1.3 Group Final Report

1. Introduction. An overview of the project and an outline of the report.

2. Description of the data set.

3. Description of the data mining and learning or cleaning algorithm or other algorithms that you used. Provide some background information on the development of the algorithm and include necessary equations and figures.

4. Experimental setup. Describe how you are going to use the data to clean and preprocess. Explain how you will implement the data mining technique in the chosen software and how you will judge the performance. Write a complete report with theoretical description and verify this mathematical concepts with applying it with actual data. Provide enough information about the codes tat you have written. Write your codes in sperate subroutines and call the functions if needed?. Explain each subroutine.

5. Results. Describe the results of your experiments, using figures and tables wherever possible. Include all results (including all figures and tables) in the main body of the report, not in appendices. Provide an explanation of each figure and table that you include. Your discussions in this section will be the most important part of the report.

6. Summary and conclusions. Summarize the results you obtained, explain what you have learned, and suggest improvements that could be made in the future.

7. References.

8. A separate appendix should contain documented computer listings.

**6103 Final Project Report – Group X**

Bradley Reardon, Kushal Ismael, Divya Parmar

The George Washington University

May 2, 2021

**Introduction**

The 2020 US Presidential Election caused a great deal of speculation surrounding voter-turnout, largely driven by the high-profile candidates and new voting methods enacted to account for Covid-19 safety precautions. Furthermore, there is a long-standing interest in what drives eligible voters to either vote or not vote, and who to vote for if a vote is cast. We decided we would like to dive deeper into the latter and explore which characteristics and features of a voter drove them to choose between Donald Trump and Joe Biden in the 2020 US Presidential Election.

In this paper, we explain how we conducted EDA on a survey results dataset, determined which features highly influenced voter decisions, and trained both a random forest and a gradient boosting classifier to predict which candidate a voter will vote for based on survey answers and demographic information.

**Dataset Description**

The dataset used in this project contains the results from a survey conducted in September 2020 by Ipsos, a multinational market research and consulting firm, and FiveThirtyEight, an American website that focuses on opinion poll analysis, politics, economics, and sports blogging. The survey focused on political subjects including what it means to be a good American, how much the survey respondent agrees or disagrees with various statements regarding systemic racism, their trust and faith in the US government, how they are affected by government policy making, thoughts on voting and past voting actions, demographic information, and who they plan to vote for in the presidential election. The raw dataset contained 119 features (columns) and 5836 observations (rows), but we reduced the number of features to 95 after deciding some were unnecessary for our project. All features are categorical due the survey questions prompting predefined answers from a given set of options. The target variable, question 23 in the survey, asked “Which presidential candidate are you planning to support?” and the given options were, “Donald Trump”, Joe Biden”, and “Unsure.”

**Data Mining/Learning Techniques**

The two machine learning techniques we decided to use are random forest and gradient boosting classifiers. Random forest and gradient boosting classifiers are similar in that they both are sets of decision trees. The two classifiers differ in two main ways: how the trees are built, and how the results are combined. In a random forest, decision trees are built independently of one another in parallel and the results are combined at the end by averaging the results from each tree. The gradient boosting classifier builds a sequence of decision trees with each tree prioritizing weak learners from the previous tree to improve its classification ability while combining results along the way.

<https://www.datasciencecentral.com/profiles/blogs/decision-tree-vs-random-forest-vs-boosted-trees-explained#:~:text=Like%20random%20forests%2C%20gradient%20boosting,one%20tree%20at%20a%20time>.

**EDA/Preprocessing**

4. Experimental setup. Describe how you are going to use the data to clean and preprocess. Explain how you will implement the data mining technique in the chosen software and how you will judge the performance. Write a complete report with theoretical description and verify this mathematical concepts with applying it with actual data. Provide enough information about the codes tat you have written. Write your codes in sperate subroutines and call the functions if needed?. Explain each subroutine.