Individual Final Report

Introduction

Our project analyzes data from a survey conducted by FiveThiryEight and Ipsos on what motivates Americans to vote in elections. This survey was conducted prior to the 2020 US Presidential Election, an election with expanded absentee and mail-in ballot procedures put into place because of the COVID-19 pandemic.

We used the survey data to determine if we could predict whether a voter would vote for Donald Trump or Joe Biden based on their response to questions in the survey. The data needed to be cleaned, explored, and processed before training and testing models for prediction. Final findings of the project would be added to a GUI.

As a group we met weekly on Tuesdays to discuss progress and findings of the project. We met to choose a final dataset, explore the data, talk about findings from EDA, determine what we planned to predict, and assign portions of the work to one another.

Individual Work

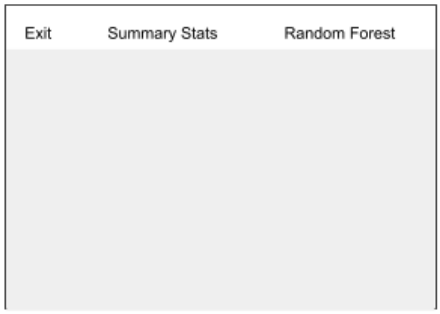
My contribution to the project was creating the GUI to showcase our exploratory data analysis and Random Forest model.

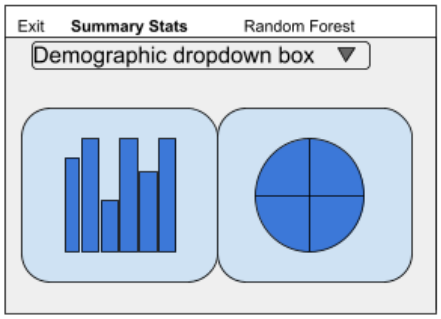
Individual Work – Detail

I learned how PyQt5 works by following the tutorial within the class files and exploring the Demo.py file to get an idea of the class structures and how they interact to display information in the GUI application. Before starting to code the GUI, I created an initial mockup of the planned GUI so that I had an idea of what I planned to build. To organize the class inheritance within the GUI and better understand how each part would work, I created a flowchart to show how each class and function would interact. (see pictures below)

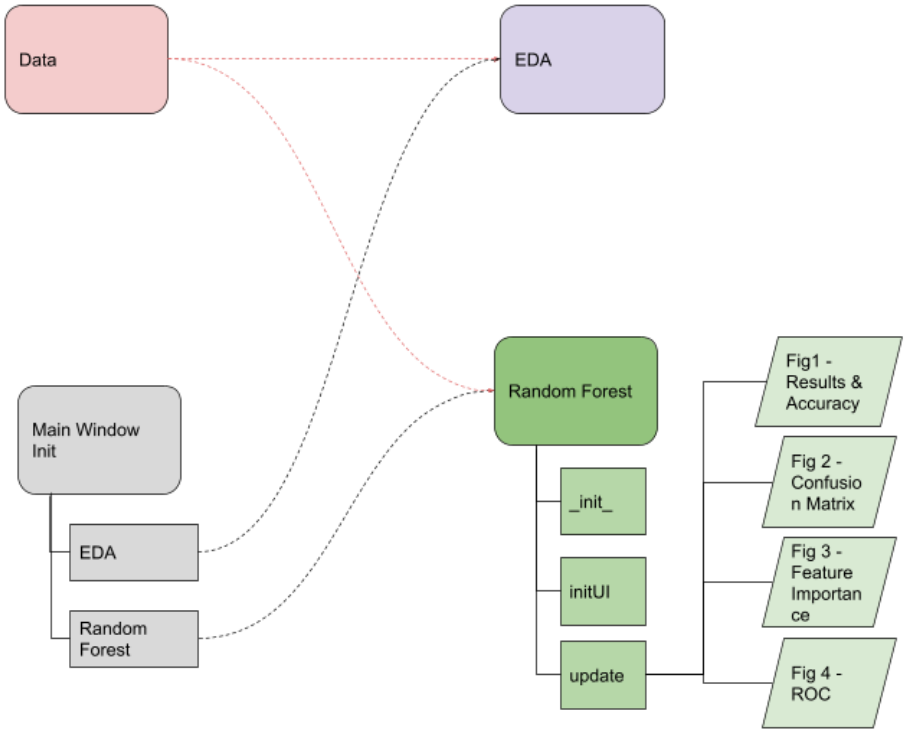
I began by creating the menu options, followed by the layouts for each of the windows so that I had a framework to add the plots and data into the GUI. I explored the Qt Widgets from the PyQt5 documentation to evaluate what widgets would work best in the GUI. After the EDA and modeling were finished by my group members I used their code to read in the data and pre-process it within the GUI. I then finalized the plots and layouts so that they run without error in the GUI.

GUI Skeleton





Classes and functions



Results

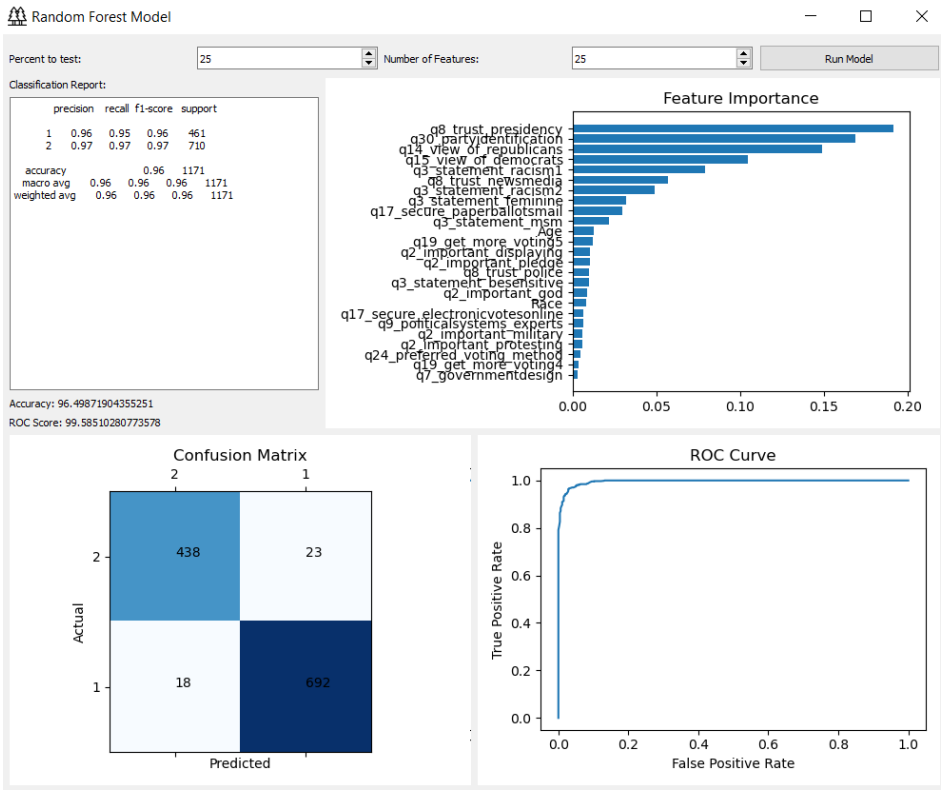
After preprocessing the data, dropping questions that were not provided to all respondents, and replacing missing values with a mean; we ran the data through four models.

The results for each of the models that we ran are as follows:

Model 1 Random Forest - Full Model	**Model 2 Random Forest - Slim Model**	**Model 3 Gradient Boosting - Full Model**	**Model 4 Gradient Boosting - Slim Model**
F1-score: 0.98 Accuracy score: 0.97	F1-score: 0.93 Accuracy score: 0.93	F1-score: 0.97 Accuracy score: 0.97	F1-score: 0.50 Accuracy score: 0.51

The Random Forest model that utilizes all of the features had the highest accuracy and F-1 score. The accuracy score determines how often the model is correct in its predictions whereas the F-1 score provides more nuance by factoring in false positives and negatives of a model. We chose to add the Random Forest model to the GUI as it had the highest accuracy and F-1 Score.

Below is a screenshot of the Random Forest model in the GUI run with 25% test and 25 features. It also has a high accuracy of 96.5%. The model was able to correctly predict 1,130 voters' selection for president and only got 41 wrong. In this instance of the model, question 8 on how much someone trusts the presidency was the most important feature, followed by question 30 on party identification.



Summary and conclusions

A person's view on politics will generally determine who they plan to vote for within presidential elections. It is not surprising that there is strong correlation between responses to questions on political views and selection of a candidate for president. The political nature of each question in relation to two polarizing presidential candidates supports the very high accuracy and f-1 scores from the random forest and gradient boosting models.

If the survey was conducted between presidential candidates of the same political party, or candidates at different levels of government then the model may not have as high an accuracy score. The model would still be extremely beneficial in determining the political issues that are most important in predicting which candidate someone will vote for in an election.

Code

0% of the code that I used in the GUI was copied from the internet. I utilized documentation from PyQt5, Matplot, pandas, and python files from class for all of my code.

References

- https://scikit-learn.org
- https://numpy.org/
- https://pandas.pydata.org/
- https://doc.qt.io/qtforpython/
- https://www.flaticon.com/
- https://matplotlib.org/