# CAPSTONE TWO
# FINAL PROJECT REPORT

Presented by Kanchanah Kannathass


# Table of Contents


- Problem Statement
- Dataset
- Data Wrangling
- Exploratory Data Analysis
- Preprocessing
- Model Building


Presented by Kanchanah Kannathass

# PROBLEM STATEMENT

How can government agencies (federal and state) promote the adoption of EV to constitute 50% of the automotive industry in the next 5 years by identifying the important factor(s) or group of factor(s) that lead to the non-adoption in the US?

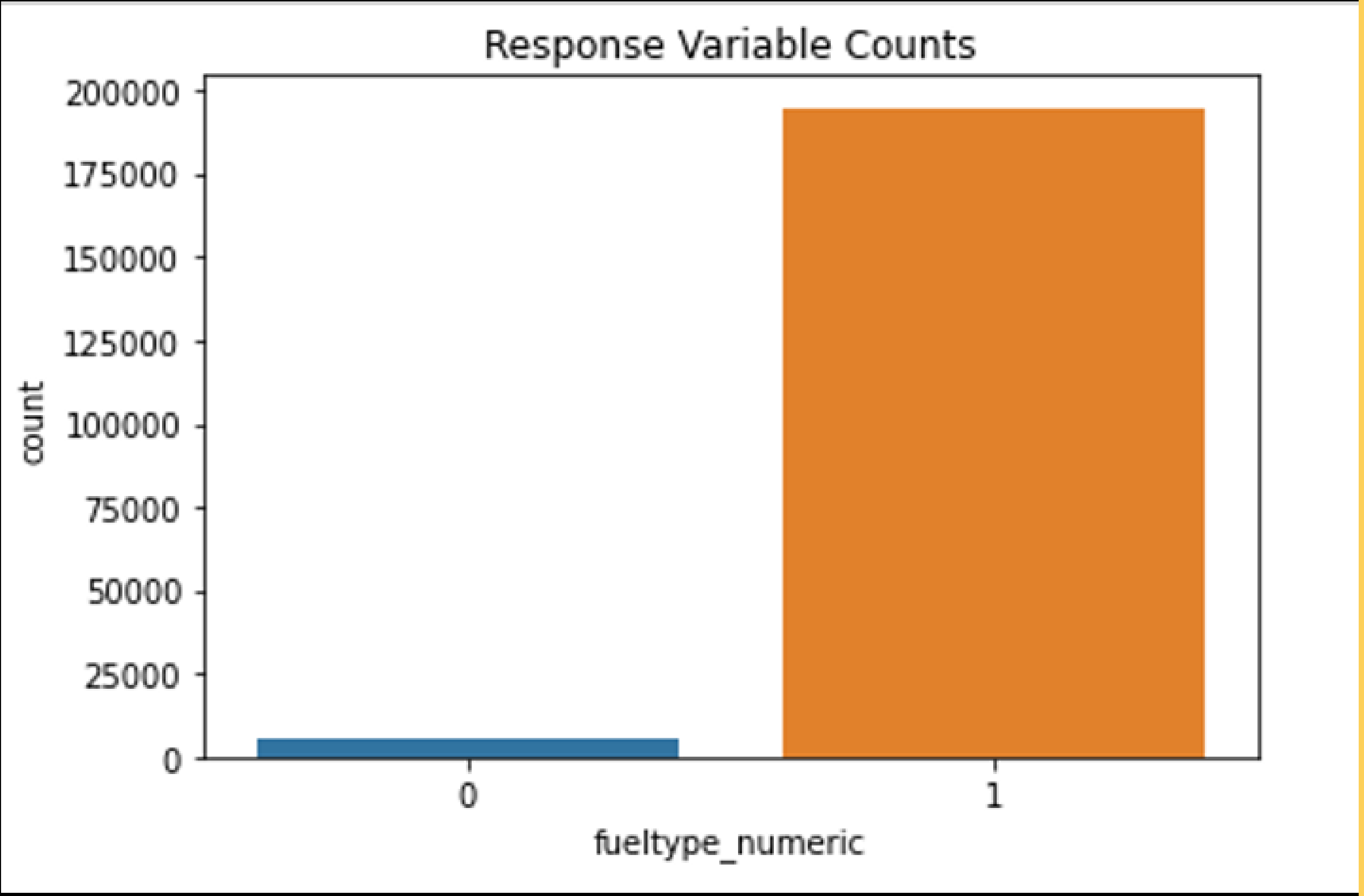
# DATASET

- 2017 National Household Travel Survey (NHTS)

- Combination of 3 datasets
  - Individual personal and household characteristics
  - Socio-economic characteristics
  - Vehicle ownership and vehicle attributes

# DATA WRANGLING

- Variable Type
  - Some variables were converted to numeric and others to characters
- Drop Duplicate Car Information
- Special Codes
  - Codes were converted to NA for the purpose of this project
- Response variable
  - Gas, Diesel and Some other fuel were recoded to non-EV
  - Hybrid, Electric or Alternative fuel were recoded to EV
- Recoding variables
  - Ordinal variables were recoded to depict ordinality and the nominal variables to reduce cardinality




# Exploratory Data Analysis - Response Variable

- About 97% own non-EV in the sample
- Most respondents still prefer to use non-EV vehicles as their mode of travel

# Correlation between Variables

## Notable Correlations

- **HHSIZE and DRVRCNT**
  - As the count of household members (HHSIZE) increase, number of drivers (DRVRCNT ) could increase as well
- **HBPPOPDN and HBRESDN**
  - Category of population density and category of housing units are positively correlated
  - As one increases, we would expect the other to increase as well
- **WRKCOUNT and DRVRCNT**
  - As the number of workers in the house increases, we would expect drivers to increase as well

# Preprocessing

- **Data Split**
  - Dataset was split 70% into the train and 30% into test dataset

- **Imputation for Missing Variables**
  - Numerical variables imputed using the median
  - Categorical variables imputed by the most frequent value

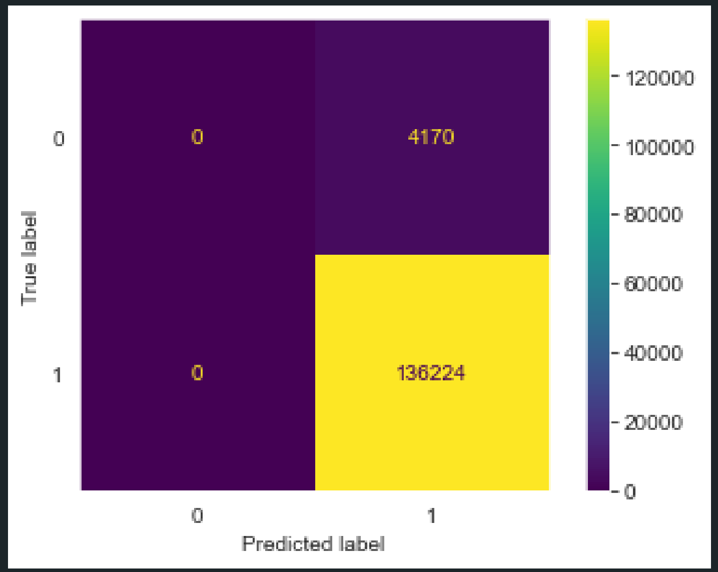
- **Scaling of Variables**
  - Numerical variables were scaled using the min max scaler

- **Dummies**
  - Dummies created for the categorical variables

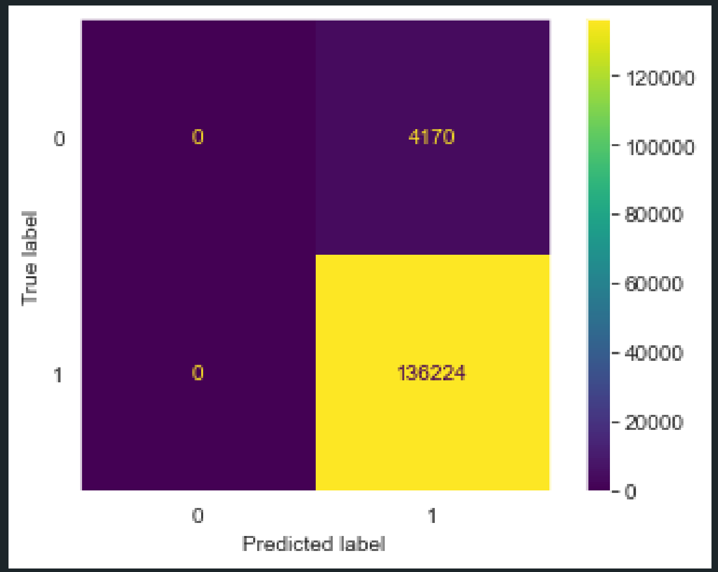
# MODEL BUILDING

## Results from Analysis

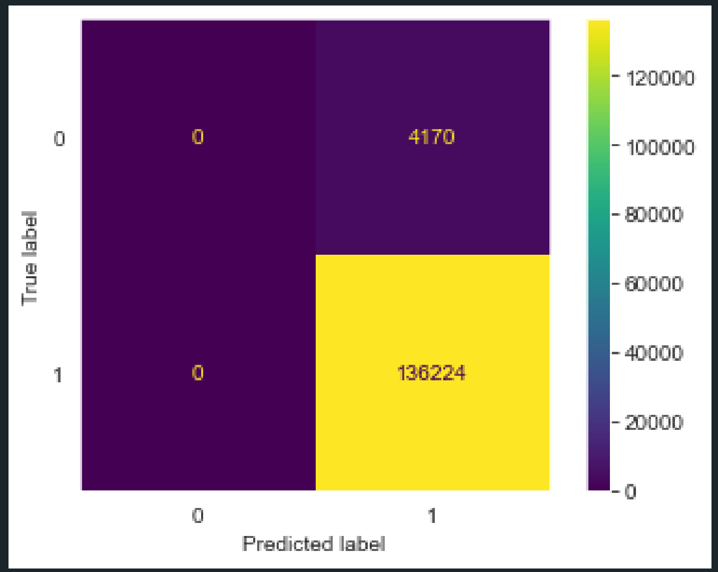
| | F1 score for Train | F1 sore for Test | Precision for Train | Precision for Test | Recall for Train | Recall for Test |
|---|---|---|---|---|---|---|
| Naïve Model | 0.956 | 0.956 | | | | |
| Logistic Regression | 0.985 | 0.985 | 0.998 | 0.999 | 1 | 1 |
| Poission Regression | 0.985 | 0.985 | 0.997 | 0.998 | 1 | 1 |
| Random Forest | 0.985 | 0.985 | 0.995 | 0.995 | 1 | 1 |




**Logistic Regression**




**Poisson Regression**

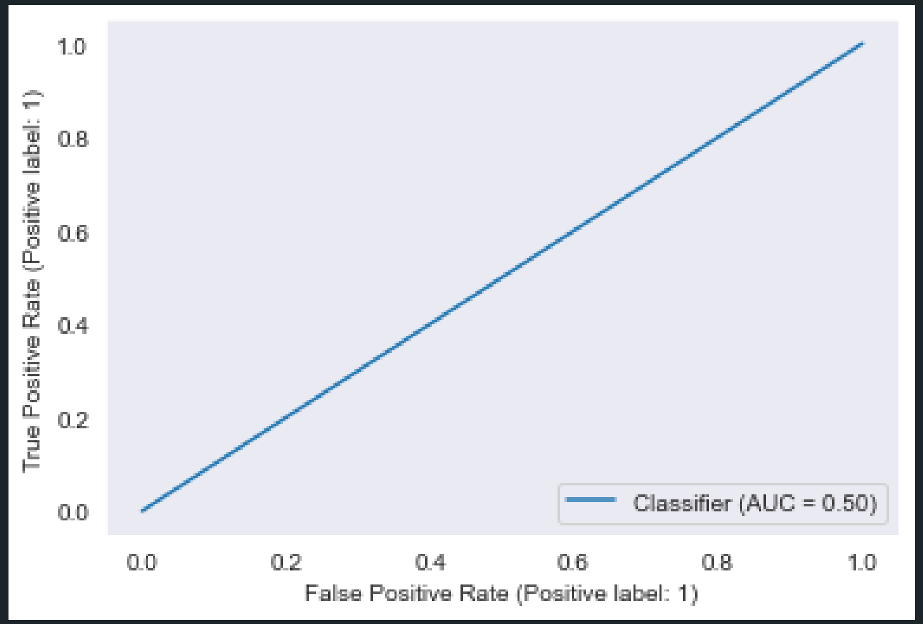



**Random Forest**

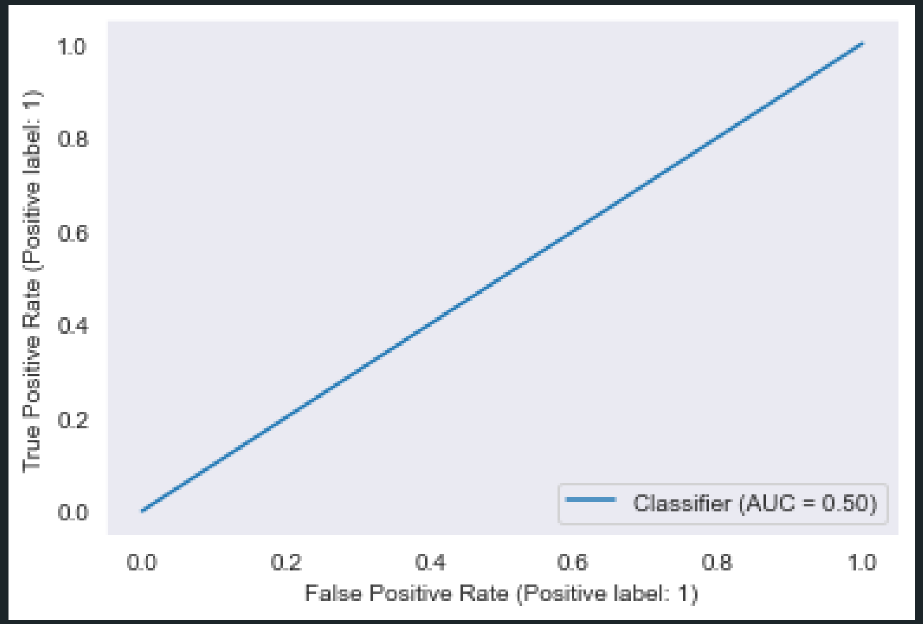
- Similar confusion matrices for all 3 models
- Further assessment was done using the ROC

# MODEL BUILDING

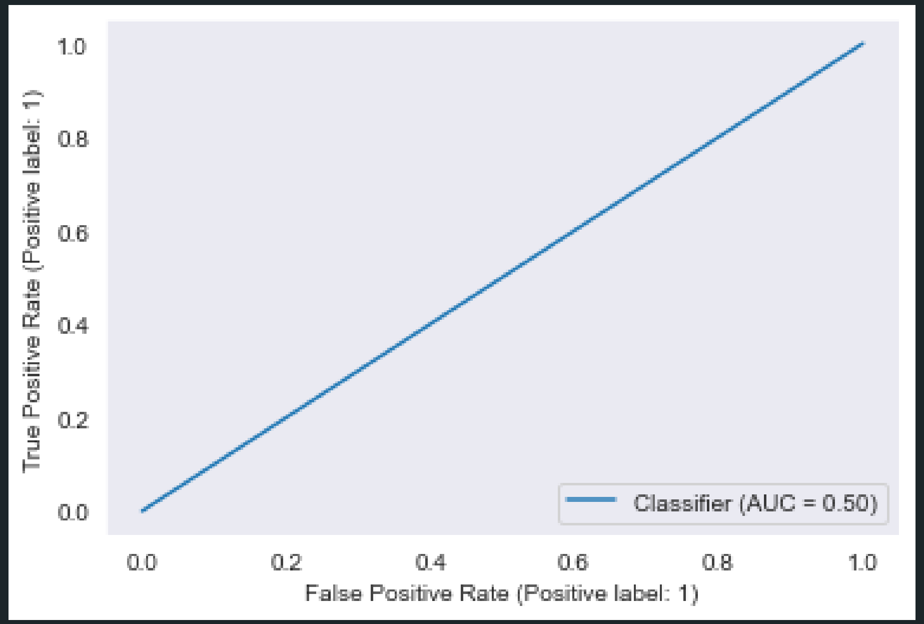
## Results from Analysis




Random Forest




Poisson Regression

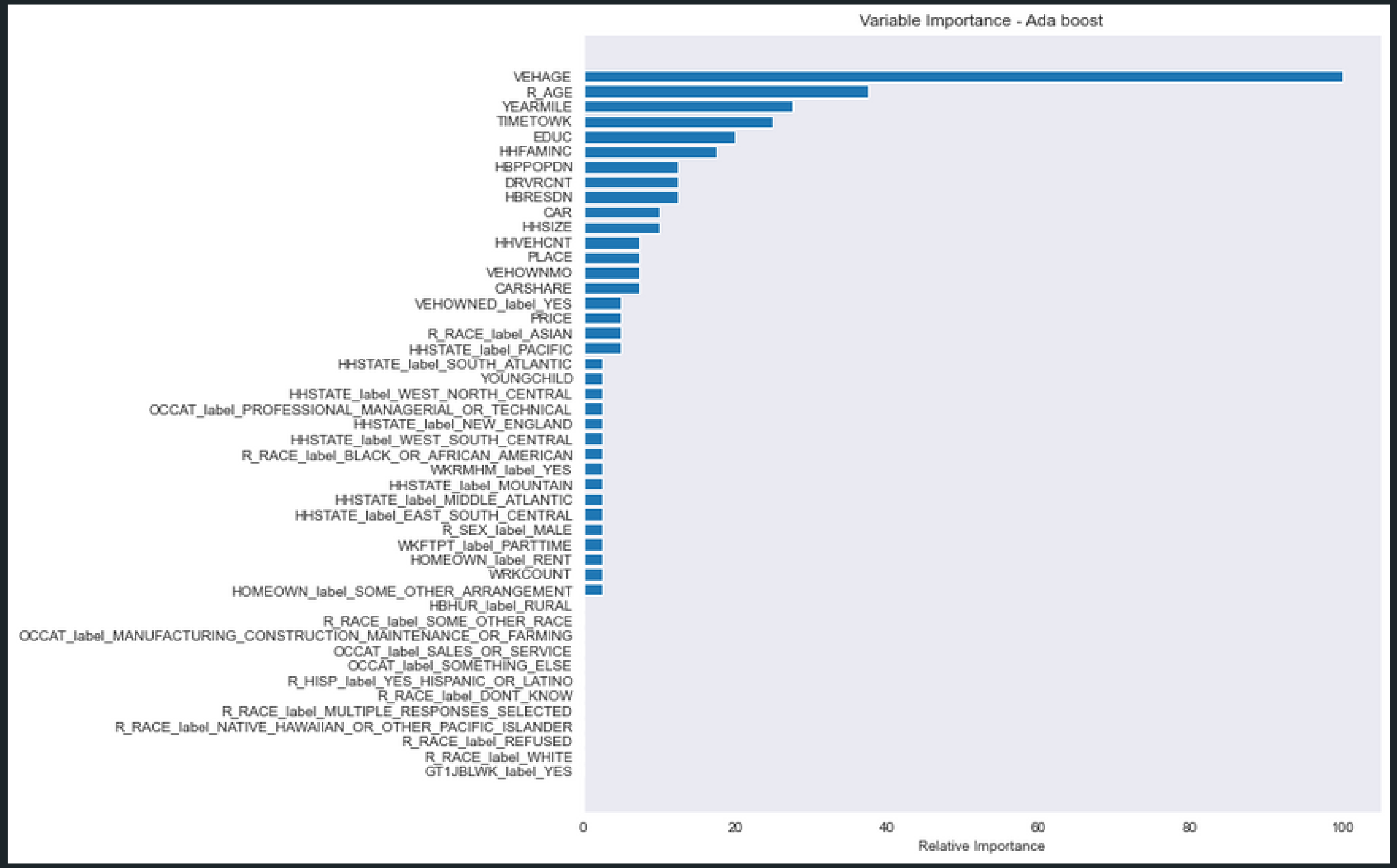



Logistic Regression

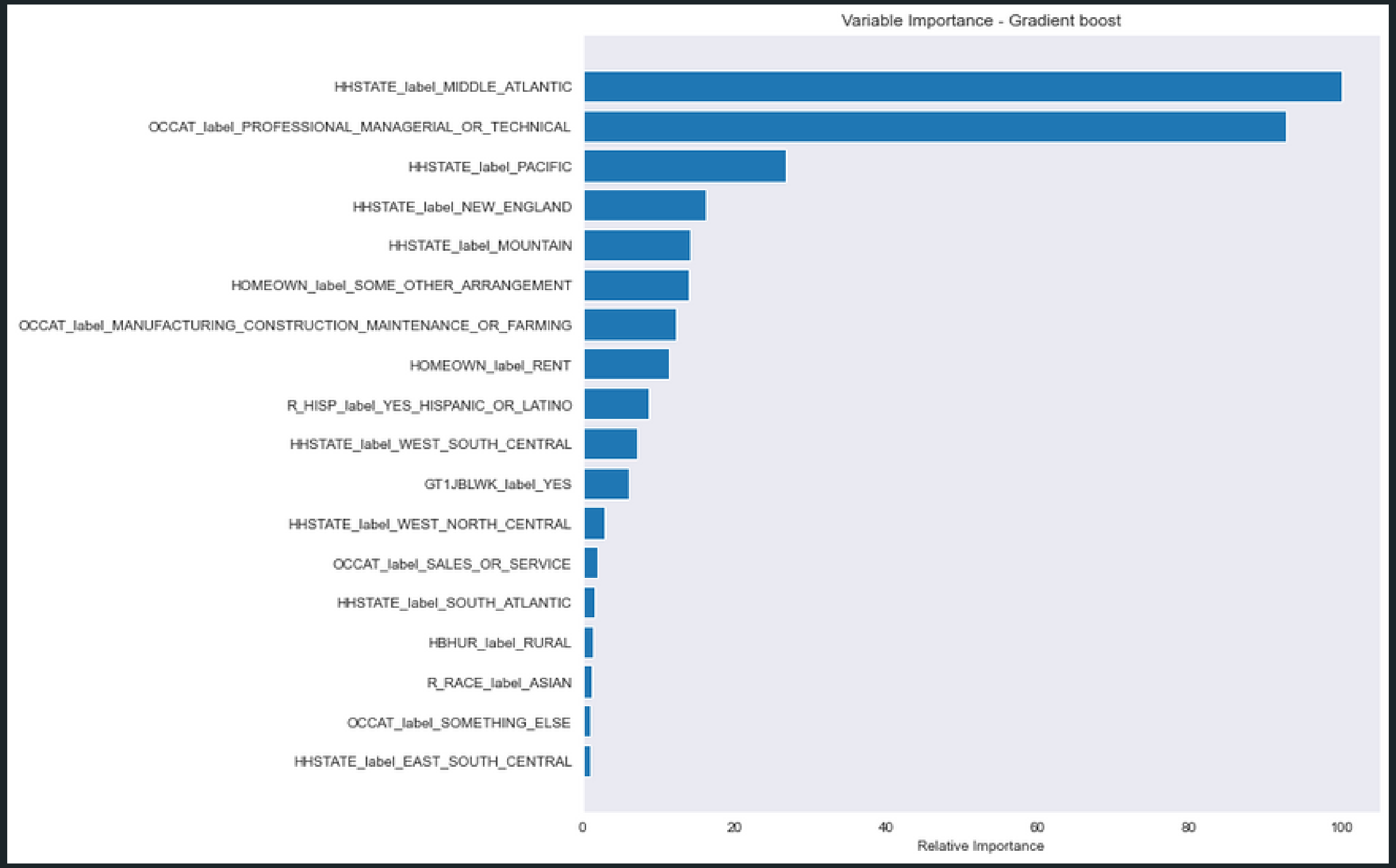
- Diagonal plot shown indicate model was guessing how to classify the data
- No pattern seen in the classification

# MODEL BUILDING

## Futher Modeling - Variable Importance

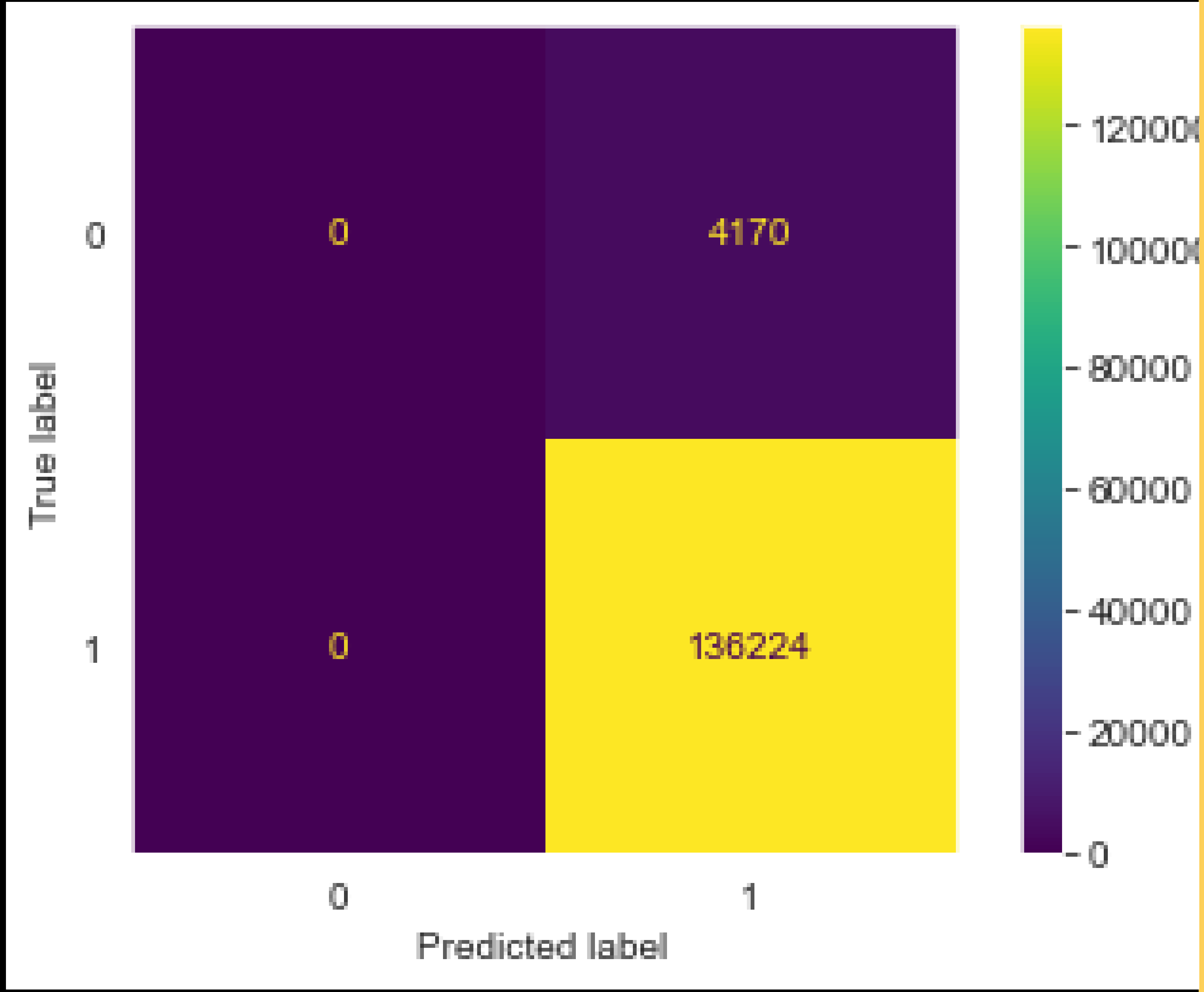



AdaBoost




Gradient Boosting

- HHSTATE_label_PACIFIC, EDUC, HHFAMINC, R_AGE, PLACE and PRICE were chosen as most important




Linear Model with SGD

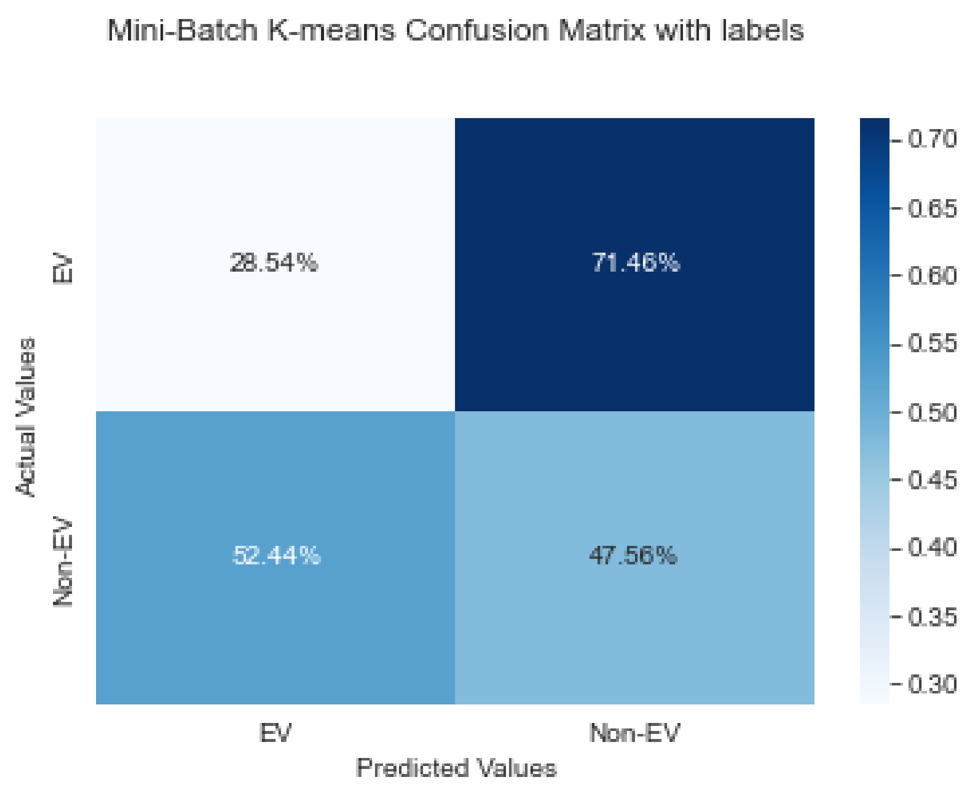
# MODEL BUILDING

## Futher Modeling - Stochastic Gradient Descend (SGD)

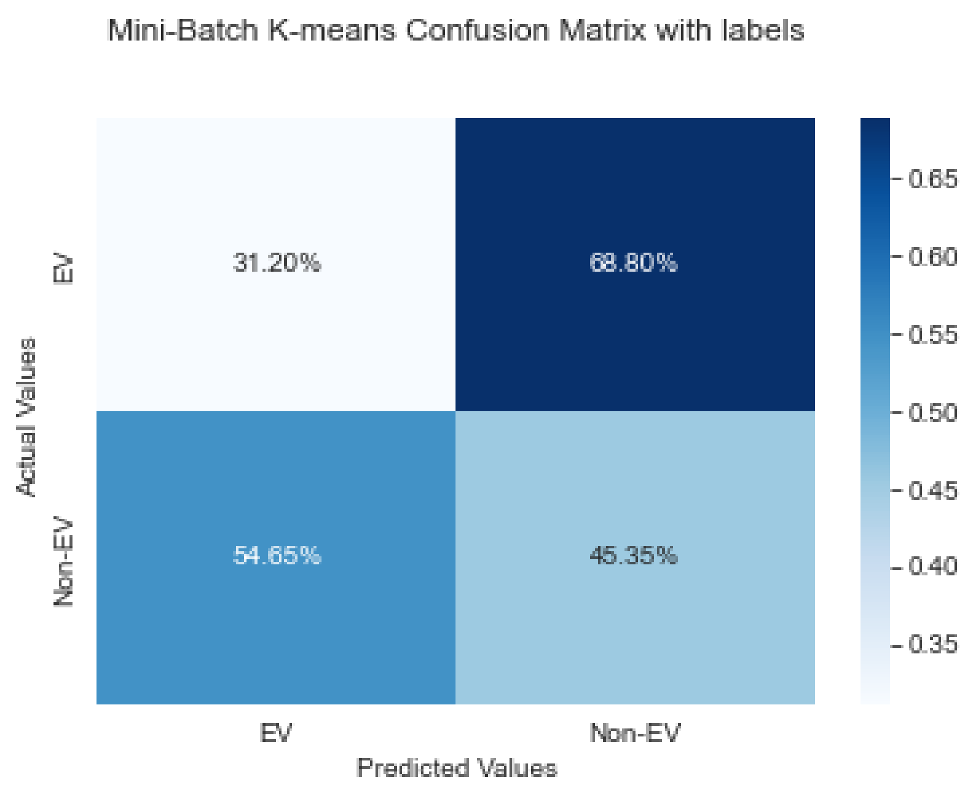
- Used most important variables from Adaboost and Gradient Boosting
- Results from this model did not differ from the previous 3 supervised models
- Another method to model the data needs to be considered

# MODEL BUILDING

## UNSUPERVISED LEARNING




**K-means**




**Mini-batch K-means**

- Results seem much better than the results from supervised learning as not all observations were predicted as non-EV

# Further Assessment of Unsupervised Models

## EACH VARIABLE WAS VISUALIZED USING A DASHBOARD

https://projectev.herokuapp.com

Presented by Kanchanah Kannathass

# Conclusion

- The dataset was particularly challenging as the response was imbalanced
- Supervised models could not classify the adequately
- Unsupervised learning model worked better than a supervised model to classify the dataset

Presented by Kanchanah Kannathass