 and after them XGBoost and Naive Bayes has the same F1-score of 0.64.

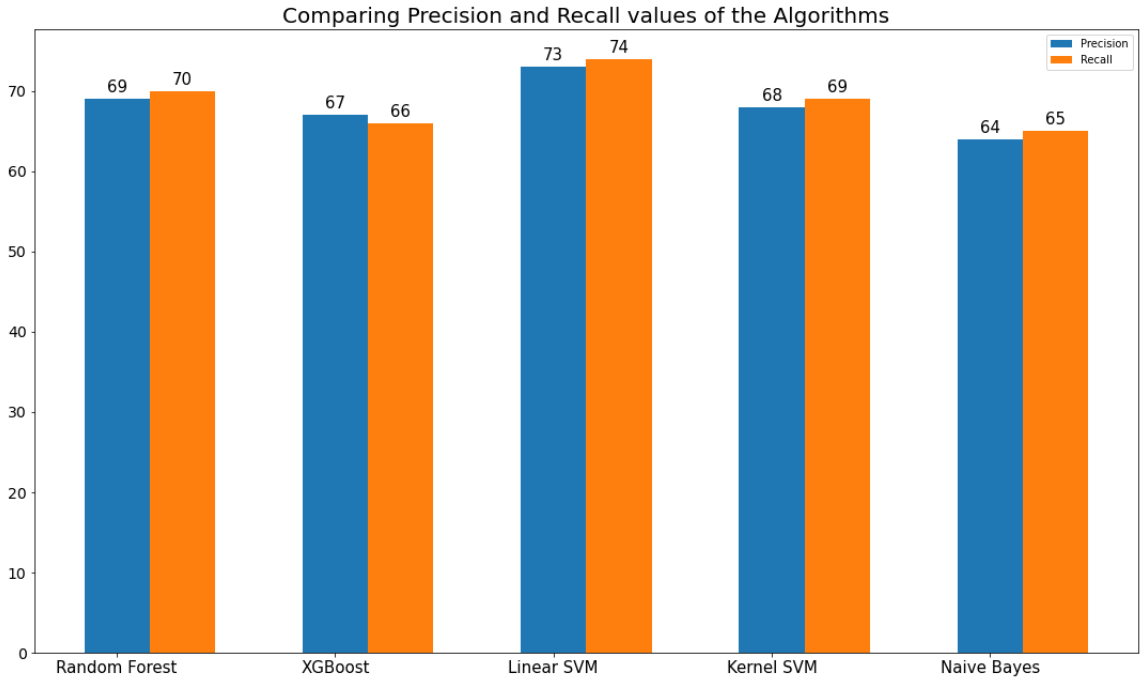


Figure 4.5: Comparing Precision and Recall values of the Algorithms

The precision and recall values on figure 4.5 has been multiplied by 100 for easier representation.
In figure 4.5 we can see the precision value of each algorithm. If we compare other metrics for example precision we can see that again Linear SVM has the highest value of precision. It has a precision value of 0.73 and the lowest value of precision is recorded by Naive Bayes which is 0.64. Random Forest has a moderately high value of precision which is 0.69. XGBoost has a precision value of 0.67 which somewhat close to Random Forest. Kernel SVM has a precision value of 0.68 which is between Random Forest and XGBoost.

In figure 4.5 we can see the recall value of each algorithm. Now as for recall the highest recall is recorded by Linear SVM and the value is 0.74. Random Forest has the highest recall value after Linear SVM which is 0.70. XGBoost has a moderately low recall value of 0.66. Kernel SVM has a recall value of 0.69 which is just a little bit less than Random Forest. The lowest value of recall is recorded by Naive Bayes which is 0.65. As mentioned before, recall is not that important to the nature of our study compared to precision. That is why we will focus less on recall compared to other performance metric.

Now if we compare how many positive labels and negative labels were predicted correctly by each algorithm then at first we will start with Random Forest. Random

Forest has predicted 82 positive labels correctly and 25 negative labels correctly. Next we will take a look at XGBoost. XGBoost predicted 75 positive labels correctly and 26 negative labels correctly. The amount of positive labels correctly is less than Random Forest. As for Linear SVM the number of correctly predicted are 89 and number of correctly predicted negative labels are 24. Here we can see that Linear SVM got more positive labels correct than Random Forest and almost as many negative labels correct as Random Forest so, in a way Linaer SVM performed a little better than Random Forest. Kernel SVM predicted 75 positive labels correctly and 30 negative labels correctly. So far, Kernel SVM got the highest amount of negative labels correctly. Naive Bayes predicted 70 positive labels correctly and 29 negative labels correctly.

4.10 Accuracy of the algorithms after applying on their over sampled Data set

Previously we have applied each algorithm on their respective reduced feature sets with only exception being Naive Bayes which was applied on the whole feature set. It was mentioned before and we will briefly discuss it again is that our data set contain class imbalance. Here what we mean by class imbalance is that: the amount of people willing to vote is far greater than the amount of people who are not willing to vote. Due to class imbalance some of the algorithm failed to perform to their full potential and give desired result and also it can be be observed from the results of the algorithms when applied on the original data set is that the amount of correctly predicted negative labels were less than the amount of the amount of wrongly predicted negative labels or also known as false negatives. It is because, the amount of positive labels are higher than negative labels. Due to this the algorithms were able to predict positive labels correctly but, failed to replicate the same performance when it came to negative labels. That is why we used a state of the art oversampling technique called SMOTE. This way we over sampled the negative class data points to counter class imbalance. Now we will apply the previously applied algorithms on these new over sampled data-set and try to analyze and compare the results.

4.10.1 Random Forest

From table 4.8 we can see that after applying SMOTE on the data set the confusion matrix is quite different from the previous iteration when we applied Random Forest on the default data set. This time the number of correctly predicted positive labels are 72, correctly labeled negative labels are 31. As for false positive labels there are 28 and number of false negative labels there are 22. As for precision we have a negative class precision of 0.53, positive class precision of 0.77. On the other hand we have a negative class recall of 0.58 and positive class recall of 0.72. Lastly F1-score: the negative class have a F1-score of 0.55 and positive class have a F1-score of 0.74. If we look closely and compare the results from our previous iteration of Random Forest we can clearly see that not only the number of correctly predicted negative points increased but also, the precision for the negative class increased and most importantly the F1-score of the negative class increased. We will discuss this

more later on in this chapter.

Accuracy: 67.32%			
	Actual No	Actual Yes	Precision
Predicted No	31	22	0.53
Predicted Yes	28	72	0.77
Recall	0.58	0.72	
F1-score	0.55	0.74	

Table 4.8: Accuracy, Precision, Recall and F1-score of Random Forest on over sampled data set

4.10.2 Extreme Gradient Boosting

From table 4.9 we can see that the performance of XGBoost has increased substantially over it's previous iteration. At first if we take a look at the accuracy it has increased from 67.32% to 69.935%. Next the number of correctly labeled positive data points are 67 and number of correctly labeled negative data points are 40. As for the false positive and false negative there are 26 false positive data points and 20 false negative data points. The precision score has also increased now the negative class boosts a precision score of 0.61 and positive class has a precision score of 0.77, As for recall positive class has a recall of 0.72 and negative class has a recall of 0.67. Now, we will take a look at the F1-score of the classes. Positive class has a F1-score of 0.74 and negative class has a F1-score of 0.63. If we compare the F1-score to the previous iteration's score we can easily notice that the F1-score of the negative class has increased from 0.50 to 0.63 which is a noticeable improvement. From comparing the two iterations of XGBoost we can say that by applying SMOTE we have improved the algorithms performance on the negative class substantially.

Accuracy: 69.935 %			
	Actual No	Actual Yes	Precision
Predicted No	40	20	0.61
Predicted Yes	26	67	0.77
Recall	0.67	0.72	
F1-score	0.63	0.74	

Table 4.9: Accuracy, Precision, Recall and F1-score of XGBoost on over sampled data set

4.10.3 Linear SVM

From table 4.10 we can see the outputs of Linear SVM. As for Liner SVM we can see that as opposed to the other two algorithm we have observed so far in this section the performance of SVM has supposedly decreased. At first the most noticeable change we can see is that the drastic change of accuracy, it has gone down from 73.856% to 64.706%. Now we will take a look at the amount of correctly labeled positive data points and correctly labeled negatively data points. The number of correctly labeled positive data points are 64 and correctly labeled negative data points are 35. Here we can see that the amount of correctly predicted positive labels have decreased but, on the other hand correctly predicted negative labels have increased. Now if we take a look at false positives and false negatives we can see that we have, 33 false positives and 21 false negatives. As for precision the negative class has a precision of 0.51 and positive class has a precision of 0.75 which also has decreased compared to it's previous iteration. Now we will take a look at the recall and F1-score. The negative class has a recall of 0.62 and the positive has a recall of 0.66. The positive class has F1-score of 0.70 and the negative class has a F1-score of 0.56. We can see that the F1-scores have also decreased,

Accuracy: 64.706 %			
	Actual No	Actual Yes	Precision
Predicted No	35	21	0.51
Predicted Yes	33	64	0.75
Recall	0.62	0.66	
F1-score	0.56	0.70	

Table 4.10: Accuracy, Precision, Recall and F1-score of Linear SVM on over sampled data set

4.10.4 Kernel SVM

From table 4.11 we can see the outputs of Kernel SVM. We will take a look at the performance of Kernel SVM. After applying SMOTE it now gives a accuracy of 69.935% which is an improvement over it's previous iteration. It predicted 73 data points correctly as positive and 34 data points correctly as negative. We have 24 false positives and 22 false negatives. If we compare this to the it's previous iteration we can see that the amount of correctly predicted negative data points have increased. Now if we take a look at the precision we have 0.59 for negative class and 0.77 for positive class. Next, if we take a look at the recall we have 0.61 for negative class and 0.75 for positive class. As for F1-score we have 0.76 for positive class and 0.60 for negative class. If we compare the F1-scores to it's previous iteration we can see that although the F1-score of the positive class is same as it was in the previous iteration but, the F1-score of the negative class has increased, previously it was 0.56 but now as mentioned before it has increased to 0.60. This indicates that the algorithm is now predicting negative class points better than it's previous iteration.

Accuracy: 69.935 %			
	Actual No	Actual Yes	Precision
Predicted No	34	22	0.59
Predicted Yes	24	73	0.77
Recall	0.61	0.75	
F1-score	0.60	0.76	

Table 4.11: Accuracy, Precision, Recall and F1-score of Kernel SVM on over sampled data set

4.10.5 Naive Bayes

From table 4.12 we can see the outputs of Naive Bayes.We also applied Naive Bayes on the over sampled data set. One thing to remember is that we did not apply any feature selection on the feature set that was used as input for Naive Bayes, we simply over sampled the feature set to counter feature imbalance. If we take a look at the accuracy is 63.4% which is a little less than what we got in the previous iteration which was 64.7%. The amount of correctly predicted negative points are 35 and the amount of correctly predicted positive points are 62. The amount of correctly predicted negative data points have increased which was 29 in the previous iteration. The amount of correctly predicted positive data points have decreased slightly. As for the amount of false positive and false negative we have 31 false positive and 25 false negative data points. If we take a look at the precision score we have 0.53 for negative class and 0.71 for positive class. The precision score for positive class has increased. As for recall the positive class has a recall of 0.67 and the negative class has a recall of 0.58. Also for the F1-score positive class has a F1-score of 0.69 and the negative class has a F1-score of 0.56.If we compare the F1-score of the negative class with the previous iteration of the algorithm when it was applied with a class imbalanced data set we can see that the F1-score of the negative class increased. So, we can observe that as the labels are balanced now so, the algorithm can correctly predict more negative points compared to it's previous iteration.

Accuracy: 63.4 %			
	Actual No	Actual Yes	Precision
Predicted No	35	25	0.53
Predicted Yes	31	62	0.71
Recall	0.58	0.67	
F1-score	0.56	0.69	

Table 4.12: Accuracy, Precision, Recall and F1-score of Naive Bayes on over sampled data set

4.11 Comparing results of the Algorithms after applying Over Sampling

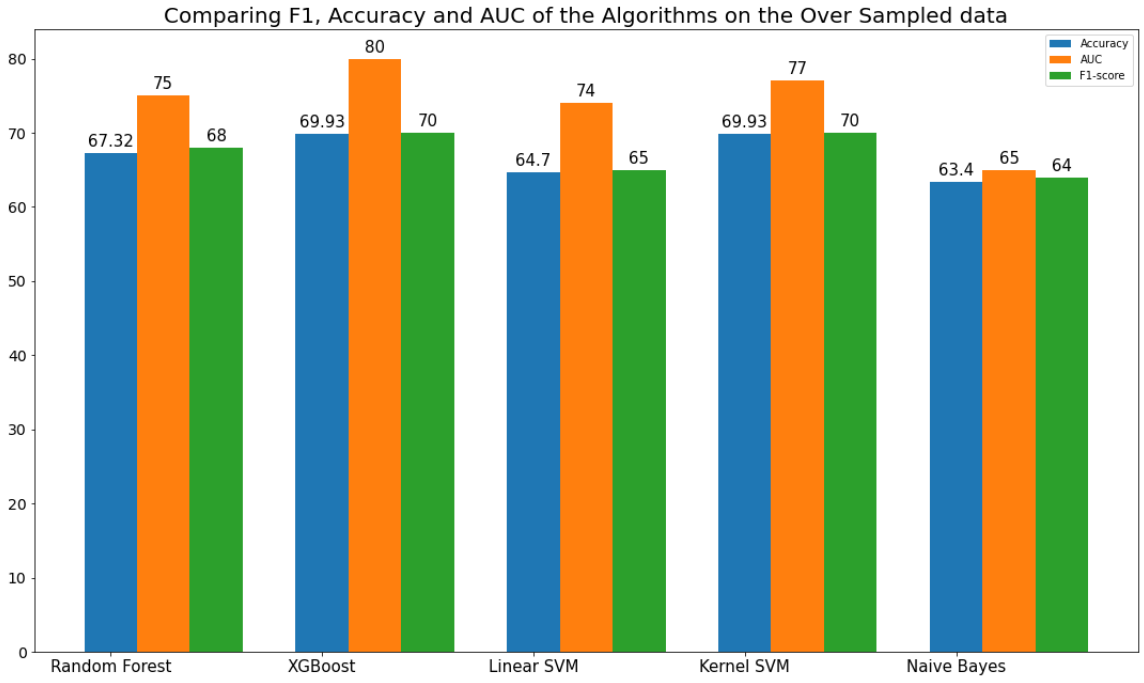


Figure 4.6: Comparing F1, Accuracy and AUC of the Algorithms on the Over Sampled data

Now we will compare each algorithms output after applying them on the over sampled data set. Just like before we will use different metrics like: accuracy, precision, recall, AUC value and F1-score rather than relying on a single metric to ensure we properly evaluate the algorithms. Figure 4.6 shows bar-chart where each bar represents an algorithm's accuracy, AUC value and F1-score. The F1-score and AUC value has been multiplied by 100 to easily represent alongside accuracy score.

At first we will analyze and discuss the accuracy value of the different algorithms. We can see that across the board the highest accuracy is achieved by XGBoost and Kernel SVM which is 69.93% for both of them. After XGBoost and Kernel SVM we have Random Forest which achieved an accuracy of 67.32%. After Random Forest we have Linear SVM and Naive Bayes. Linear SVM has an accuracy score of 64.7% and Naive Bayes has an accuracy score of 63.4%. If we compare each algorithm to their previous counterpart which is when they were applied using the default data set we can see that the accuracy of Random Forest has decreased slightly, on the other hand accuracy of XGBoost has increased from 67.32% to 69.93%. As for Linaer SVM it's accuracy has decreased drastically from 73% to 64.7%. Accuracy of Kernel SVM has increased from 68.627% to 69.93%. Naive Bayes previously had an accuracy score of 64.7% which has decreased slightly to 63.4%.

Now we will take a look at F1-score. Highest value of F1-score was achieved by XGBoost and Kernel SVM which is 0.70 for both of them. Random Forest has a F1-score of 0.68. Linear SVM has a F1-score of 0.65 and Naive Bayes has a F1-score

of 0.64. If we compare the F1-score with when the algorithms were applied with the default data set we can see that F1-score of Random Forest is same as before. As for XGBoost we can see that it's F1-score has increased from 0.64 to 0.70 which is a noticeable increase. On the other hand for F1-score of Linear SVM it has decreased substantially, it was 0.73 before but now it is 0.5. The F1-score of Kernel SVM has increase slightly it was 0.68 previously now it is 0.70. The F1-score of Naive Bayes remained unchanged.

As we are done with Accuracy and F1-score now we will take a look at AUC values. Highest Value of AUC is recorded by XGBoost and Kernel SVM. XGBoost has an AUC value of 0.80 and Kernel SVM has a AUC value of 0.77. Both of them had a increase in their AUC value as XGBoost had an AUC value of 0.71 and Kernel SVM had an AUC value of 0.72 on the default data set. Random Forest has an AUC value of 0.75 which is an improvement compared to it's previous AUC value on the default data set which was 0.72. Linear SVM has an AUC value of 0.74 which is a subtle increase compared to it's AUC value on the default data set. Finally, Naive Bayes got an AUC value of 0.65 which is a decrease compared to it's AUC value on the default data set.

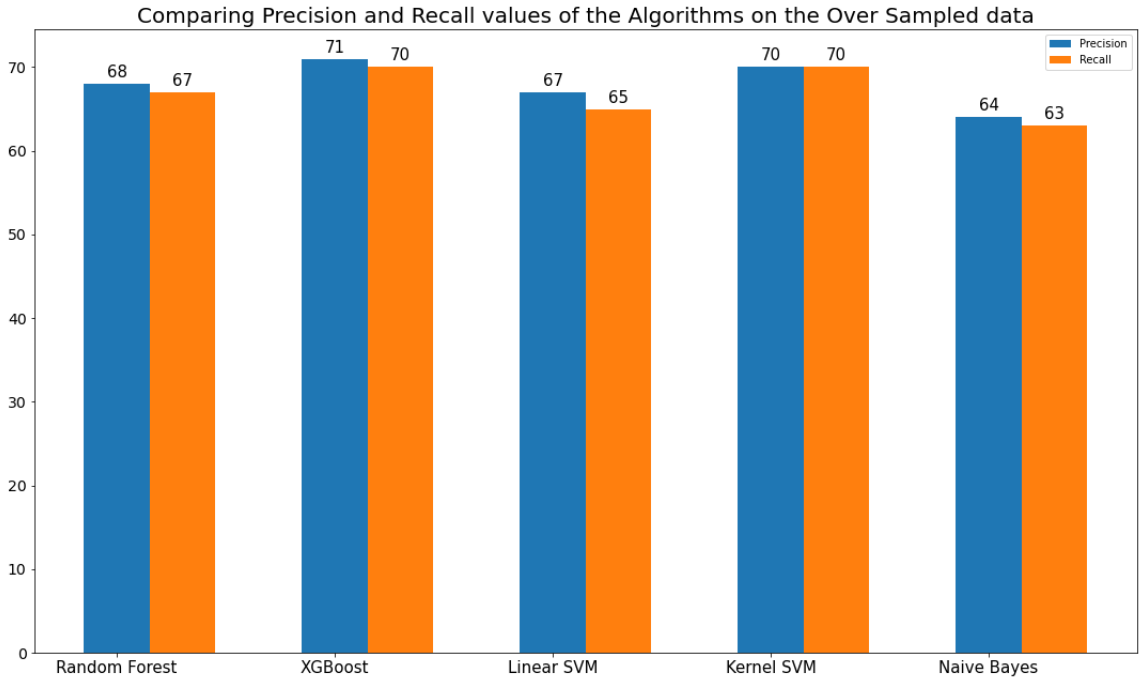


Figure 4.7: Comparing Precision and Recall values of the Algorithms on the Over Sampled data

The precision and recall values on figure 4.7 has been multiplied by 100 for easier representation.
Now we will take a look the precision value of each algorithm to determine which algorithm performed better. XGBoost achieved an precision value of 0.71 which is highest across the board. After XGBoost the highest value of precision was achieved by Kernel SVM which is 0.70. Linear SVM and Random Forest achieved a somewhat close value which are 0.67 and 0.68 for Linear SVM and Random Forest respectively.

Lowest value of precision was recorded by Naive Bayes which is 0.64. From these observation we can say that XGBoost and Kernel SVM performed better than other algorithm and Naive Bayes performed the worst at least according to the precision value.

As we are done with accuracy and precision now all that is left is recall and F1-score. We will take a look at recall first then we will move on to F1-score. As for recall the highest value was recorded by XGBoost and Kernel SVM which is 0.70 for both of them and after that Random Forest comes in with a value of 0.67. Linear SVM has a recall value of 0.65 and Naive Bayes has a recall value of 0.63.

Now if we take a look number of correctly predicted negative and positive data points Random Forest predicted 72 positive labels correctly and 31 negative labels correctly. XGBoost predicted 67 positive labels correctly and 40 negative labels correctly. XGBoost predicted correct positive labels a little less than Random Forest but substantially more negative labels correctly compared to Random Forest. As for Linear SVM it predicted 64 positive labels correctly and 35 negative labels correctly which are somewhat moderate. Kernel SVM predicted 73 positive labels correctly and 34 negative labels correctly which is very close to what Random Forest's performance. Naive Bayes predicted 62 positive labels correctly and 35 negative labels correctly which is less than all other algorithms comparatively.

4.12 Default data set vs. Over Sampled Data set

Now we will compare results of various metrics we used to measure the performance of the algorithm both on the default data set and the over sampled data set.

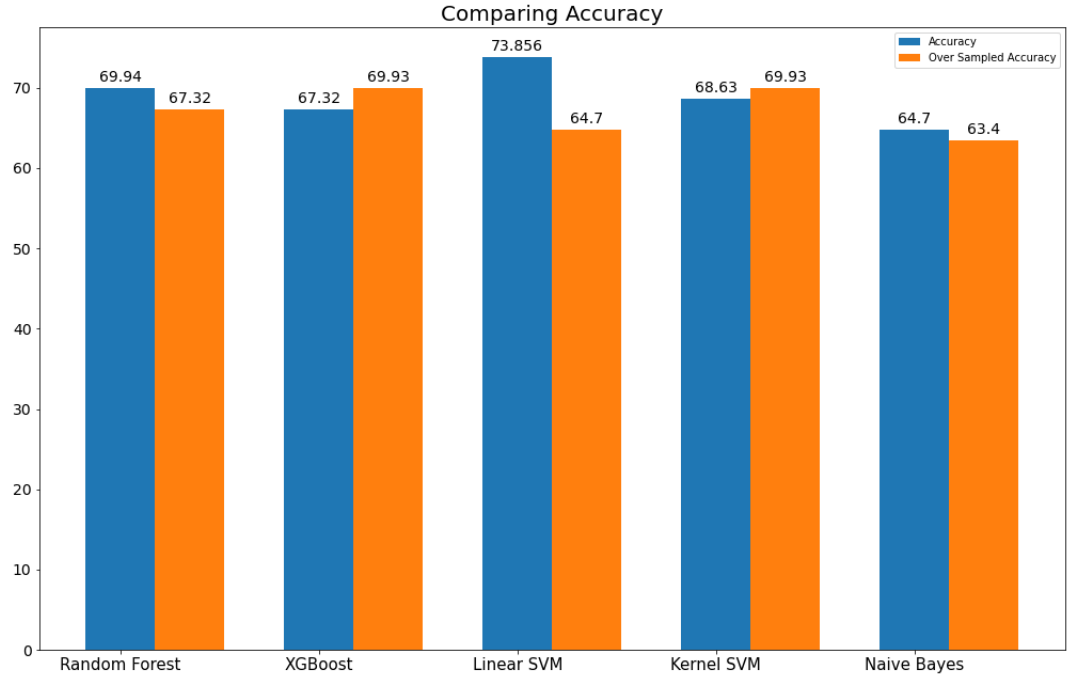


Figure 4.8: Comparing accuracy between default and Over Sampled Data set

From figure 4.8 we can see the comparison of the accuracy scores of all the algorithms before and after over sampling. At first if we take a look at the accuracy score of Random Forest we can see that it decreased slightly after applying over sampling compared to when it was applied on the default data set. Random Forest had an accuracy score of 69.935% and after applying over sampling it got an accuracy score of 67.32%. Now as for XGBoost we can see that it's accuracy score increased after applying it on the over sampled data set compared to when it was applied on the default data set. XGBoost achieved an accuracy score of 67.32% on the default data set and 69.93% on the over sampled data set. On the other hand Linear SVM's accuracy score decreased drastically after applying it on the over sampled data set compared to when it was applied on the default data set. Linear SVM got an accuracy score of 73.856% on the default data set and 64.7% on the over sampled data set. Kernel SVM's accuracy increased slightly. It was previously 68.627 on the default data set and now it is 69.93% on the over sampled data set. Naive Bayes's accuracy decreased from 64.7% to 63.4%.

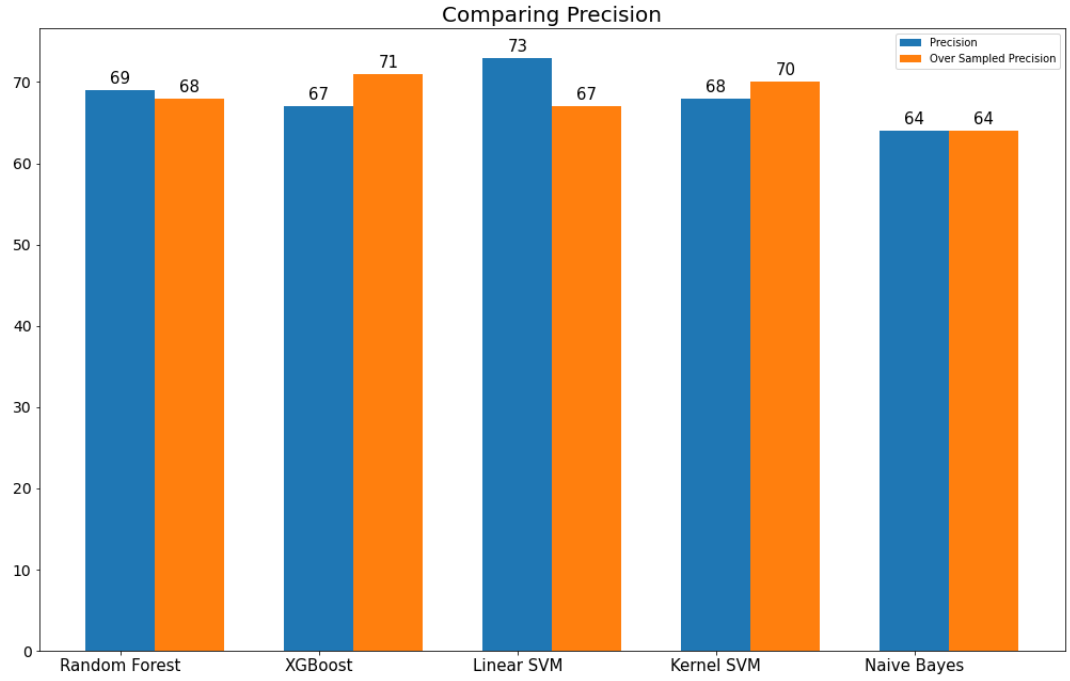


Figure 4.9: Comparing precision between default and Over Sampled Data set

Figure 4.9 shows an comparison of precision values of each algorithm when applied on the default data set and the over sampled data set. Precision values are multiplied by 100 for better representation in figure 4.9. Random Forest had a precision value of 0.69 on the default data set and it got a little less on the over sampled data set which is 0.68. XGBoost's precision increased dramatically after applying over sampling compared to it's precision value on the default data set. XGBoost got a precision value of 0.67 on the default data set and 0.71 on the over sampled data set. Linear SVM saw a substantial decrease in it's precision value after being applied on the over sampled data set. Linear SVM had a precision value of 73 on the default data set and it got a precision value of 0.67 on the over sampled data set. Kernel SVM's precision value also increased after applying over sampling. It got a precision

value of 0.70 compared to it's precision value on default data set which was 0.68. Naive Baye's precision value remained the same after applying over sampling. In both the cases Naive Bayes had a precision value of 0.64.

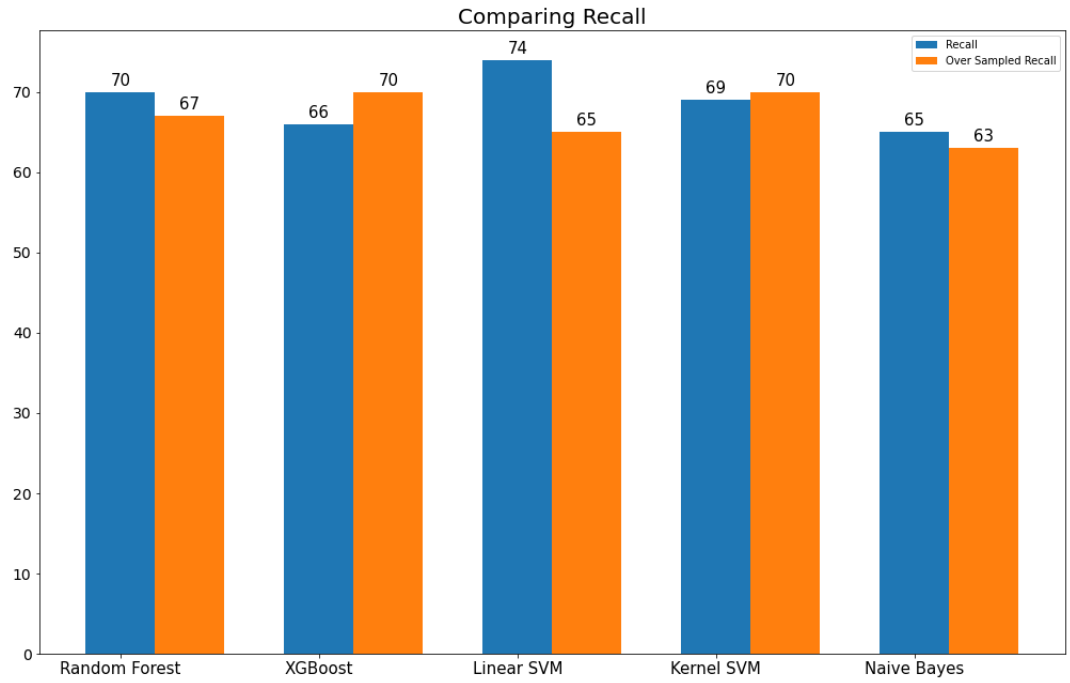


Figure 4.10: Comparing recall between default and Over Sampled Data set

In figure 4.10 we can take a look at the comparison of recall values of respective algorithms before and after applying on the over sampled data set. Recall values are multiplied by 100 for better representation in figure 4.10. At first we can see that the recall value of Random Forest decreased when it was applied using the over sampled data set compared to when it was applied on the default data set. It achieved a recall score of 0.70 when it was applied using the default data set and it achieved a recall value of 0.67 when it was applied on the over sampled data set. Now as for XGBoost we can see that there is an increase in the value of recall when it is applied on the over sampled data set compared to when it was applied using the default data set. On the default data set XGBoost has a recall score of 0.66 and on the over sampled data set XGBoost has a recall score of 0.70. Now if we take a look at the recall score of Linear SVM we can see that it's recall score decreased after applying on the over sampled data set. It got an recall score of 0.65 on the over sampled data set and 0.74 on the default data set. Recall score increased slightly for Kernel SVM after applying on the over sampled data set compared to default data set. Kernel SVM got a recall score of 0.69 on the default data set and 0.70 on the over sampled data set. On the other hand recall score decreased slightly for Naive Bayes when applied on the over sampled data set. Naive Bayes had a recall score of 0.65 on the default data set and it got a recall score of 0.63 on the over sampled data set.

If we take a look at figure 4.11 we can see a comparison between the F1-scores of the each algorithm applied on the default data set and over sampled data set. F1-scores

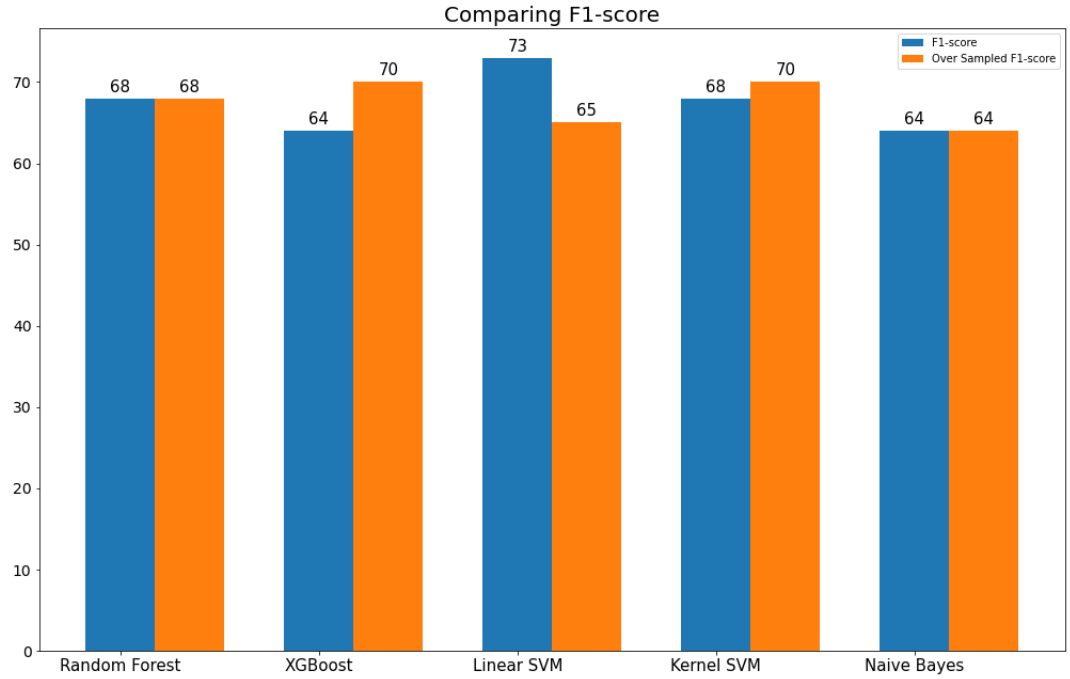


Figure 4.11: Comparing F1-score between default and Over Sampled Data set

are multiplied by 100 for better representation in figure 4.11. At first we will take a look at Random Forest. We can see that F1-score of Random Forest is unchanged compared to when it was applied on the default data set and over sampled data set. Random Forest had a F1-score of 0.68 on both cases. As for XGBoost we can see that it has gained a noticeable boost in F1-score after being applied on the over sampled data set compared to when it was applied on the default data set. XGBoost had a F1-score of 0.64 on the default data set and a F1-score of 0.70 on the over sampled data set. On the other hand Linear SVM's F1-score decreased substantially after applying on the over sampled data set. Previously it was 0.73 now it is 0.65. Kernel SVM saw a slight increase in it's F1-score after being applied on the over sampled data set. Kernel SVM achieved a F1-score of 0.68 on the default data set and 0.70 on the over sampled data set. Naive Bayes had the same F1-score it had on both the default data set and the over sampled data set. The value of Naive Bayes's F1-score is 0.64 on both case.

If we take a look at figure 4.12 we can see the AUC values of each algorithm before and after applying over sampling. In figure 4.12 the AUC values were multiplied by 100 for better representation. Random Forest had an AUC value of 0.72 on the default data set and it got an AUC value of 0.75 on the over sampled data set. XGBoost had an AUC value of 0.71 on the default data set and it got an AUC value of 0.80 on the over sampled data set. Linear SVM had an AUC value of 0.71 on the default data set and it got an AUC value of 0.74 on the over sampled data set. Kernel SVM got an AUC value of 0.72 on the default data set and it got an AUC value of 0.77 on the over sampled data set. Naive Bayes had an AUC value of 0.69 on the default data set and it got an AUC value of 0.65 on the over sampled data set. From the comparison we can see that XGBoost and Kernel SVM achieved better AUC values after over sampling that is why they performed better than other

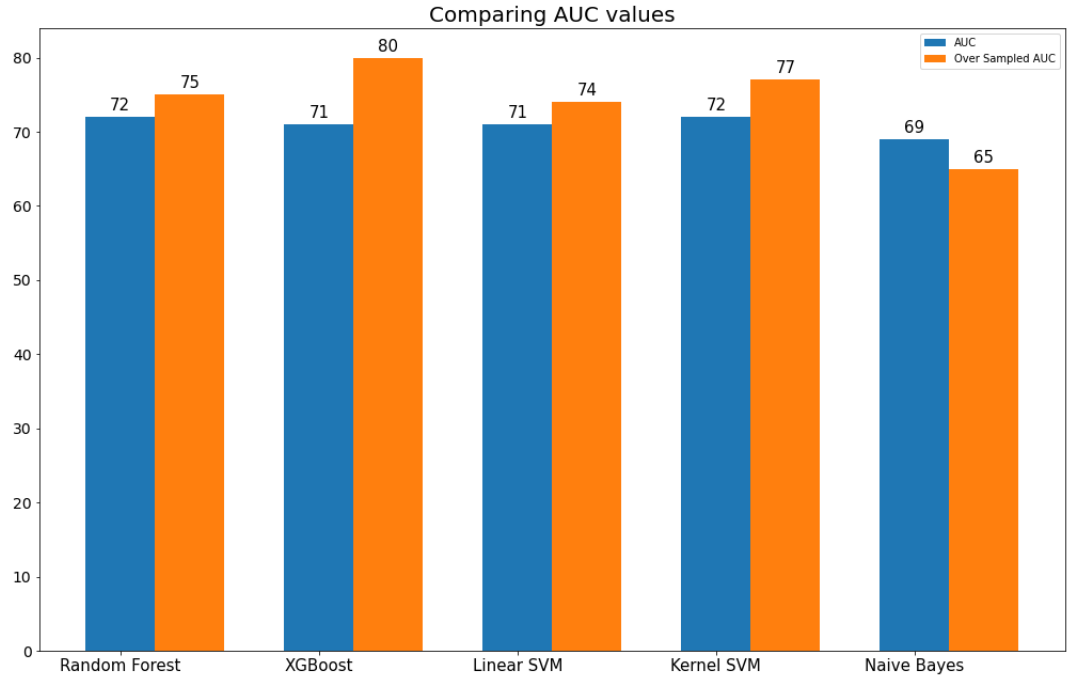


Figure 4.12: Comparing AUC values between default and Over Sampled Data set

classifiers. All of the classifier saw an increase in their AUC value except Naive Bayes after over sampling. Naive Bayes's AUC value decreased after being applied on the over sampled data set as it was previously 0.69 and after over sampling it became 0.65.

4.13 Final Verdict on the result analysis

By analyzing all the metrics for each algorithm before and after applying over sampling we saw that XGBoost and Kernel SVM outperformed all the other algorithms. In this section we will discuss the behaviour of the algorithms. Firstly, we will start with Linear SVM. Linear SVM gave a very high accuracy and performed well on the default data set but gave rather low accuracy and performed poorly on the over sampled data set. This happened because, we know that, Linear SVM works better on linearly separable data set [4] [38]. Before applying over sampling our data set had a lot of positive class instances but very less negative class instances. In other word it was somewhat linear. That is why Linear SVM outperformed other algorithms on the default data set but, after applying over sampling the data set was no more linear. That is why Linear SVM failed to replicate it's performance on the over sampled data set. Now, as for XGBoost we know that, at it's core it is a tree based algorithm and also uses boosting mechanism [21]. How XGBoost works is that, it creates a decision tree and then boosts it in different layer with each layer taking feedback from previous layer to rectify mistakes and increase performance [36]. When XGBoost was applied on the default data set it could not utilize it's full potential due to presence of class imbalance. After applying over sampling it performed way better compared to it's performance on the default data set. Kernel SVM also performed better after over sampling. Random Forest had a consistent performance on both the default data set and over sampled data set. As for Naive

Bayes it could not perform better on any of the iteration because, of how it works. Naive Bayes considers each feature independently and thus it is unable to capture any complex relation between the features [28] and as we already know that our data set is quite complex and the relations between the features are not simple at all. That is why XGBoost and Kernel SVM outperformed other algorithms and proved to be best for predicting whether someone will vote or not.

Chapter 5

Final Remark

This chapter restates the problem, summarizes our finding and discusses what can be done to further improve this work and what is our future plan with this work.

5.1 Conclusion

Voting is a very important duty of a citizen. It is necessary for a person to participate in the election process to make the democratic system effective. As mentioned before we know that a significant portion of the voter are new voters and for some reasons they are reluctant to vote. There are some social, cultural and psychological reasons embedded behind this reluctance to vote and with this research we tried to find out those reasons behind not participating in the voting process. The objective of this research was to find out the non-political factors behind young voter enthusiasm because, we noticed that among our peers or more generally people around us who are new voters or will be new voter are somewhat uninterested to vote. If a new voter is uninterested to vote then chances are when they grow up or become old they will never participate in the voting process, this way a country can never achieve perfect democracy. Not many work has been done on this topic all over the world and almost none in Bangladesh. Our system can be used by the government and the election commission to determine the factors that repel or motivate the young voters and the government can fine tune those factors to ensure maximum voter turnout and achieve perfect democracy. Due to shortage of man power and resources we did our study on the students of BRAC University only but we intend to extend our work and collect data from other people. In our system again as mentioned before, we used Random Forest, Linear SVM and Extreme Gradient Boosting along with RFECV to perform feature selection. We then unified those feature set to find the absolute feature set which consists all the features that are responsible for young voter enthusiasm. We then took each reduced feature set and applied a hyper-parameter tuned variant of the algorithm on it and tried to predict whether a person will vote or not. For example: we applied a hyper-parameter Random Forest on the feature set we generated using RFECV and Random Forest. We applied both Linear SVM and Kernel SVM on the feature set generated by Linear SVM and RFECV and we applied Naive Bayes on the whole data set. We also over sampled the feature sets and again applied the algorithms. We saw that out of the five algorithms used Extreme Gradient Boosting and Kernel SVM outperformed all the other algorithms. Another thing we noticed from doing this research is that even

after explicitly stating the fact that we were looking specifically for the non-political factors most of the time people confused it with politics. Lastly in our work most of the people who participated in our survey wanted to vote. The amount of people who wanted to participate in the voting process were way more than people who did not wanted to participate. Due to this situation we had to perform over sampling on the data set.

5.2 Regarding status collected as part of survey

As a part of our survey we collected status from our participants regarding their reflection about the voting system which is meant for a social media as we had plans to process those status and use them as an additional feature to help in our prediction process but, we noticed that an exceptionally large amount of those status were negative. This is why we could not use them. One interesting thing to notice is that even if someone was willing to vote but, their status would reflect otherwise. This is another example of the participants not taking the survey seriously. That is why we had to drop any plan we had to process those status and use them as feature.

5.3 Limitations

Some of the limitations of our work are as follows:

- We could not collect a lot of data
- There were a lot of noise in the data
- The amount of people who did not want to vote or less likely to vote were significantly less than people who wanted to vote
- We could only collect data from BRAC University due lack of resources

Due to lack of similar work done before and lack of pre-established data set we could not achieve better results. When performing survey we faced many problems such as: lack of interest of the participants and that is why many of them did not take their time to fill the survey properly. That is why we had a lot of noise in our data set. Another problem we faced was that due to lack of resources and misconception of people about our study we could not reach broader target audience. Our data set is also not that big compared to the amount of features in our data set that is why we could not get better result and we could not apply more powerful algorithms like Neural Network.

5.4 Future Work

Some of future work we have planned are as follows:

- Improve and enhance our survey
- Collect more data

- Add more features if necessary
- Apply more powerful learning algorithm such as: Neural Network
- Collect status from participants and perform sentiment analysis

As mentioned before due to many limitations we could not perform many of our planned tasks that we plan to do in the future. Most important of them are collecting more data and. As mentioned before our data set has a lot of noise. In our future work we will try to collect data more efficiently so that we get less noise also we want to collect more data so, that we can apply more powerful learning algorithm such as: Neural Network to get better result. Another thing we could not do is that we could not perform sentiment analysis on the status we collected from the survey participants due to most of them being negative. In our future work we will try to collect statuses more efficiently so that, we can perform sentiment analysis on them and enrich our system.

Bibliography

[1] R. E. Ward, "Urban-rural differences and the process of political modernization in japan: A case study", *Economic Development and Cultural Change*, vol. 9, no. 1, Part 2, pp. 135–165, 1960. DOI: 10.1086/449884. eprint: https://doi.org/10.1086/449884. [Online]. Available: https://doi.org/10.1086/449884.

[2] W. H. Riker and P. C. Ordeshook, "A theory of the calculus of voting", *American political science review*, vol. 62, no. 1, pp. 25–42, 1968.

[3] S. J. Rosenstone, "Economic adversity and voter turnout", *American Journal of Political Science*, vol. 26, no. 1, pp. 25–46, 1982, ISSN: 00925853, 15405907. [Online]. Available: http://www.jstor.org/stable/2110837.

[4] C. Cortes and V. Vapnik, "Support-vector networks", *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[5] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/a:1010933404324.

[6] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications", vol. 2049, Jan. 2001, pp. 249–257. DOI: 10.1007/3-540-44673-7_12.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique", *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[8] ——, "Smote: Synthetic minority over-sampling technique", *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[9] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs.", *Icwsm*, vol. 7, no. 21, pp. 219–222, 2007.

[10] V. Larcinese, "Does political knowledge increase turnout? evidence from the 1997 british general election", *Public Choice*, vol. 131, no. 3-4, pp. 387–411, Nov. 2007. DOI: 10.1007/s11127-006-9122-0.

[11] H. McIntosh, D. Hart, and J. Youniss, "The influence of family political discussion on youth civic development: Which parent qualities matter?", *PS: Political Science & Politics*, vol. 40, no. 3, pp. 495–499, 2007. DOI: 10.1017/S1049096507070758.

[12] M. G. Marshall and B. R. Cole, *Global report 2009: Conflict, governance, and state fragility*. Center for Systemic Peace, 2009.

[13] A. J. Berinsky and G. S. Lenz, "Education and political participation: Exploring the causal link", *Political Behavior*, vol. 33, no. 3, pp. 357–373, 2010. DOI: 10.1007/s11109-010-9134-9.

[14] G. I. Webb, "Naïve bayes", in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 713–714, ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_576. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_576.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] D. M. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation", 2011.

[17] D. Bangladesh *et al.*, *Bangladesh demographic and health survey 2014*, 2013.

[18] K. Holt, A. Shehata, J. Strömbäck, and E. Ljungberg, "Age and the effects of news media attention and social media use on political interest and participation: Do social media function as leveller?", *European Journal of Communication*, vol. 28, no. 1, pp. 19–34, 2013. DOI: 10.1177/0267323112465369. eprint: https://doi.org/10.1177/0267323112465369. [Online]. Available: https://doi.org/10.1177/0267323112465369.

[19] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation", *Pattern Recognition*, vol. 48, no. 9, pp. 2839–2846, 2015.

[20] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Opinion mining and sentiment polarity on twitter and correlation between events and sentiment", in *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, IEEE, 2016, pp. 52–57.

[21] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system", in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[22] A. GREENBLATT, *What would happen if america made voting mandatory?*, Feb. 2016. [Online]. Available: https://www.governing.com/topics/elections/gov-compulsory-voting-switzerland.html.

[23] P. Sharma and T.-S. Moh, "Prediction of indian election using sentiment analysis on hindi twitter", in *2016 IEEE International Conference on Big Data (Big Data)*, IEEE, 2016, pp. 1966–1971.

[24] U. C. Bureau, *Census bureau projects u.s. and world populations on new year's day*, Jul. 2017. [Online]. Available: https://www.census.gov/newsroom/press-releases/2016/cb16-tps158.html.

[25] ——, *Voting in america: A look at the 2016 presidential election*, May 2017. [Online]. Available: https://www.census.gov/newsroom/blogs/random-samplings/2017/05/voting_in_america.html.

[26] B. Joyce and J. Deng, "Sentiment analysis of tweets for the 2016 us presidential election", in *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, IEEE, 2017, pp. 1–4.

[27] J. M. Krogstad and M. H. Lopez, *Black voter turnout fell in 2016 us election*, May 2017. [Online]. Available: https://www.pewresearch.org/fact-tank/2017/05/12/black-voter-turnout-fell-in-2016-even-as-a-record-number-of-americans-cast-ballots/.

[28] R. Gandhi, *Naive bayes classifier*, May 2018. [Online]. Available: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.

[29] Shiv, *K-fold cross validation*, Jun. 2018. [Online]. Available: https://medium.com/@shivam.somani09/k-fold-cross-validation-cb85632dba3.

[30] N. I. Tripto and M. E. Ali, "Detecting multilabel sentiment and emotions from bangla youtube comments", in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, IEEE, 2018, pp. 1–6.

[31] S. Canada, *Population and dwelling count highlight tables, 2016 census*, Feb. 2019. [Online]. Available: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=101&S=50&O=A.

[32] D. DeSilver, *Despite global concerns about democracy, more than half of countries are democratic*, May 2019. [Online]. Available: https://www.pewresearch.org/fact-tank/2019/05/14/more-than-half-of-countries-are-democratic/.

[33] *F-score*, May 2019. [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/f-score.

[34] R. Hidayatillah, M. Mirwan, M. Hakam, and A. Nugroho, "Levels of political participation based on naive bayes classifier", *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 1, pp. 73–82, 2019.

[35] R. Khandelwal, *K fold and other cross-validation techniques*, Jan. 2019. [Online]. Available: https://medium.com/datadriveninvestor/k-fold-and-other-cross-validation-techniques-6c03a2563f1e.

[36] V. Morde, *Xgboost algorithm: Long may she reign!*, Apr. 2019. [Online]. Available: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.

[37] S. Narkhede, *Understanding auc - roc curve*, May 2019. [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[38] R. Pupale, *Support vector machines(svm) - an overview*, Feb. 2019. [Online]. Available: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989.

[39] O. Shalev, *Recall, precision, f1, roc, auc, and everything*, May 2019. [Online]. Available: https://medium.com/swlh/recall-precision-f1-roc-auc-and-everything-542aedf322b9.

[40] E. I. Unit, "Democracy index 2019: A year of democratic setbacks and popular protest", 2019.

[41] M. Abdullah, *Cec: Voter turnout not over 30%*, Feb. 2020. [Online]. Available: https://www.dhakatribune.com/bangladesh/election/2020/02/01/cec-voter-turnout-not-over-30.

[42] [Online]. Available: http://bdlaws.minlaw.gov.bd/act-367/section-24681.html.

[43] [Online]. Available: https://www.indexmundi.com/bangladesh/demographics_profile.html.

[44] *3.2. tuning the hyper-parameters of an estimator*. [Online]. Available: https://scikit-learn.org/stable/modules/grid_search.html.

[45] *Bangladesh age structure*. [Online]. Available: https://www.indexmundi.com/bangladesh/age_structure.html.

[46] *Government of the people, by the people, for the people*. [Online]. Available: https://www.clingendael.org/publication/government-people-people-people.

[47] *Sklearn.model_selection.gridsearchcv*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[48] *Sklearn.preprocessing.onehotencoder*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

[49] *Undata — record view — population by age, sex and urban/rural residence*. [Online]. Available: http://data.un.org/Data.aspx?d=POP&f=tableCode:22.

[50] *Vote (verb) definition and synonyms: Macmillan dictionary*. [Online]. Available: https://www.macmillandictionary.com/dictionary/british/vote_1.

[51] *Voter turnout by age group - elections canada*. [Online]. Available: https://elections.ca/content.aspx?section=res&dir=rec/eval/pes2015/vtsa&document=table1&lang=e.

[52] *Voting*. [Online]. Available: https://www.dictionary.com/browse/voting.

[53] *Voting: Meaning in the cambridge english dictionary*. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/voting.

Appendix A

5.5 List of Features in Data set

Feature	Description
age_18-20	Participants aged between 18 to 20 years old
age_21-23	Participants aged between 21 to 23 years old
age_24-26	Participants aged between 24 to 26 years old
age_26	Participants aged 26 years and above
demo_support	Participant's support for democratic system
fam_pol_Yes	Presence of political person in family
fam_pol_No	Presence of no political person in family
fam_pol_2_No	Presence of political family member(s) does not influences participants voter participation
fam_pol_2_Yes	Presence of political family member(s) influences participants voter participation
father_edu_HSC or equivalent	Participant's father's educational qualification HSC or equivalent
father_edu_Honours or equivalent	Participant's father's educational qualification Honours or equivalent or higher
father_edu_SSC or equivalent	Participant's father's educational qualification SSC or equivalent
father_rep_vote_He does not discuss about voting	Participant's father does not talk about voting
father_rep_vote_Voting as a trend	Participant's father represent voting as a trend
father_rep_vote_Voting is important duty as a citizen	Participant's father represents the participation in voting as an important duty as a citizen
finance_1_Owned	Participant's family lives in their own house
finance_1_Rented	Participant's family lives in a rented house
finance_2_No	Participant's family does not own a car
finance_2_yes	Participant's family owns a car
high_school_Inside Dhaka	Participant completed high school inside Dhaka
high_school_outside Dhaka	Participant completed high school outside Dhaka

mother_edu_HSC or equivalent	Participant's mother's educational qualification HSC or equivalent
mother_edu_Honours or equivalent	Participant's mother's educational qualification Honours or equivalent or higher
mother_edu_SSC or equivalent	Participant's mother's educational qualification SSC or equivalent
mother_rep_vote_He does not discuss about voting	Participant's mother does not talk about voting
mother_rep_vote_Voting as a trend	Participant's mother represent voting as a trend
mother_rep_vote_Voting is important duty as a citizen	Participant's mother represents the participation in voting as an important duty as a citizen
newspaper_No subscription	Participant has a newspaper subscription
newspaper_There is subscription but I don't read them daily but my family members do	Participant has a newspaper subscription he/she don't read them but their family members do.
newspaper_There is subscription and me and my family member read them daily	Participant who have a newspaper subscription and he/she and their family members read daily
parents_aware_Yes	Participant's Parents are politically aware
parents_aware_No	Participant's Parents are not politically aware
parents_enthu_No	Participant's Parents are not enthusiastic to vote
parents_enthu_Yes	Participant's Parents are enthusiastic to vote
social_13 hours	Participant's social media usage between 1 to 3 hours
social_1 hour	Participant's social media usage less than a hour
social_3 hour	Participant's social media usage greater than three hour
social_No, I don't use social media	Participants who do not use social media
pol_aware	Participant's Political awareness
parliament	Participant's Knowledge about functionality of parliament system
judicial	Participant's Knowledge about functionality of judicial system
label	Whether the participant will vote or not (target variable)
vote_xp	Participant's previous experience of voting