Predicting Potential Customers for Caravan Insurance

Capstone Project 1 Report

Introduction

Selling insurance is not easy, particularly if one is selling a niche product. What if one can figure out the right customer segments to market to? Would it be more optimal to market to the right crowd than indiscriminately selling to each person?

Then the core questions are: 1) Who or what customer with which characteristics have a higher likelihood to buy caravan insurance? 2) If we can identify the right customers to market to, how should we approach them? 3) What could be the ROI or efficiency gain on such a campaign?

Client

If I were to be hired as a consultant, I expect insurance companies to be my major clients. Automobile Insurance market is typically highly competitive and more often than not price driven. Thus, if a company can identify a niche space, that company may potentially add value to the book of business by 1) more efficient cost structure or 2) gain market share.

Data

Potential Approach

Typically, I start with studying the variables, and its definitions. Then I plan to look at descriptive statistics, such as correlations, mean and variance. I could also get a bird's eye view on how each variable may correlate via some N x N variable plots.

Depending on how the data looks, I may or may not group certain variables where data may be thin for a range, such as people who are age 65+ tend to be less represented in a company's book of business.

Potential Model Choice

1. Logistic Regression - create a marketing rating based on selected variables
2. CARTs
3. Discriminant Analysis?

Data collection and wrangling

The data is from a previous Kaggle posting and is owned and supplied by the Dutch datamining company Sentient Machine Research and is based on real world business data. The data itself are well organized and already cleaned, so there is not much for cleaning.

The URL to the page is: https://www.kaggle.com/uciml/caravan-insurance-challenge.

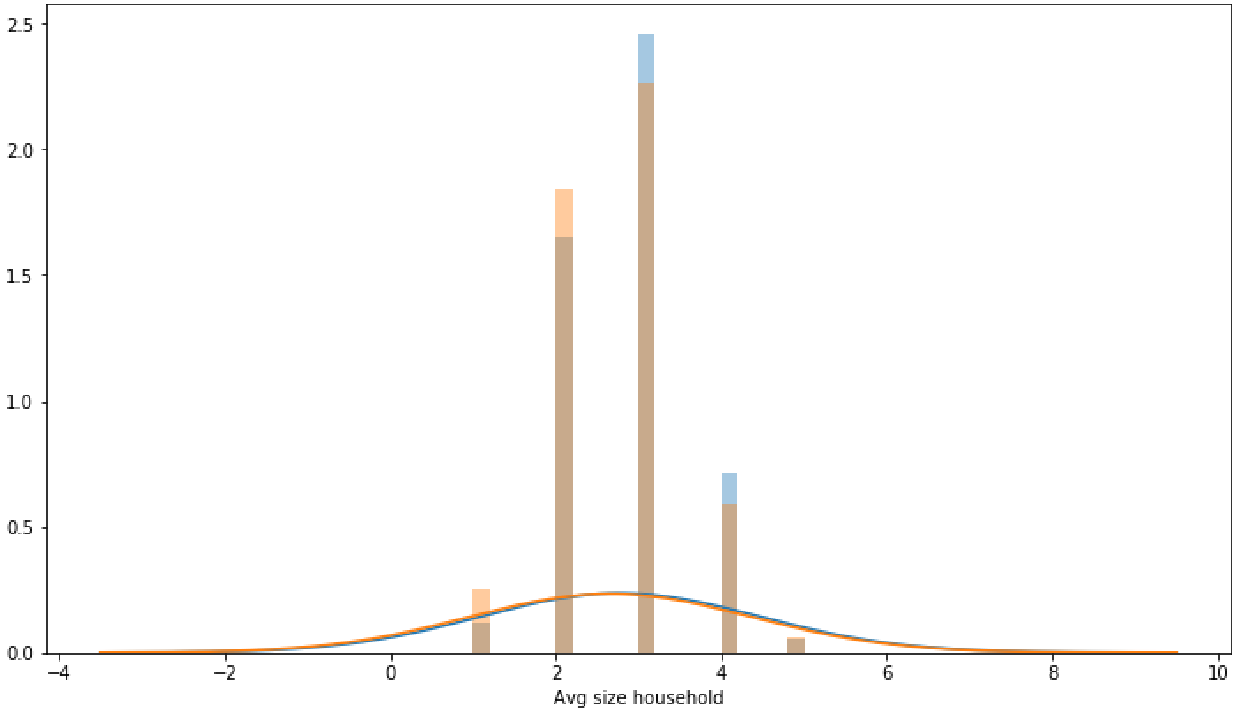
The data dimension is 5822 observations and 85 variables. A lot of the variable are banded, which in effect become discrete. Important thing to note is that each observation is looking at a zip code. Also, the data is imbalanced, which means that we have significantly more observation with no caravan insurance than those with, because in part not everyone own caravan. Therefore, the need for caravan insurance is a niche area. Also, each column would refer to a key table.

While the data is clean, the column names are in a different language, presumably Dutch. Thus, I remapped the column names to English version for clarity. Secondly, I split the data into training and testing set. The training set is what I will build the models on and test the performance on the testing data set.

Exploratory Data Analysis

Once the data is ready, the first step is to have an idea about the data itself. I dive in with visualization and describe the data with measure such as means and charts.

Using histograms, I can see how the data is distributed for each variable. The following is average size of household in each zip code:

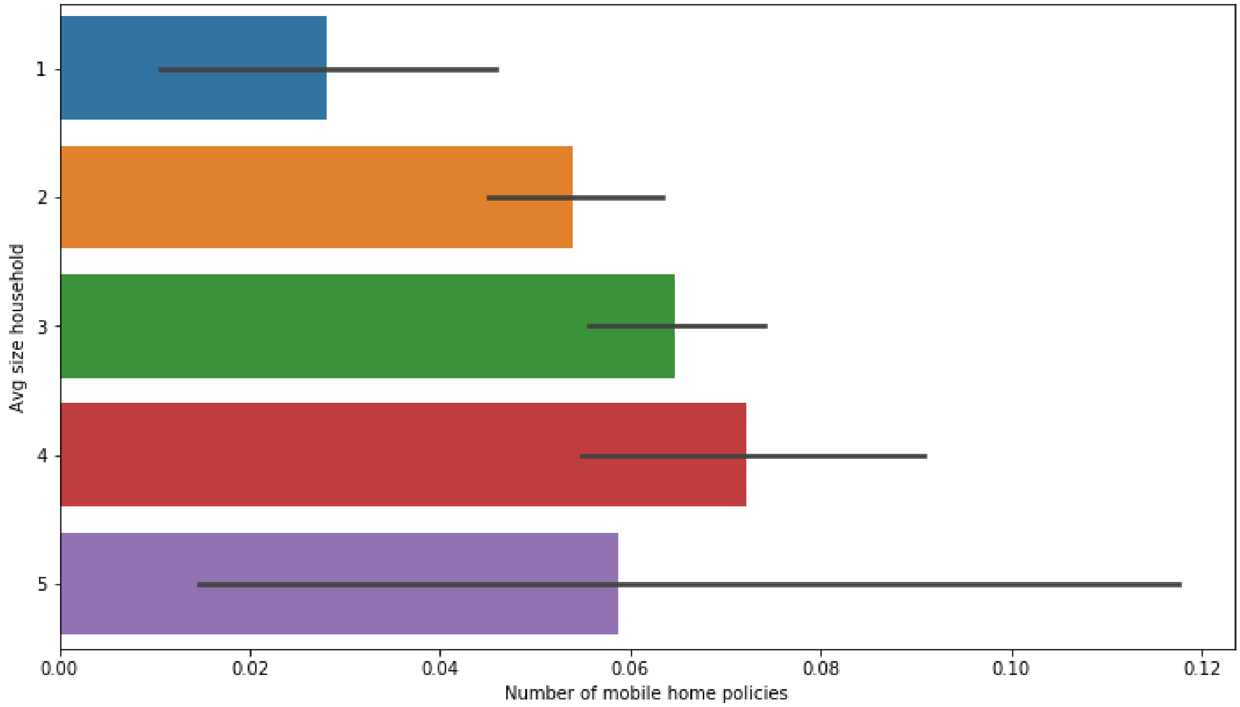


From the Histogram above, we can see 2 hues. The blue is bars shows the histogram of those observation with no Caravan Insurance and the orange one does. This histogram has been standardized, meaning they compared on the same scale after adjusting for the record size. We see they are close. This variable may visually suggest that this variable may not be predictive

The table below summarize the statistic information. First column ‘Count’ tell use how many observations. ‘Mean’ tells us the observed average rate of observation has caravan insurance. ‘Std’ is the standard deviation, which is a measure to give us a sense on how ‘spread out’ from our mean is. Then the ‘Sum’ shows how many of those zip code has caravan insurance.

Avg Size Household	Count	Mean	Std	Sum
1	284	0.0282	0.1657	8
2	2131	0.0540	0.2260	115
3	2646	0.0646	0.2459	171
4	693	0.0722	0.2589	50
5	68	0.0588	0.2370	4

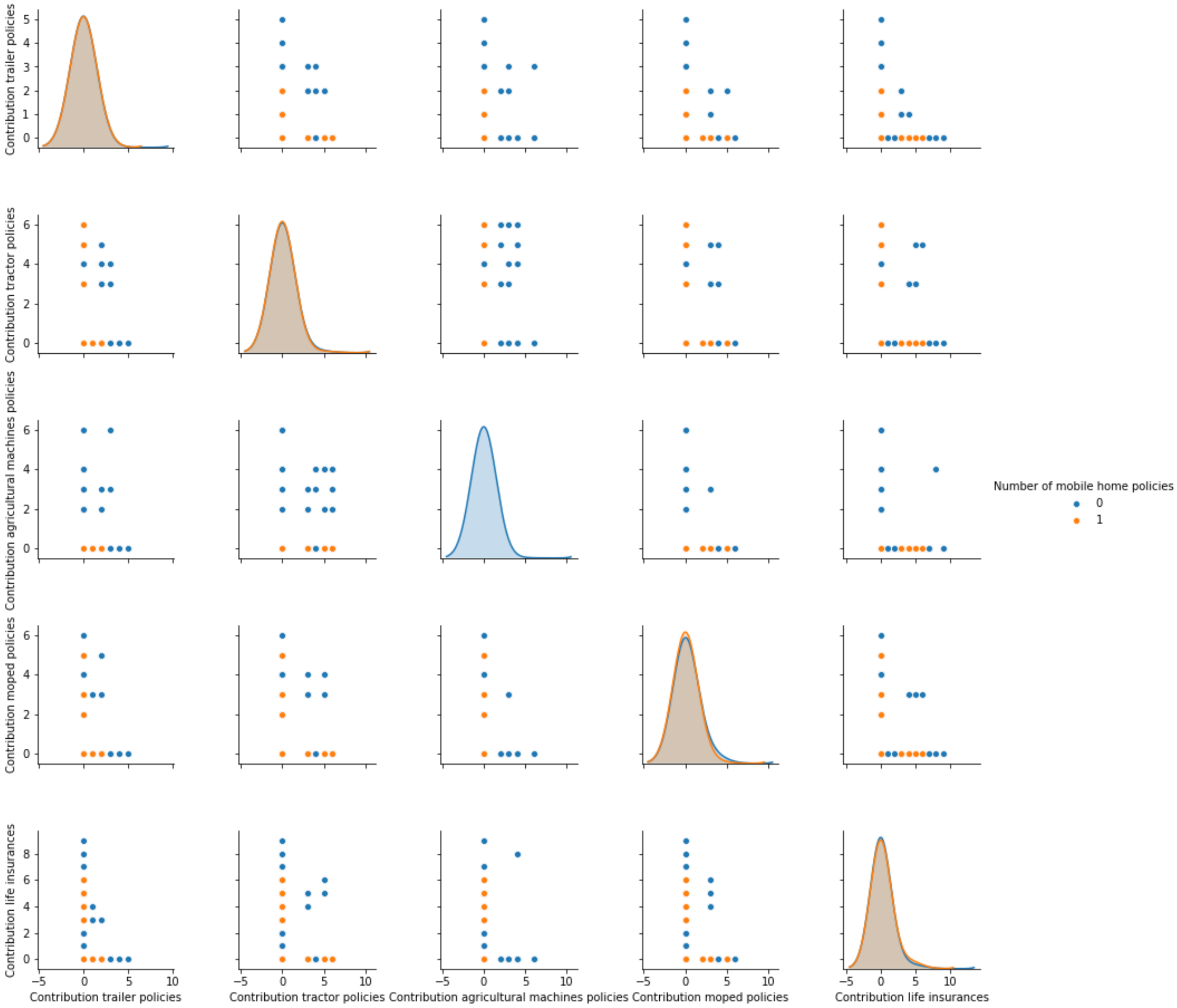
The chart below visualizes the mean column. This is to also visually see if the mean measure varies by the classification of this variable. Briefly it seems there are some differentiation between different household size. The caveat here though is that some classification has relatively small records, and thus may not be credible.



All these three tables and charts are applied to the 85 variables. Based on my EDA, there was not obvious evidence of separability.

The next part of my EDA, I also looked at how each independent variable may have correlate with one another. This is to prevent multicollinearity. This is important to in a way not double count on the effect. My first attempt is to look at the pair plot.

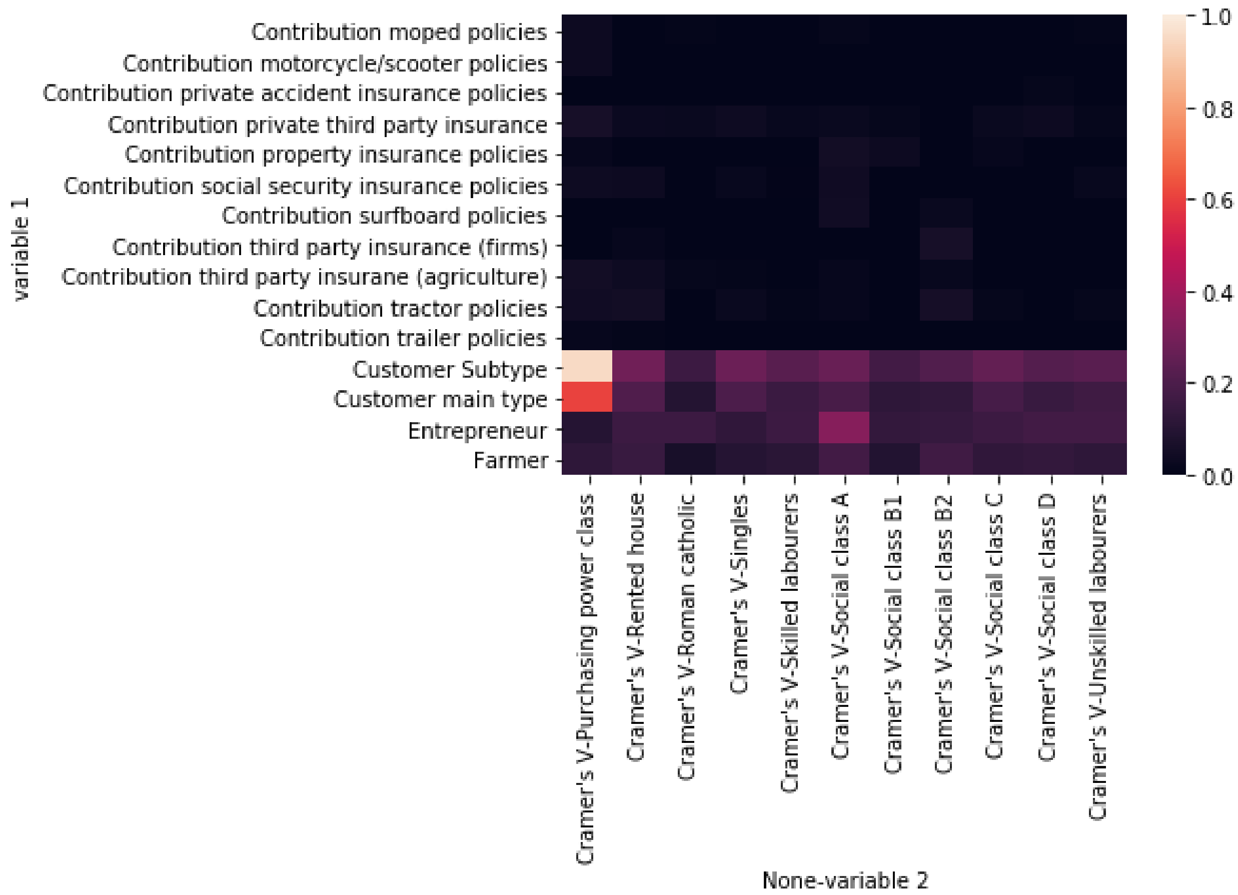
A sample as below:



By looking the pair-wise scatter we are trying to find out if there is a pattern of a spread. This help me see if we should avoid have certain combination of variables being together when building models. Luckily, in the example above, we do not seem to see as recognizable pattern suggesting collinearity.

Furthermore, I visualized the collinearity effect with heat map using Cramer's V. Normally, when we look at covariance and correlation, we use Pearson's. However, since my target variable is binary result, the Cramer V has been suggested as a more suitable method.

An example is below:



The heat map quickly identifies the pair with hi correlation. The darker the cell is the less correlated and vice versa. It seems like that we are not having much problem with collinearity and majority of variable pairs are close to black.

After the EDA, I dive deeper into the data with statistical inferences. Since this project is a classification problem and the target variable and most of the independent variables are categorical, I felt that the chi-square test for between each independent variable and the dependent variable.

The null hypothesis (H0) is that there is no difference between the proportion of the zip code that has caravan insurance for each variable than the overall population, without discerning specific variable(s).

The alternative hypothesis (H1) is that there is a difference.

The formal expression is:

H0: U_xi = U_y

H1: U_xi != U_y

I leverage the Sci-Kit Learn built in Chi-Square Contingency table to get the Chi Square statistic and calculate the P-value from a chi-square distribution. I do it one variable against the target variable at a time using a loop.

Below is the chi-square test result for the first 9 variables, out of total 85.

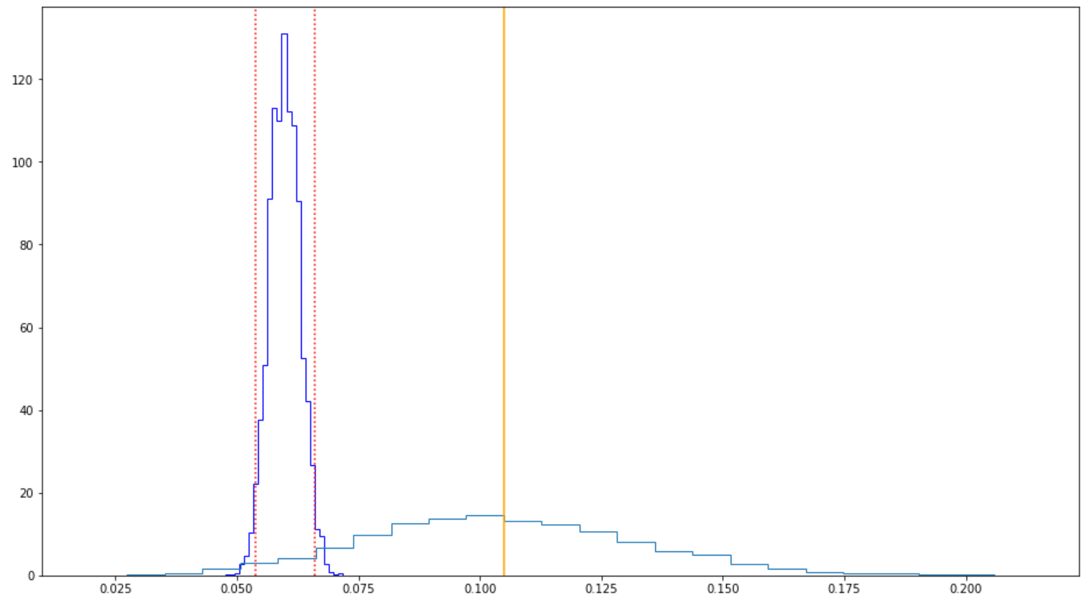
variable name	chi square stat	p-value
Customer Subtype	124.81	0.001
Number of houses	3.46	0.9999
Avg size household	9.33	0.5015
Avg age	3.29	0.9931
Customer main type	88.66	0
Roman catholic	9.12	0.9815
Protestant	22.4	0.3192
Other religion	11.74	0.4671
No religion	19.4	0.4959

From that table, we may infer that the variable ‘Customer Subtype’, for example, can be statistically significant, because the p-value is significantly smaller than 5%.

Bootstrap Inference

Beside from the Chi-Square test, I also applied bootstrap inference as my second approach. The main idea here is to compare the overall mean or proportion of the sample population that has caravan insurance.

As an example, here is a visual example.



The dark blue histogram is the bootstrap distribution of the mean and the red vertical lines are the boundaries for the 95% confidence intervals of the mean. The orange line is the mean of the subset of the first independent variable with classification as 1. The first variable has 42 different classes. The light blue line is the bootstrap distribution of the subset. From this chart, we may conclude that the mean of this subset seems significant, though the variance may be quite high.

Results and In-depth analysis using machine learning

The next phase of this project is to apply machine learning techniques to build predictive models on the Caravan Insurance data set. The following are an overview of the steps I took to analyze my data.

First, I narrow down the variables I am to model on by the P-value I did for the chi-square test. I narrow down from 85 variables to 28 variables, which saves time on building and to tune the hyper-parameters.

My starting point is for this section of this project is continuing from the result obtained from Statistical Inference. From our chi-square test, we may narrow down what variables I think could be more relevant. The selection criteria include the variable if its chi-square p-value is less than 5%, which means this variable statistically can have a meaningful difference from the average.

The data set has already split into training and test data set in the previous section of this project. I will build models from the training data set and test the models' performances using the testing data set.

Second, I list down possible models for my data. Since my target variable is categorical, I felt CART, Random Forest, GBM, and Logistic Regression seem to be the best possible fits. The problem is a classification problem. Initially, I took the models as is and apply them without tweaking the hyperparameters.

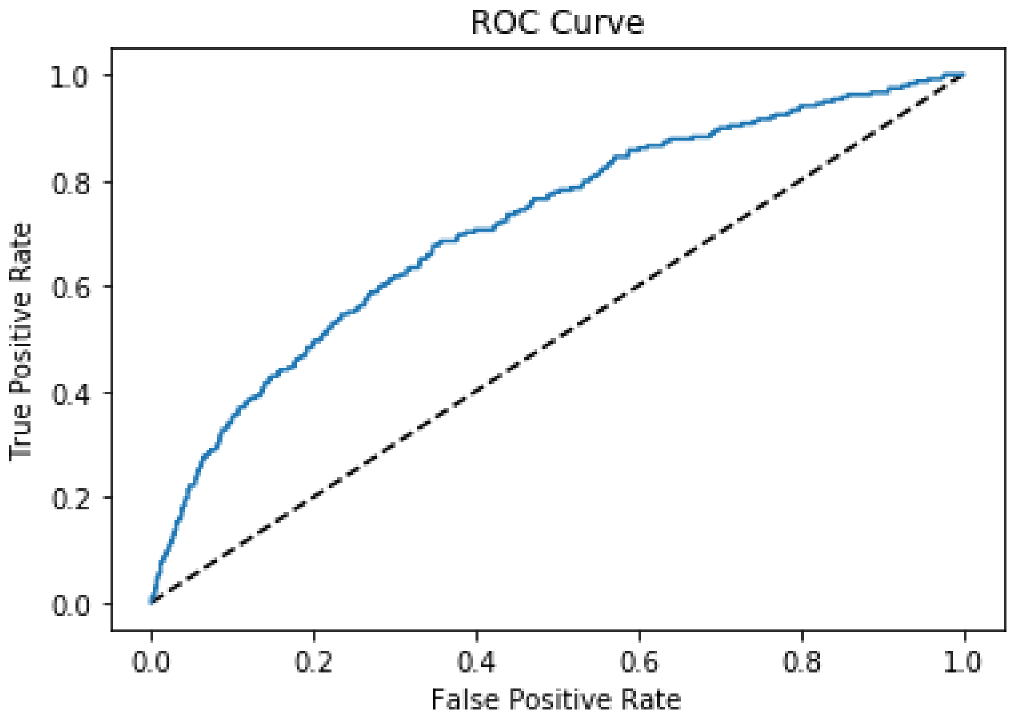
To evaluate how my model is performing, I turned to use the confusion matrix initially. The three metrics I am most interested in the confusion matrix are accuracy, precision, and recall. Accuracy shows how much of the sample classified correctly, percentagewise. This metric shows the model doing well in both training and test data, but different look into precision or recall indicates that the model's failure at predicting true positives.

The following is an example of that from my logistic regression model:

```
      0  1
0  5470  4
1   346  2

 accuracy:      0.94
 precision:     0.333
 recall:        0.006

      0  1
0  3759  3
1   235  3

 accuracy:      0.94
 precision:     0.5
 recall:        0.013

CV Score: 0.9395
```

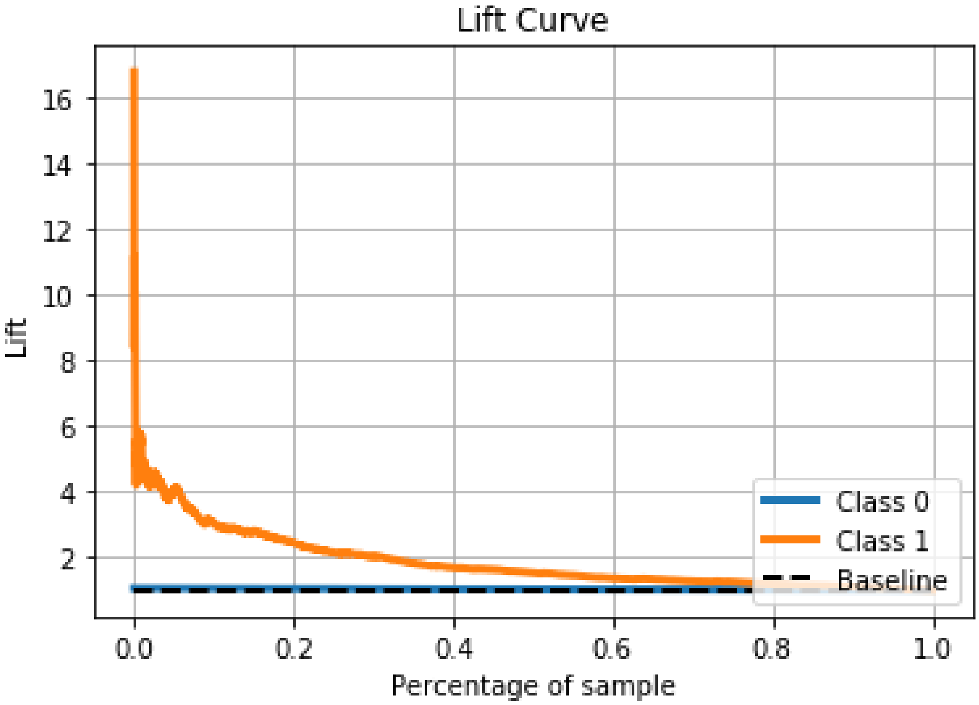
This phenomenon is due to an imbalance in the data that I mentioned early in the report. We see a significant proportion of the sample population does not have caravan insurance than those who do. Therefore, the models generally are great at predict who wouldn't buy caravan insurance and poorly at those who might be interested.

Another perspective is to look at the ROC and AUC. ROC is a probability curve, and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The AUC stands for area under the curve, which goes from 0.5 to 1.0, where 1.0 is perfect and 0.5 is no different from taking a random observation and thus not effective.



I also observe the lift and cumulative gains chart. Lift charts measure the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model. This helps us to see which proportion of the population only needed to be contracted for soliciting caravan insurance. Ideally, the implication of these measures is to find the most gain for the model with the least amount of customer needed to be contacted, thus getting the most bang for your buck.

i.e Lift chart for Logistic Regression



The fourth step is to fine-tune the models using grid search or random search on the hyperparameters. I mainly use grid search try to get some improvements of these models. Unfortunately, the gain is minimal so it might not be worth the time spent to further optimize the model. Perhaps other projects could have different results. Also depending on the ML algorithm, the time spent to do grid search may not be worthwhile as these processes could take a long time just to yield a minor improvement.

From this exercise we can find some way to model our data. However, the result is far from ideal. The best performing model has an AUC of 0.713, meaning that there a lot of space for the model, if possible, to improve. If I can devote more time, I will attempt to address the imbalance in the data between positive and negative response in our target variable.