Data Exploration on Caravan Insurance

Needle in a Haystack

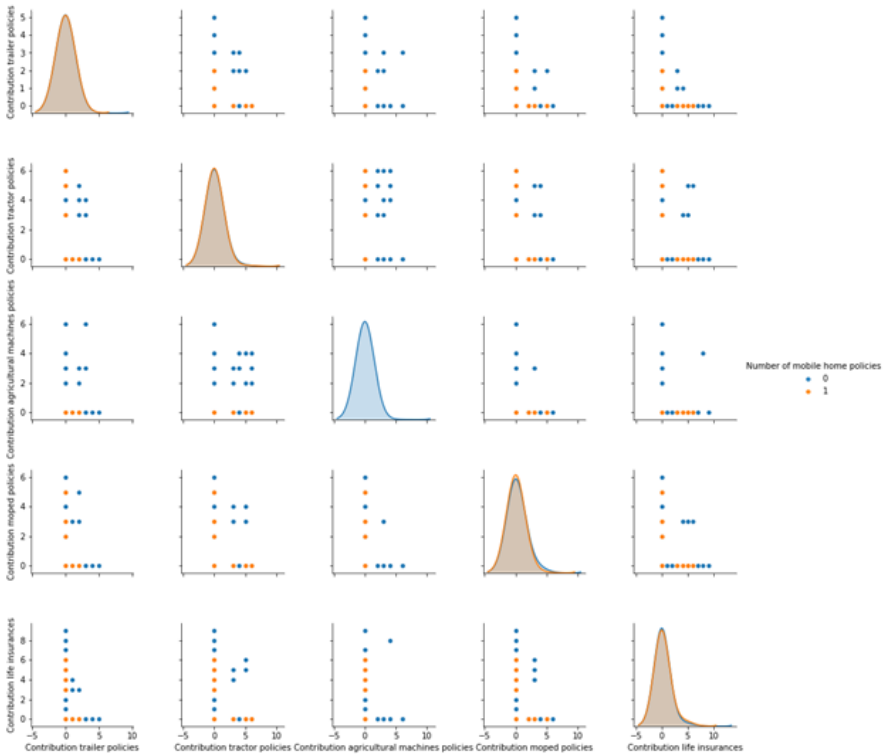
- Purpose of this project
  - Seeking potential buyers of Caravan Insurance
  - Who and why they may be at least interested

Data

- Source: https://www.kaggle.com/uciml/caravan-insurance-challenge
- Variables
  - 85 Variable
  - Categorical Variables problem

EDA

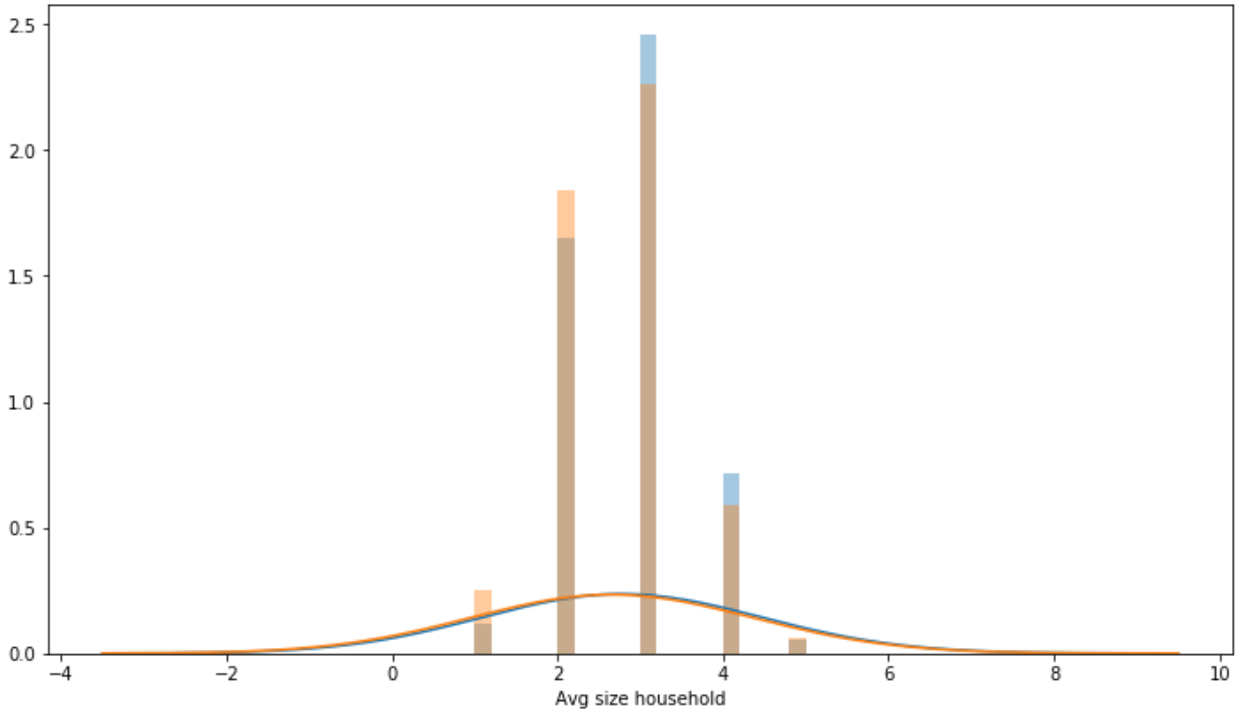
- Pair Plot



- The pair plot shows how 1 variable may or may not correlate with the other
- When a variable paired with itself, the histogram is produced instead
- The plot is showing groupings with observations with target variable = 0, in blue, or target variable = 1, in orange

EDA

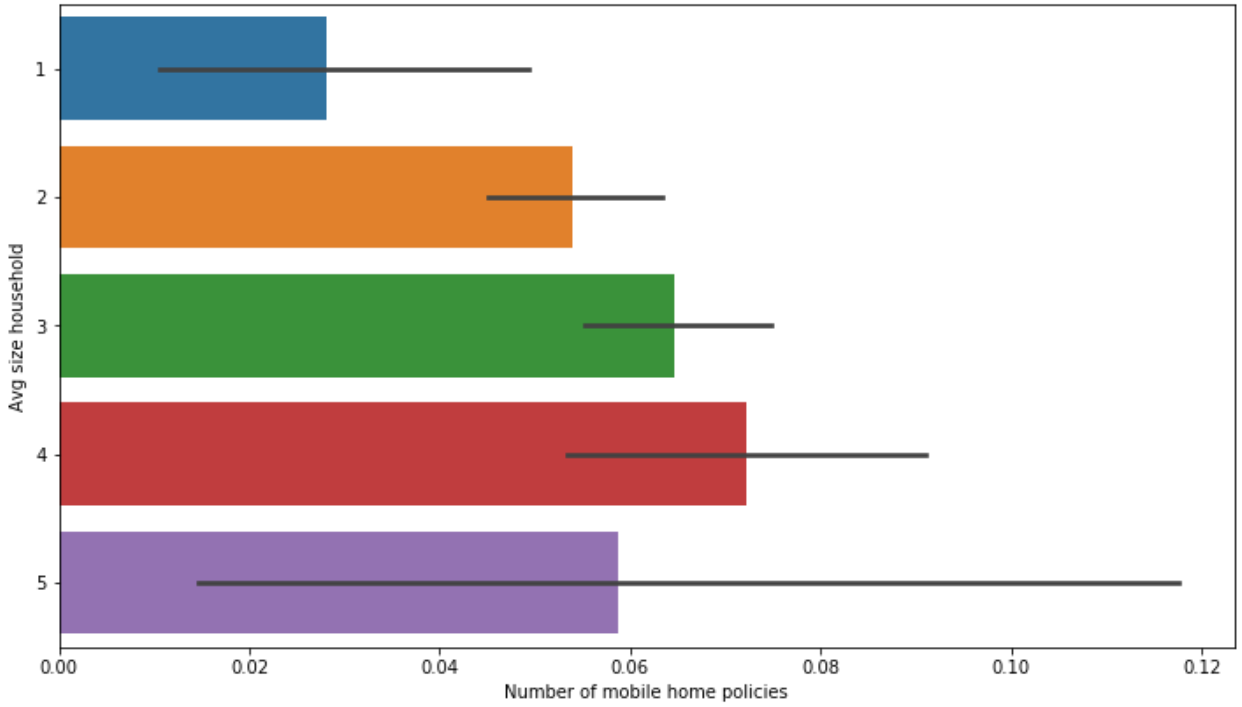
- Histogram



- Exploring data with histograms
- Look closer by variable
- Trying to see if visually there's a noticeable difference those who have Caravan Insurance to those who don't

EDA

- Bar Plot



- Another perspective to explore data
- Looking at the mean proportion to see if there's a difference by class within a given variable

Statistical Inference

Chi Square test

- Infer whether a variable has its mean estimate different from the overall mean

Bootstrap

- Simulate the data and graphically represent

Analysis with Machine Learning Models

- Data for modeling
  - Reducing dimension from 85 variable to 23 variable
  - Based on the P-value on each variable in the Chi-Square test
- Classification Problem
  - Yes/No vs. Continuous

Models for consideration

- Decision Tree
  - Basic Model for Reference
- Random Forest
  - Expansion of Decision Tree
- Gradient Boost Machine
- Logistic Regression

Understanding model performance

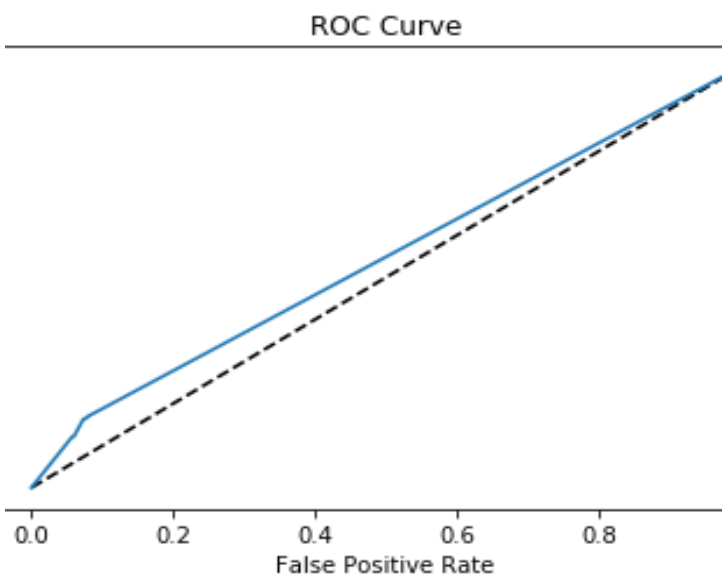
- Confusion Matrix, Precision, Recall and Accuracy
- ROC and AUC
- Lift charts and Cumulative Gain Chart

Performance Measure from Confusion Matrix

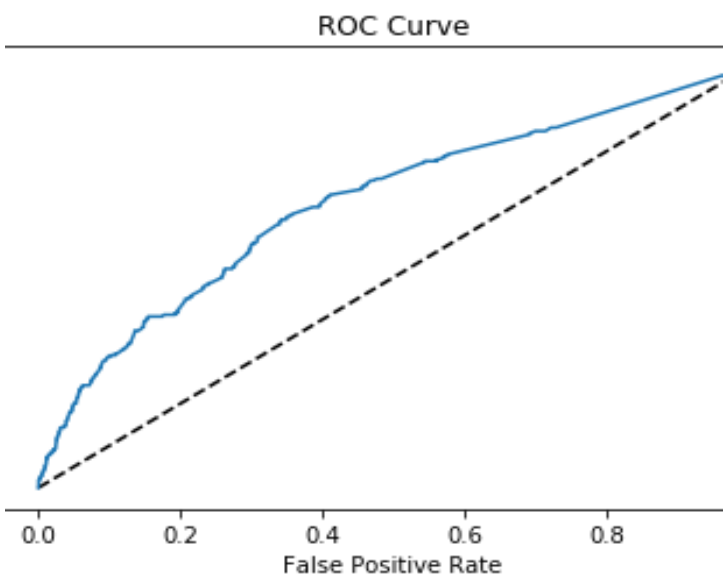
Comparison on Test Data	Classification Tree	Random Forest	GBM	Logistic Regression
Accuracy	0.891	0.934	0.94	0.94
Precision	0.114	0.25	0.429	0.5
Recall	0.122	0.055	0.013	0.013

ROC and AUC

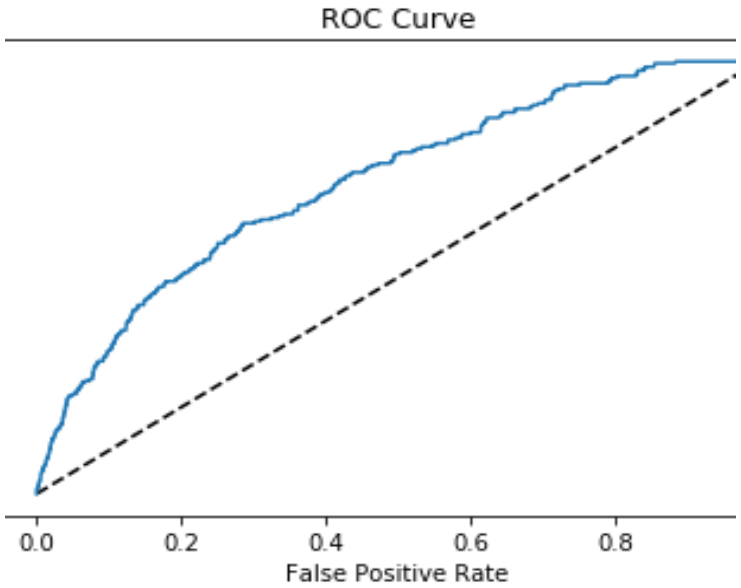
Decision Tree



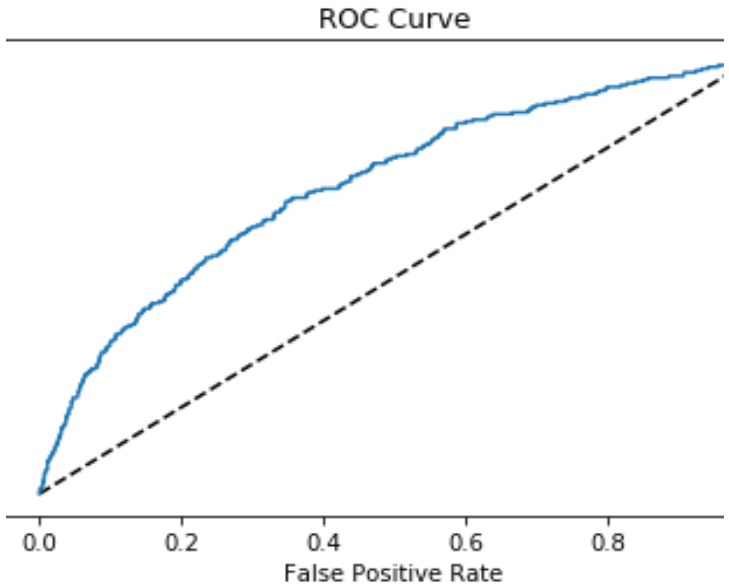
Random Forest



GBM

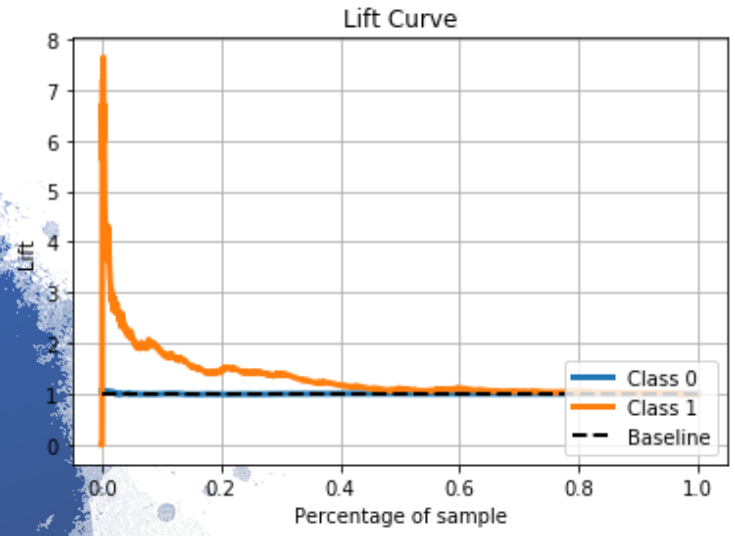


Logistic Regression

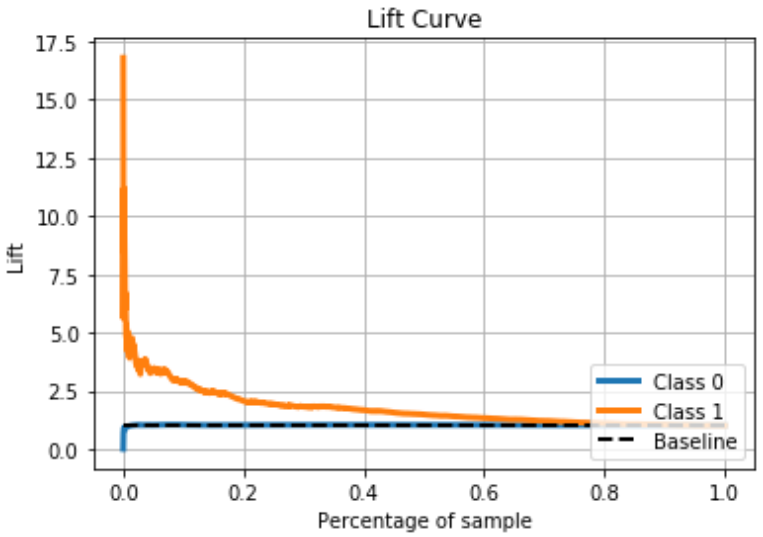


Lift Charts

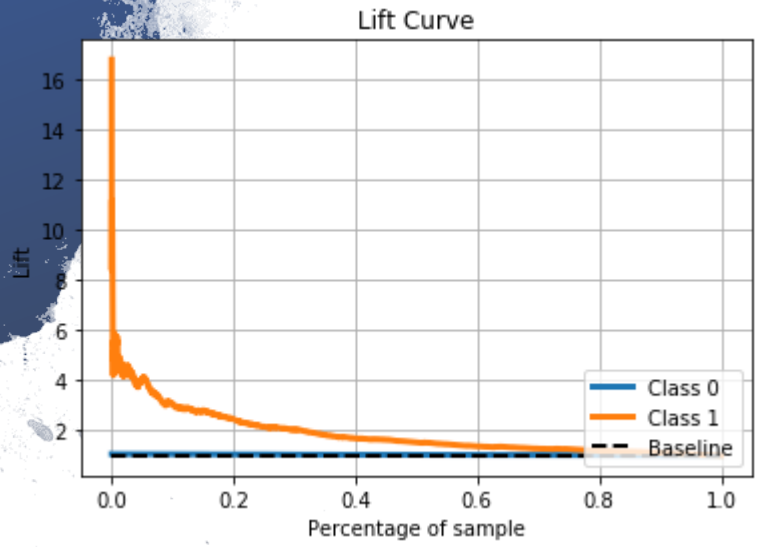
Decision Tree



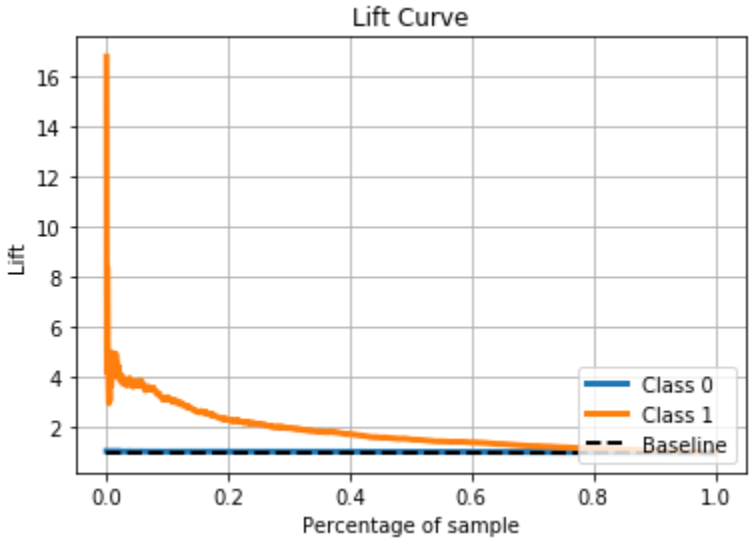
Random Forest



GBM



Logistic Regression



Future Considerations

- Addressing Imbalance of Data
- Experiment with other models