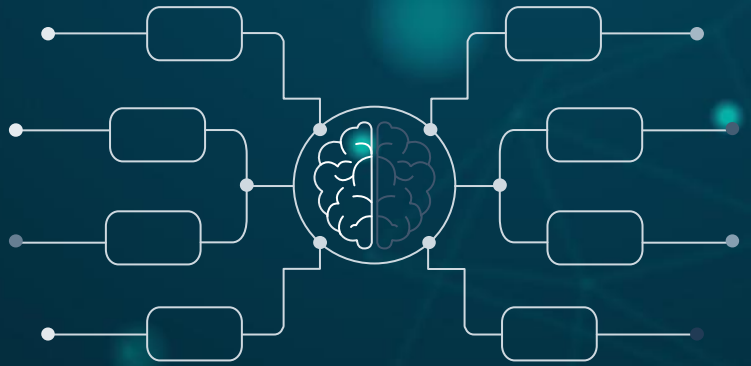
# Voter Prediction

Kushal Ismael, Divya Parmar, Bradley Reardon

# Table of Contents





# Introduction 01

# Project Overview

Using machine learning classification techniques, we predict which candidate a voter plans to vote for based on self-reported survey data.

We chose this dataset because the topic is highly relevant, there was an opportunity to apply EDA and preprocessing to a large number of features, and the dataset was mostly clean and well-documented courtesy of FiveThirtyEight.

# Dataset Overview

# 02

# Dataset Source

FiveThirtyEight

An American data journalism website that focuses on opinion poll analysis, politics, economics, sports, and predictive models

Ipsos

A multinational market research and consulting firm

# Dataset Overview

## Nonvoters Survey Question Subjects

- What it means to be a good American
- Level of agreeance towards systemic racism statements
- Faith (or lack thereof) in the US Government
- How affected one feels by government policy making
- Thoughts on voting, and past voting actions
- Demographic information


## Dataset Structure

- 5,836 observations
- 199 features – all categorical

## Target Feature

- Question 23 – “Which presidential candidate are you planning to support?”
- Possible answers: “Donald Trump”, “Joe Biden”, “Unsure”

# EDA & Preprocessing

# 03

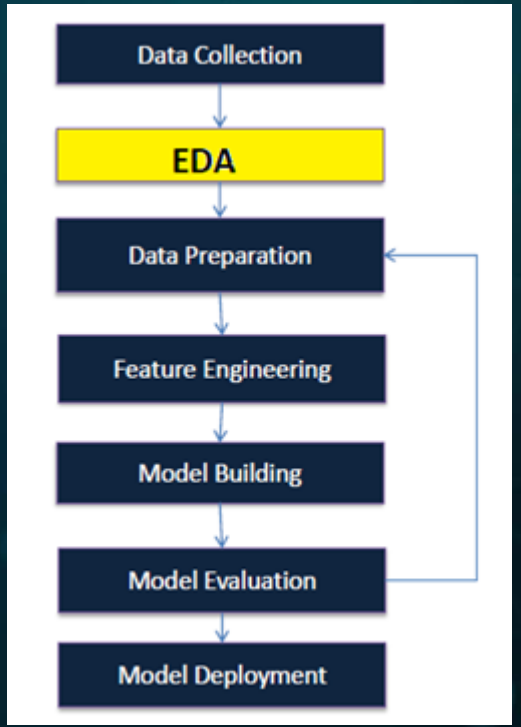
# EDA

## Cleanup

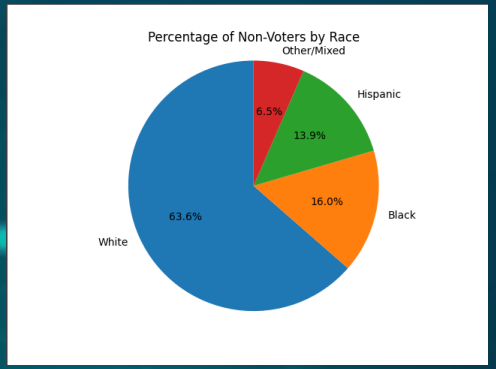
- Renamed features to increase readability
- Dropped 23 features that had only partial responses (i.e. “Which type of Republican are you?” would be null for Democrats in the sample)

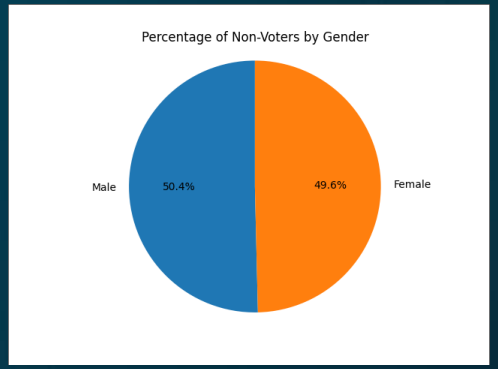
## Visualization

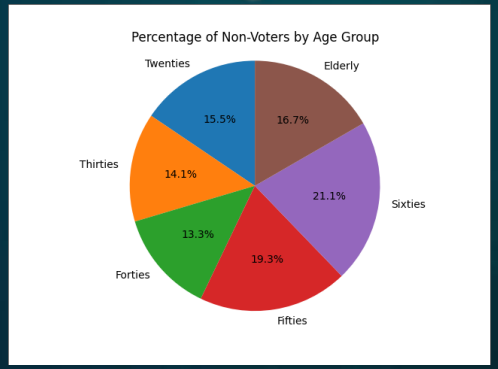
- Added Age_Group feature to see distribution of non-voters by age group
- Plotted demographic information (race, gender, education, etc) to check for balance

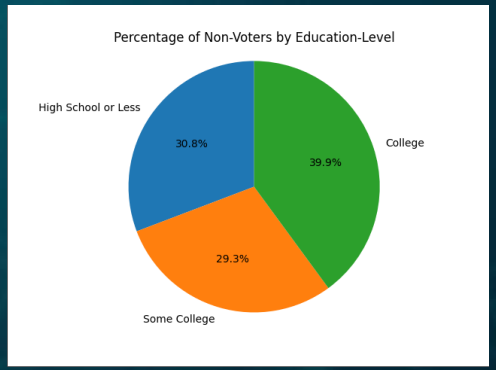


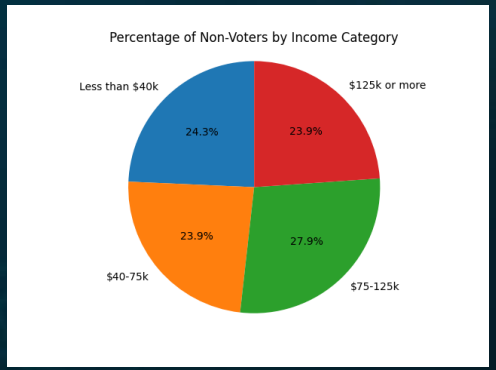
# Normality Check – Pie Charts



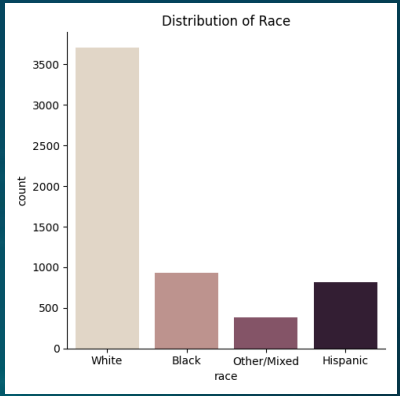



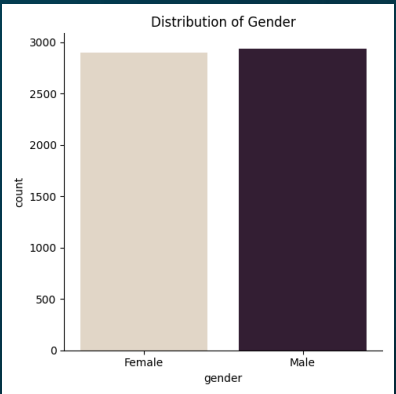



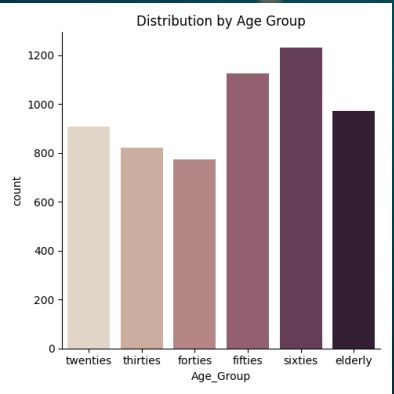



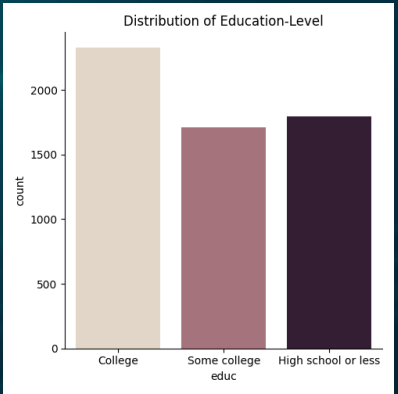



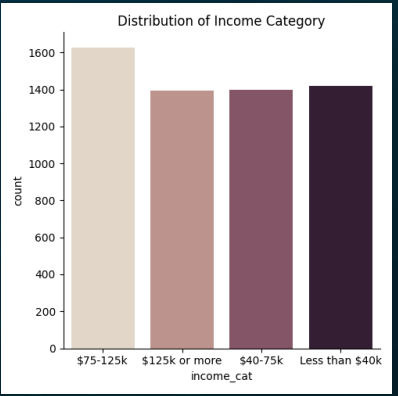
# Normality Check - Histograms



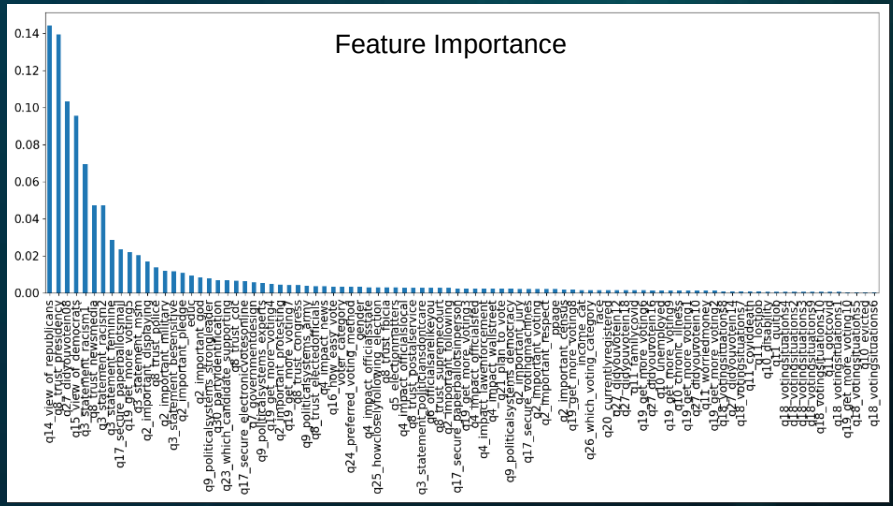



# Preprocessing

- Label encoded categorical features
- Dropped observations where person supported third-party candidate or refused to answer
- Replaced non-answers with the mean response of the individual's demographic group
- Calculated feature importance
- Dropped features that were direct proxy for our target variable ("What is your view of Republicans?", "Do you trust the presidency?")



# Modeling 04

# Random Forest Classifier

## Model 1 Specifications - Full Model

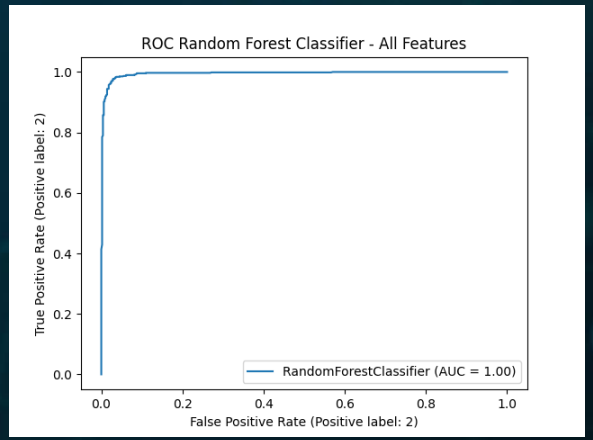
- Standard random forest classifier
- Trained on full dataset
- N estimators = 500

```
Results Using All Features:

Classification Report:
              precision    recall  f1-score   support

           1       0.98      0.96      0.97       491
           2       0.97      0.98      0.98       680

    accuracy                           0.97      1171
   macro avg       0.97      0.97      0.97      1171
weighted avg       0.97      0.97      0.97      1171
```



# Random Forest Classifier

## Model 2 Specifications: Slim Model

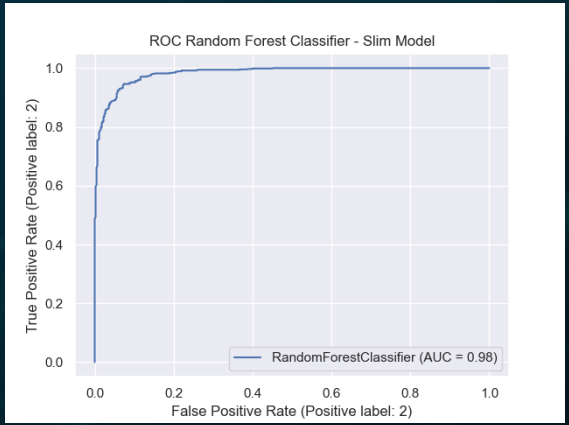
- Standard random forest classifier
- Dropped top features
- Determined via feature importance calculation
- N estimators = 500

```
Results For Slim Model:

Classification Report:
              precision    recall  f1-score   support

           1       0.93      0.90      0.91       449
           2       0.94      0.96      0.95       722

    accuracy                           0.93      1171
   macro avg       0.93      0.93      0.93      1171
weighted avg       0.93      0.93      0.93      1171
```



# Gradient Boosting Classifier

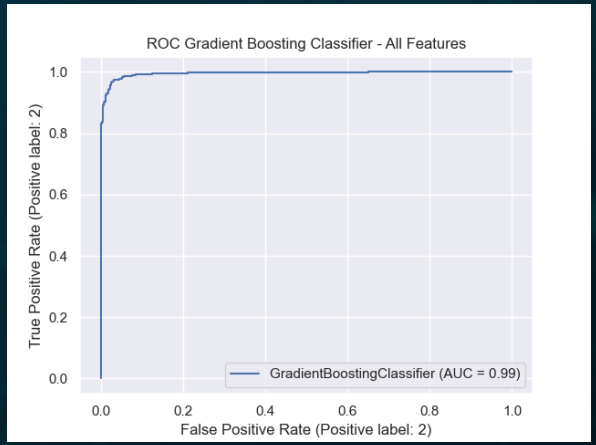
## Model 3 Specifications - Full Model

- Standard gradient boosting classifier
- Trained on full dataset
- N estimators = 500
- Learning rate = 0.05

Results Using Gradient Boosting & All Features:

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.97 | 0.97 | 0.97 | 491 |
| 2 | 0.97 | 0.97 | 0.97 | 680 |
| accuracy | | | 0.97 | 1171 |
| macro avg | 0.97 | 0.97 | 0.97 | 1171 |
| weighted avg | 0.97 | 0.97 | 0.97 | 1171 |



# Gradient Boosting Classifier

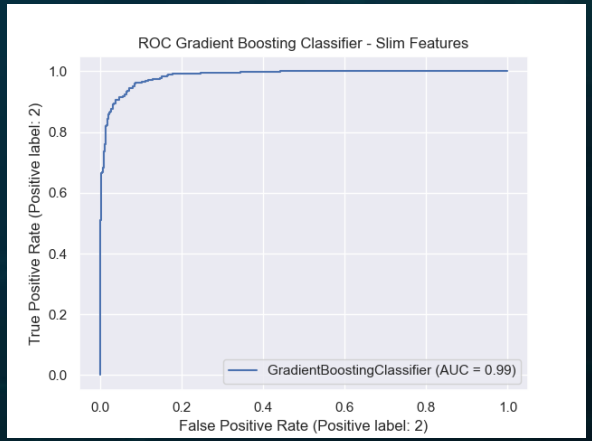
## Model 4 Specifications - Slim Model

- Standard gradient boosting classifier
- Trained on full dataset
- N estimators = 500
- Learning rate = 0.05

Results Using Gradient Boosting & Slim Dataframe:

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.41 | 0.37 | 0.39 | 491 |
| 2 | 0.58 | 0.62 | 0.60 | 680 |
| accuracy | | | 0.51 | 1171 |
| macro avg | 0.49 | 0.49 | 0.49 | 1171 |
| weighted avg | 0.51 | 0.51 | 0.51 | 1171 |

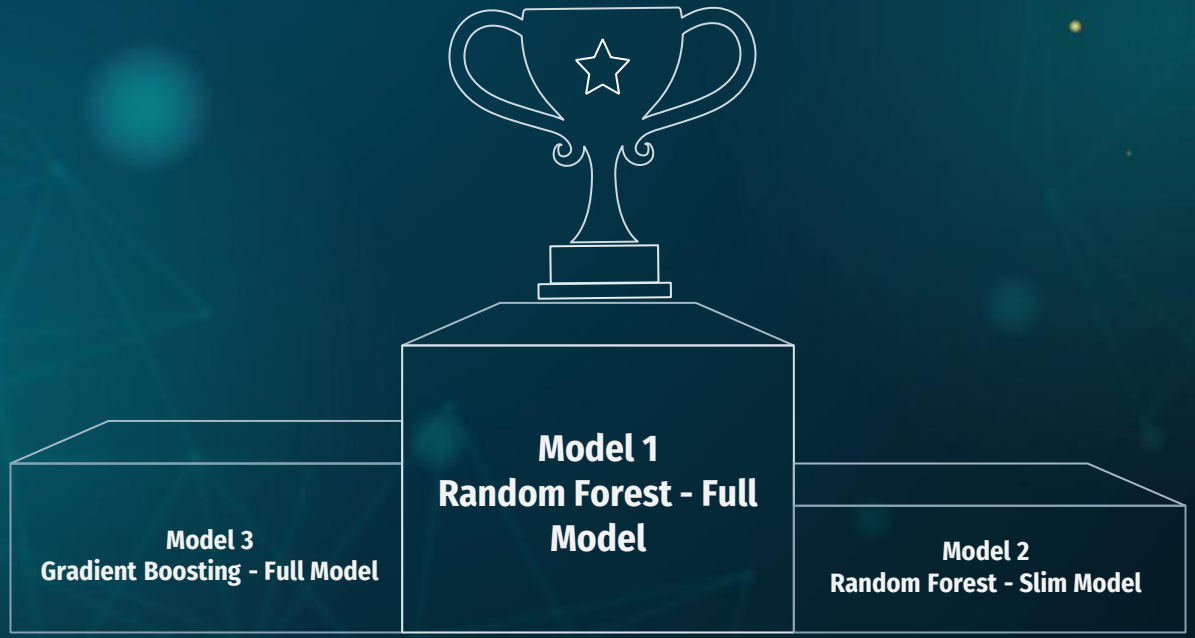


# Results 05

# Model Comparison

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

# Model Comparison



# Conclusion 06

# Conclusion

- Using FiveThirtyEight survey data, we decided to predict who voters would vote for president based on their survey answers
- We conducted exploratory data analysis to better understand the group of voters and make sure the classes were balanced
- We preprocessed our data - label encoding, dropping columns and observations
- We fit both random forest and gradient boosting models, and we ran on “full” and “slim” feature sets
- We saw extremely high accuracy and f1 scores
- Random forest did better than gradient boosting, and “full” feature models did better than “slim” feature models

# Caveats

- Although our models had high accuracy, the impact of our work is limited. Knowing how someone votes based on their opinion of Democrats and Republicans, view of police, view of systemic racism, etc. is not a novel finding.  Predicting elections involves deciding how individuals will vote, while having much less information than we do here, as well as trying to predict turnout rates. A small, controlled sample cannot substitute for this.
- Another question we could have considered is voter turnout. We could have set the target variable as whether someone planned to vote or not, then analyzed it by political affiliation, demographics, etc. This would get to the heart of this survey, which wanted to know why eligible voters don't vote.

# Citations


- https://scikit-learn.org
- https://numpy.org/
- https://pandas.pydata.org/
- https://pypi.org/project/PyQt5/
- https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56
- 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
- https://morningconsult.com/opinions/to-persuade-or-to-turn-out-voters-is-that-the-question/
- https://www.bloomberg.com/graphics/2020-us-election-results/methodology


# GUI Demonstration

# 07

# GUI Demonstration

*Insert GUI Demo vid*

# Thank you!