Identifying High Risk Motor Insurance Customers Through the Use of Data Visualisation and Machine Learning.

Luke Birkett

Words: 2000

Objective and Project Overview

The aim of this project is to present a methodology for identifying the customers that may be at an enhanced risk of filing a claim during their policy. The ability to effectively identify high risk customers is imperative for accurate pricing of policies and profitability. Claims are an expense to insurance companies, thus, to consistently under-price high risk customers will almost definitely lead to a net loss over time. Alternatively, to over-price low/lower risk customers will drive potential customers away to cheaper policies in the market.

This project starts with a variety of variable focused visualisations to provide contextual insight. Following this, a random forest algorithm is tuned to the data with the goal of providing each customer with a probability that they will launch a claim during their policy. The probabilities generated are evaluated against a naïve baseline claim rate given by the percentage of claims found in the dataset, i.e every customer is equally as risky. Finally, the results (probabilities) are presented in boxplot visualisations in order to explore whether there are trends within each variable even after all other variables have been considered.

1 Introduction

1.1 Data Sources and Description

The data used for this project was obtained from the 'CASdatasets' package in R. This package holds a number of datasets, originally used in conjunction with the book 'Computational Actuarial Science with R' by Arthur Charpentier. The data used here is from the `ausprivauto` library which holds 3 automobile claim datasets for Australia where third-party insurance is compulsory. The particular dataset used within this library is the `ausprivauto0405` which holds information on 67,856 policies of up to 1 year in length for the years 2004 to 2005.

Table 1. Variable descriptions and data types.

Variable	Variable Type	Description
Exposure	Numerical	The number of policy years.
VehValue	Numerical	The vehicle value in thousands of AUD.
VehAge	Factor	The vehicle age group.
VehBody	Factor	The vehicle body group
Gender	Factor	The gender of the policyholder.
DrivAge	Factor	The age of the policyholder.
ClaimOcc	Integer	Indicates occurrence of a claim.
ClaimNb	Integer	The number of claims.
ClaimAmount	Numerical	The sum of claim payments.

Table 2. Head view of the data.

Exposure	VehValue	VehAge	VehBody	Gender	DrivAge	ClaimOcc	ClaimNb	ClaimAmount
0.303	1.06	old cars	Hatchback	Female	young people	0	0	0
0.648	1.03	young cars	Hatchback	Female	older work. People	0	0	0
0.569	3.26	young cars	Utility	Female	young people	0	0	0
0.317	4.14	young cars	Station wagon	Female	young people	0	0	0
0.648	0.72	oldest cars	Hatchback	Female	young people	0	0	0
0.854	2.01	old cars	Hardtop	Male	older work. People	0	0	0

Table 3. Continuous variable summary statistics.

Statistic	Exposure	VehValue	ClaimOcc	ClaimAmount	ClaimNb
Min	0.002	0.000	0.000	0.00	0.000
1st Quartile	0.219	1.010	0.000	0.00	0.000
Median	0.446	1.500	0.000	0.00	0.000
Mean	0.468	1.777	0.0681	137.30	0.07276
3rd Quartile	0.709	2.150	0.000	0.00	0.000
Max	0.999	34.560	1.000	55922.10	4.000
Variance	0.0841	1.452	0.064	1115765	0.077
SD	0.290	1.205	0.252	1056.29	0.278
Sum	31800.82	120581.5	4624	9314604	4937

1.2 Descriptive Statistics and Exploratory Analysis.

1.2.1 Exposure

In this dataset, exposure refers to the length of the policy. Table 3 shows that policy lengths range from the given hours of a working day to the entire year. The statistics imply there is a fairly even distribution of policies throughout the year, however, Figure 1 shows there to be a huge spike at the one-year mark, though this is to be expected as most personal policies tend to be yearly. Longer policies mean more time on the road so this variable can be expected to show up as a strong indicator of claims. This is supported by Figure 2 which shows the long policies to have a higher claim rate. It should be noted that this variable is not the main focus of this project, instead the characteristics of the driver and vehicle are more interesting. That being said, this is a useful variable for the modelling stage as it removes a lot of noise from the dataset, i.e. without this variable the random forest algorithm would likely attribute policy claims to certain prevailing characteristics when in fact the claim was a by-product of simply being on the road a lot.

Figure 1. Policy lengths distribution.

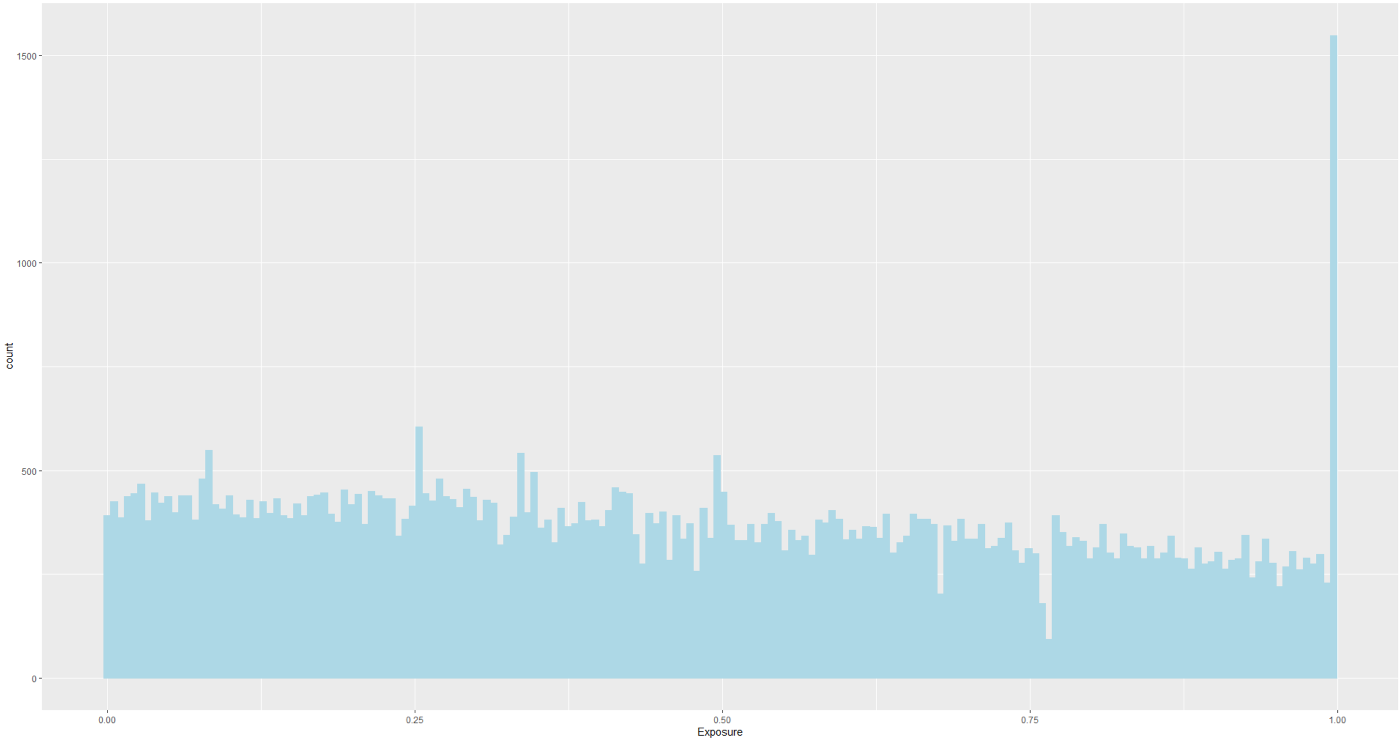
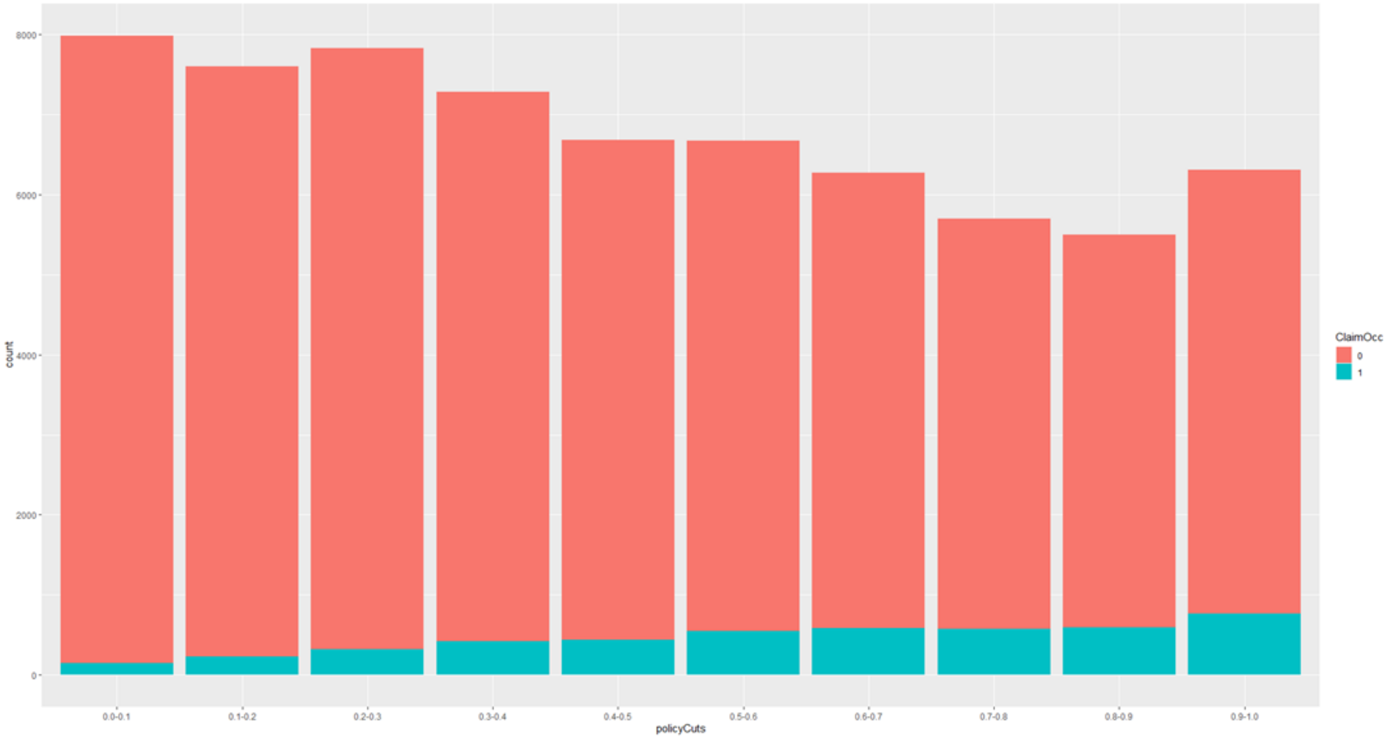


Figure 2. Policy length against claim occurrence.



1.2.2 Vehicle Value

The vehicle values range from worthless (0) to 34.560. It is not entirely clear what this scale is but from Figure 3 we can see that the data is heavily right skewed. This is further represented in Table 1 where the 3rd quartile cut-off is only 2.15. Anecdotally, I expect vehicle value to effect claim rate in two ways. Firstly, cheaper cars may have a lower claim rate due to the likelihood that repairs may outweigh the cost of the vehicle itself. Conversely, more expensive cars may show a lower claim rate as drivers may be more cautious due to the cost of the vehicle. Figure 4 shows a histogram of the vehicle value on a log scale. There appears to be a spike in the centre and a flattening in the extremes, however, this is in accordance with the overall distribution of the data so no clear hypothesis can be drawn at this stage of the analysis.

Figure 3. Vehicle value distribution

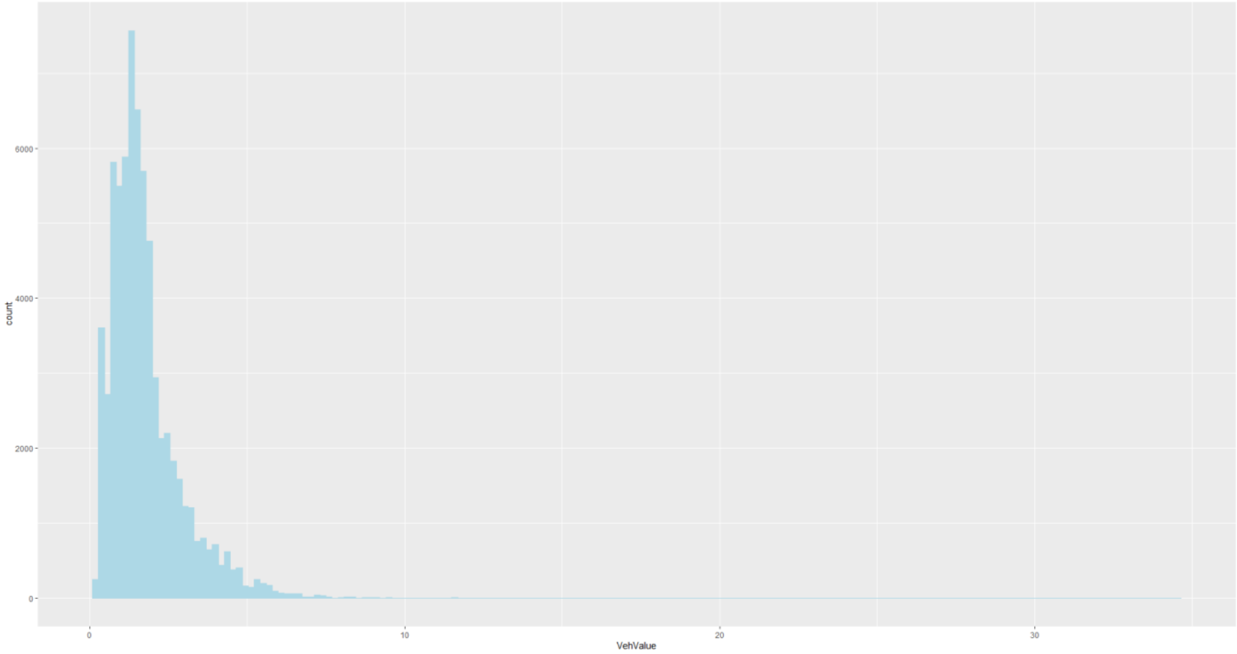
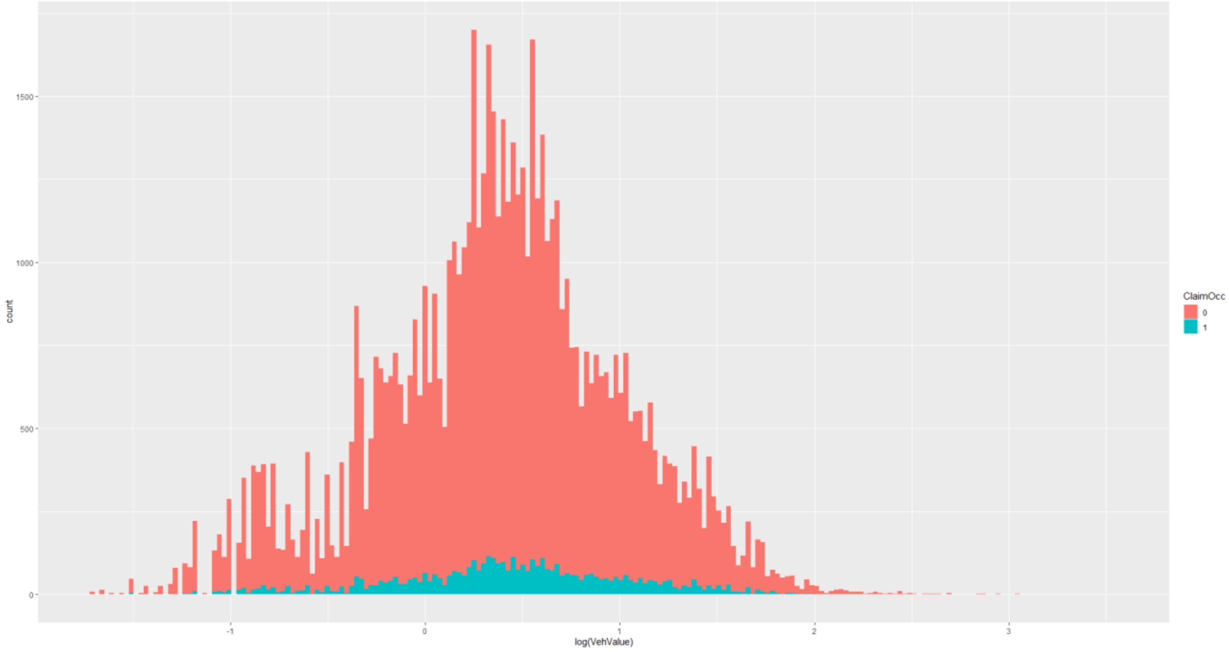


Figure 4. Vehicle value claim occurrence histogram log



1.2.3 Vehicle Age

Conceptually, it would be intuitive to assume that vehicle age would behave the same as vehicle value in terms of claimant risk due to the correlation between cost and age. However, in this dataset vehicle age is a categorical variable split into: youngest, young, old and oldest cars. Figure 4 shows the distribution of each category along with the claim counts. It's not easy to infer any information from the graph due to the differing heights of each bar. Figure 6 presents the data in proportional bars. At first glance their appears to be little variation between the categories. That said, the rates for the oldest cars are lowest and the youngest cars have the second lowest claim rates, indicating some sort of trend.

Figure 5. Vehicle age claim occurrence counts

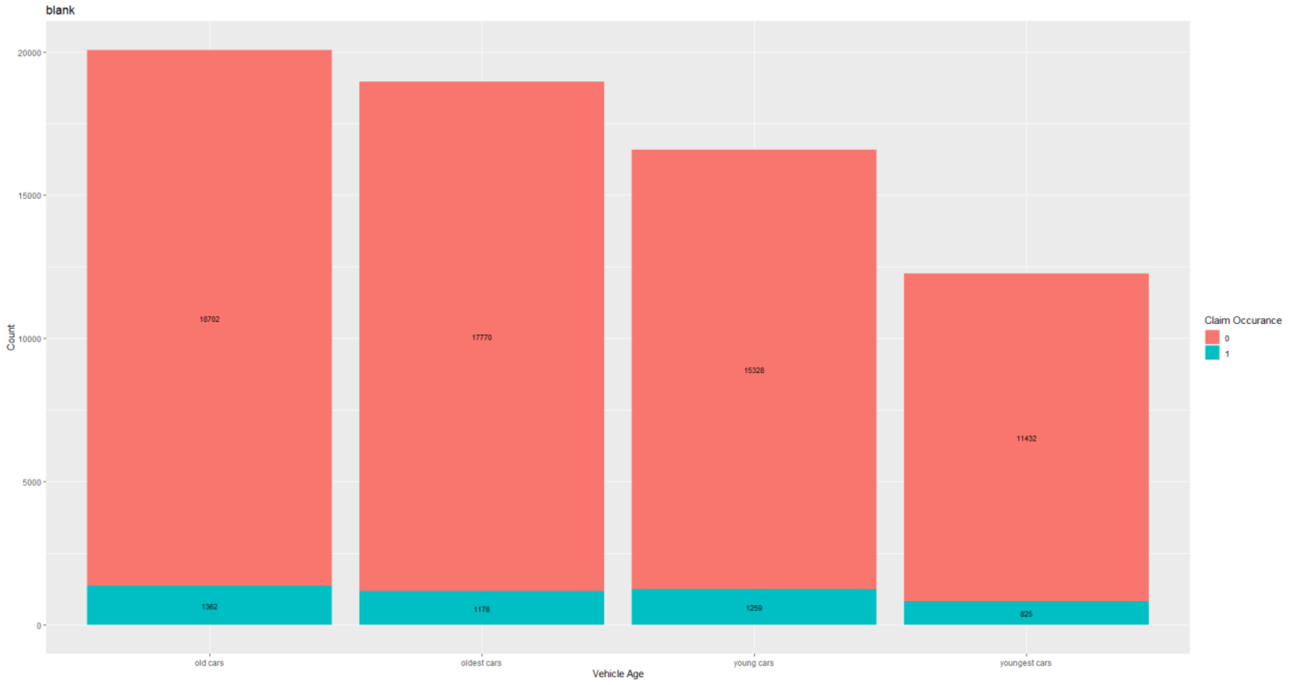
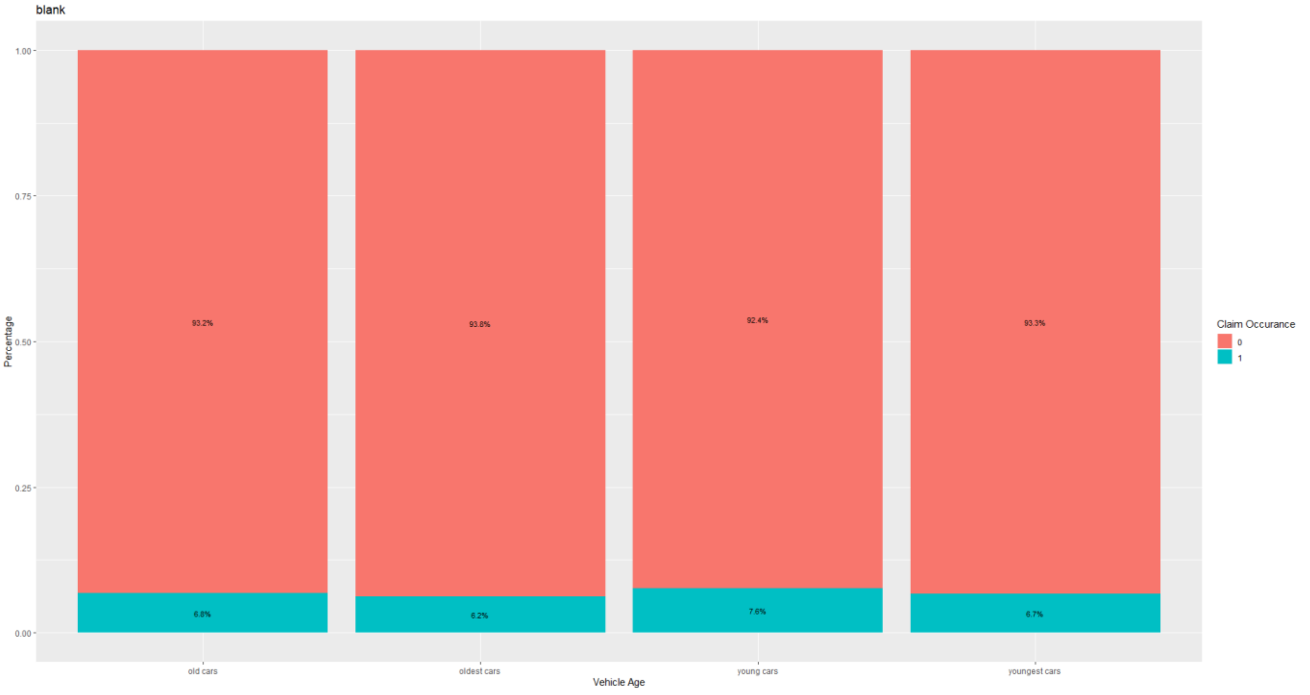


Figure 6. Vehicle age claim occurrence rates



1.2.4 Vehicle Body

The type of vehicle driven is often an indicator of the drivers characteristics. For example, young people tend to drive hatchbacks as they are cheaper and easier to drive. Convertibles tend to be expensive, luxury items thus driven by wealthy individuals. In this dataset there are 13 categories as shown in Figure 7. An issue with this data is the sparseness of some of the categories. Small sample sizes can lead to skewed results meaning we should be weary of drawing conclusions from these categories. From Figure 8 we can see that there is little variation between the most popular body types, hatchback and sedan, coming in at 6.7% and 6.8%. However, the station wagons, the 3rd most popular body type, have an occurrence rate of 7.2%, perhaps indicating larger work vehicles represent more risky policies.

Figure 7. Vehicle body claim occurrence

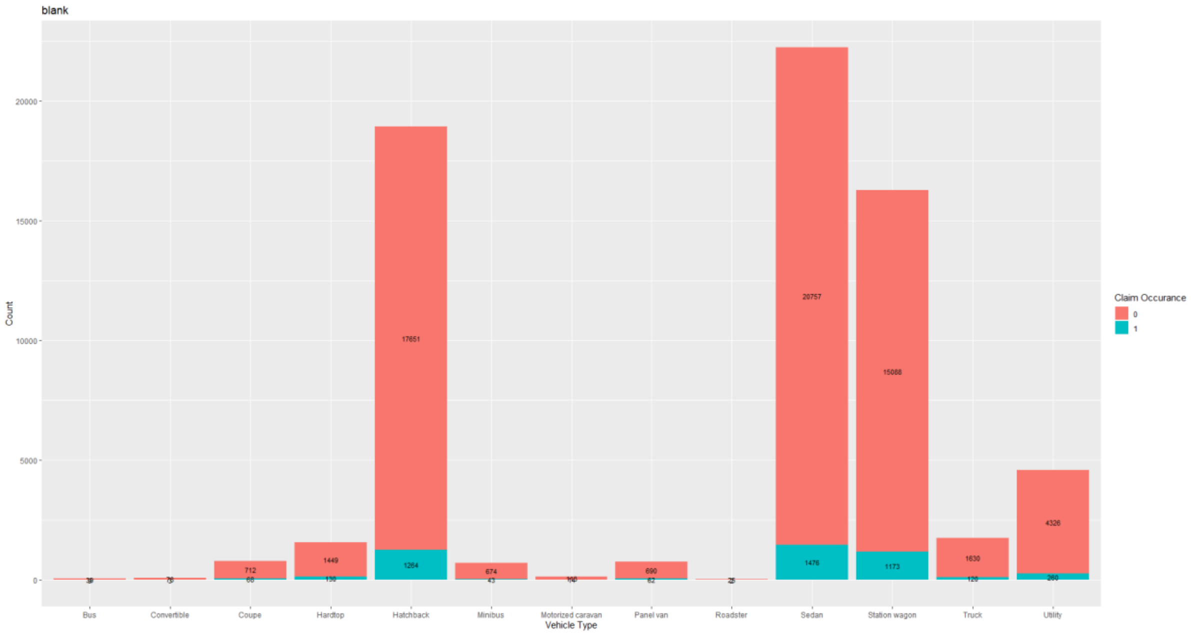
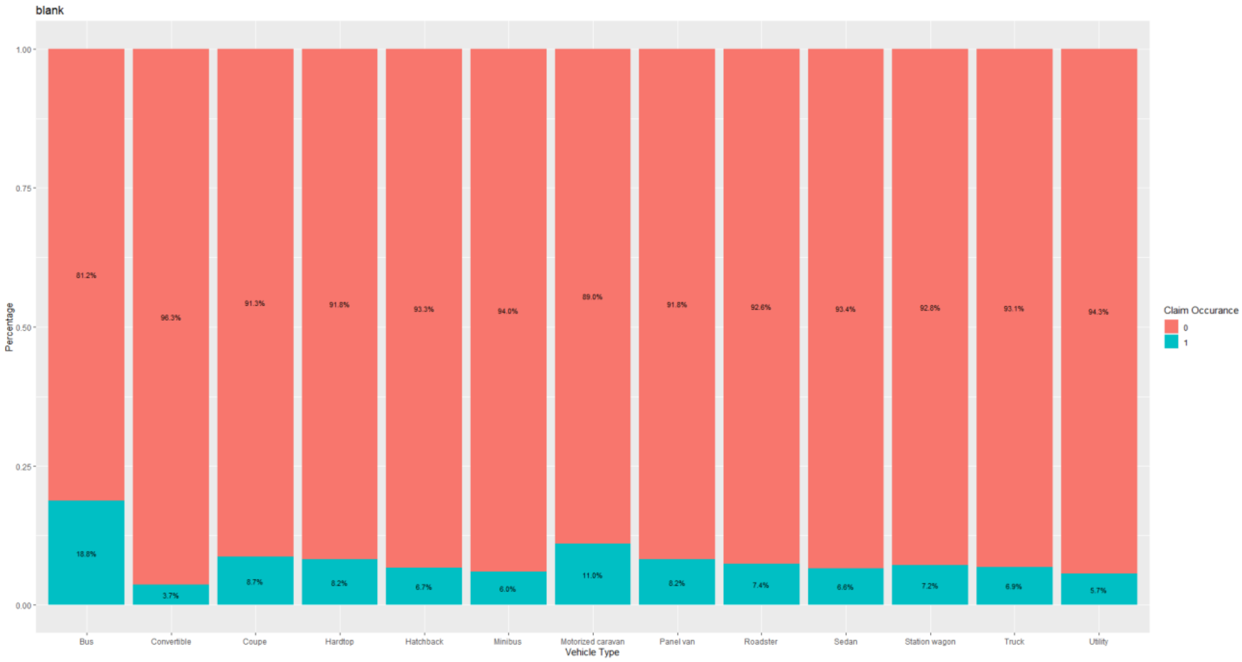


Figure 8. Vehicle body claim occurrence



1.2.5 Driver Age

Generally, younger drivers are considered to be of higher risk on the road due to their lack of on road experience. Figure 9 supports this with young and youngest people having the highest claim rates at 7.2% and 8.6%. Additionally, there appears to be a spike in occurrence rates for working people with general working people claiming at a 7.1% rate and older working people claiming at a 6.7% rate compared to 5.7% for normal older driver. This finding may support the hypothesis that vehicles associated with work are at a higher risk of claiming.

Figure 9. Driver age count

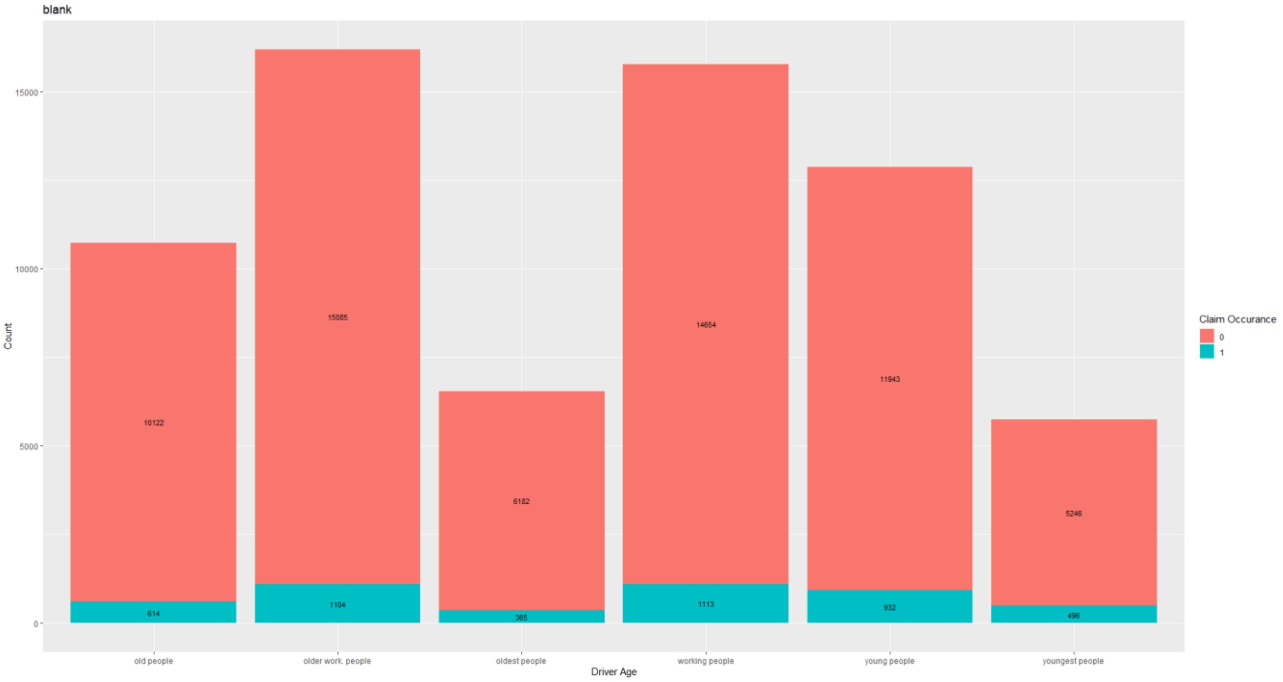
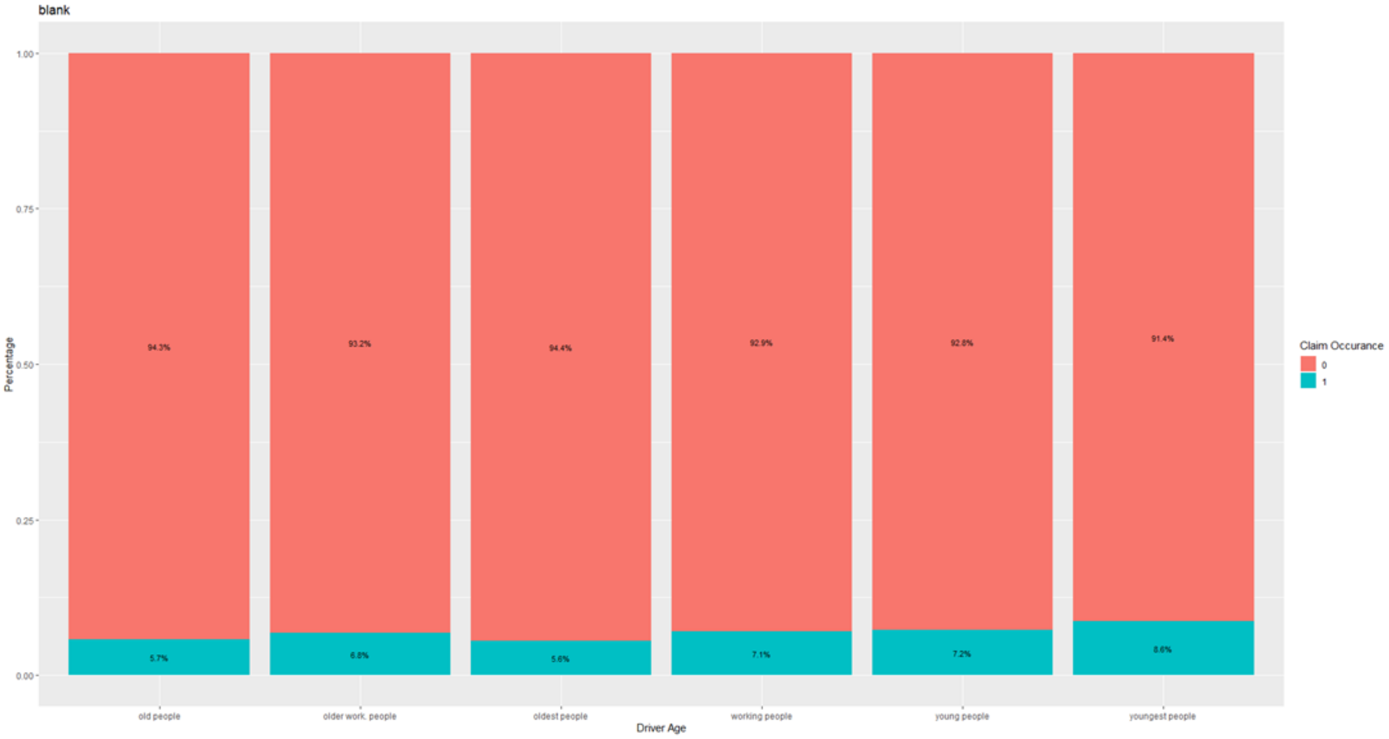


Figure 10. Driver age claim occurrence rate



1.2.6 Gender

With respect to vehicle insurance, gender is typically a topic of interest when considering younger, male drivers though not so much outside of this scope. In this particular dataset there are more female drivers than male, but Figure 11 shows the claim rates are almost identical.

Figure 11. Gender claim count

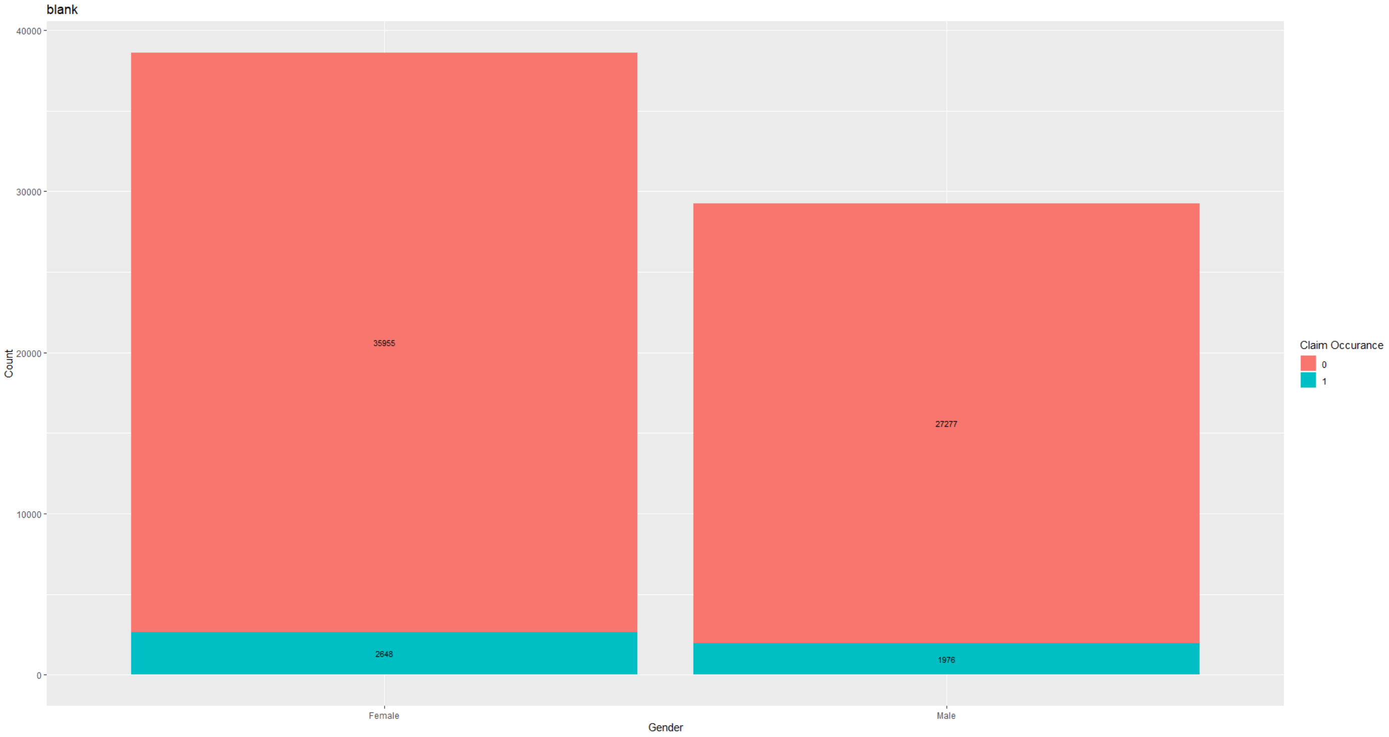
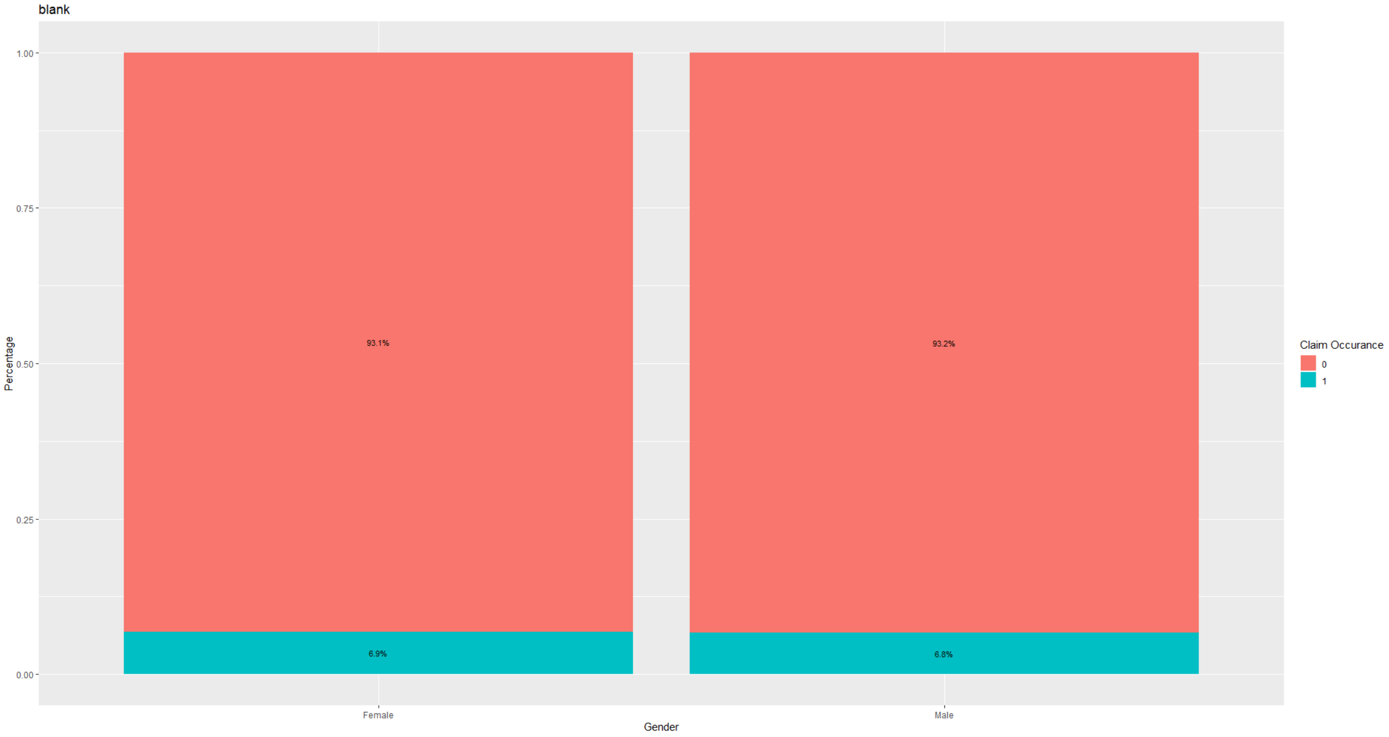


Figure 12. Gender claim rate



2 Random Forest Modelling

To assign each observation a probability (risk rate) I use a random forest classifier trained on 70% of the dataset and tested on the remaining 30%. I opted for this model because it is simple to understand and can handle variables that have a non-linear relationship with the outcome variable which I expect some variables may exhibit. To optimize this model, I tune two hyperparameters, mtry (number of variables used for each split) and minimum node size (depth of each tree), using the grid search method. I compare the results of my model to a baseline rate of 6.8% assigned to each customer and evaluate the results using the Brier Score (MSE). The results of the grid search can be seen in Table 3, the minimum error is obtained when mtry is 2-3 and node size is around 750-950. Importantly, the models MSE outperforms the baseline measure. For the final model I opted for an mtry of 2 and a node size of 500. I chose 500 as there is very little variation compared to a node size of 950 and this made run times a lot shorter. Figure 13 presents the tree tuning plots. Random forests cannot overfit from number of trees, but there becomes a point when the decrease in MSE per increase of trees is not worth the additional run time. The point at where this occurs is when the test MSE appears to flatline which is around 3000+ here. Finally, Figure 14 presents the variable importance. As expected, policy length is the most importance. Interestingly, value is the second most important by a long way. Followed by body type, driver age, vehicle age and gender. When applied to the entire dataset the model has an MSE of 0.0604 compared to a baseline of 0.0635.

Table 4. Hyperparameter grid search

Mtry	Node size	OOB MSE	Model MSE	Baseline MSE
2	950	0.0623	0.0623	0.0634
2	850	0.0623	0.0623	0.0634
3	950	0.0623	0.0623	0.0634
2	750	0.0623	0.0623	0.0634
3	850	0.0624	0.0623	0.0634

Figure 13. Tree tuning MSE plot

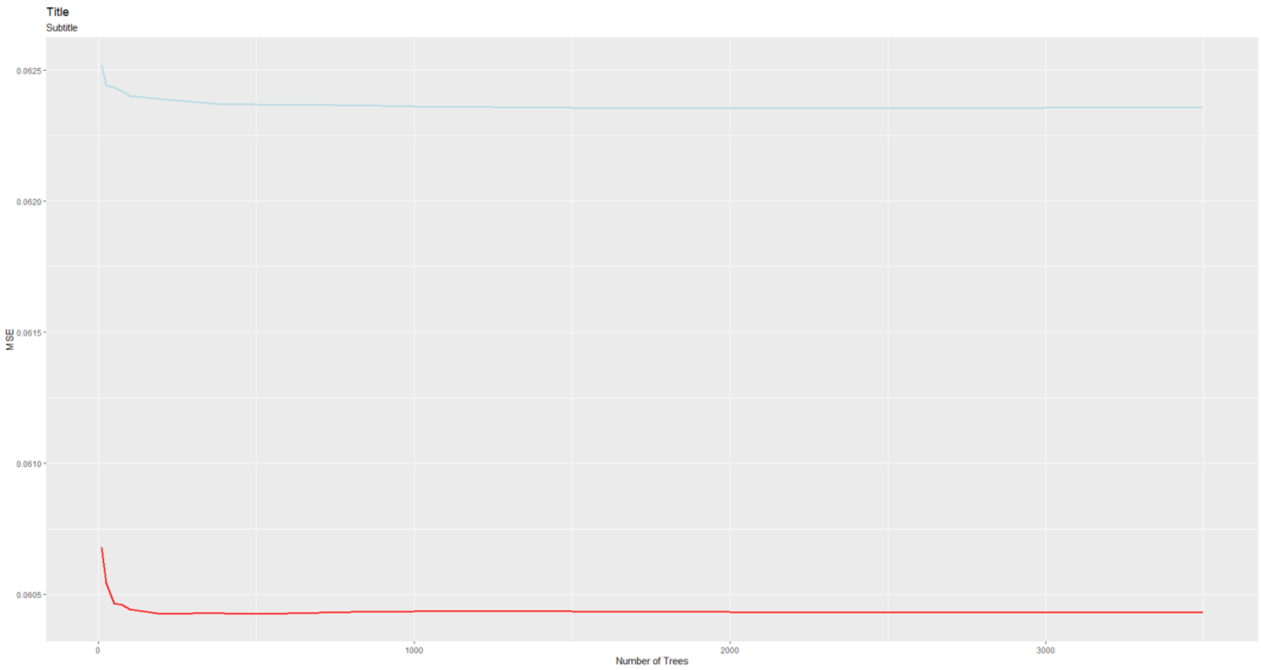
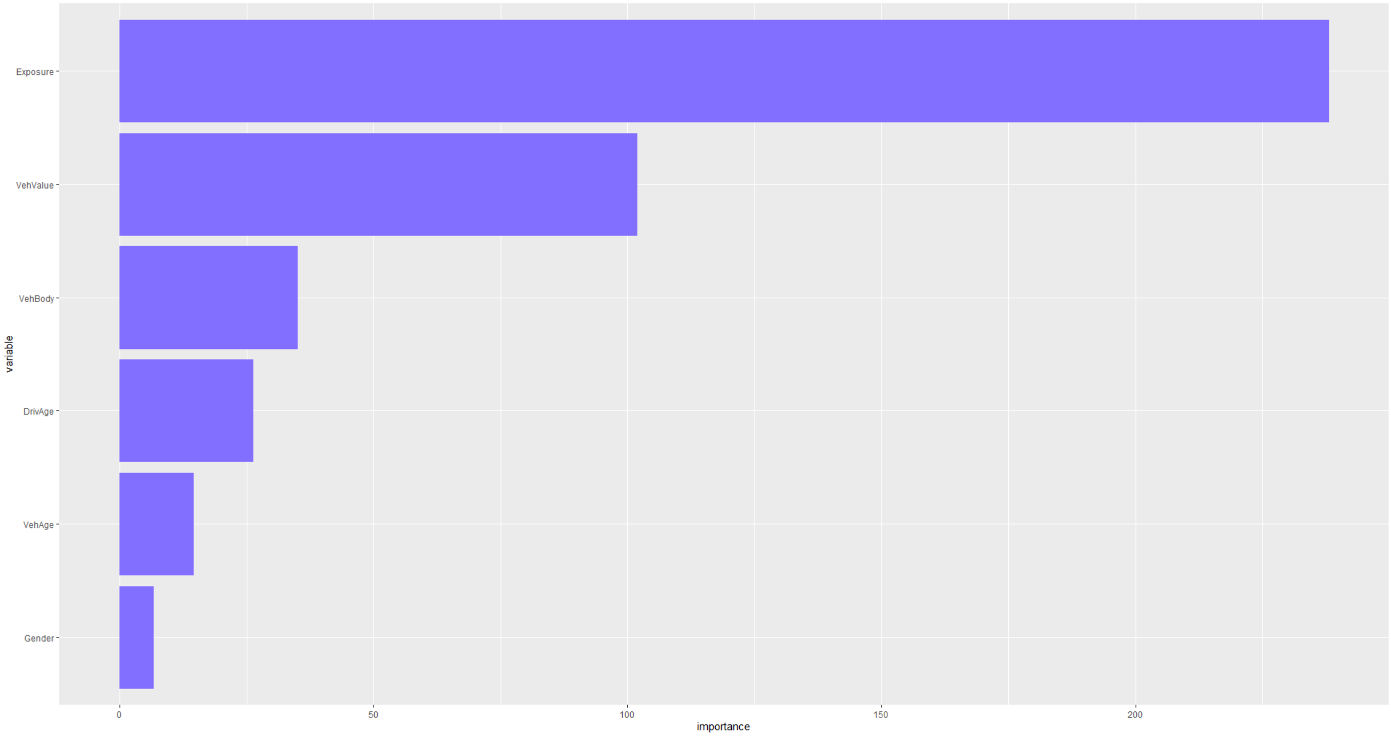


Figure 13. Variable importance plot



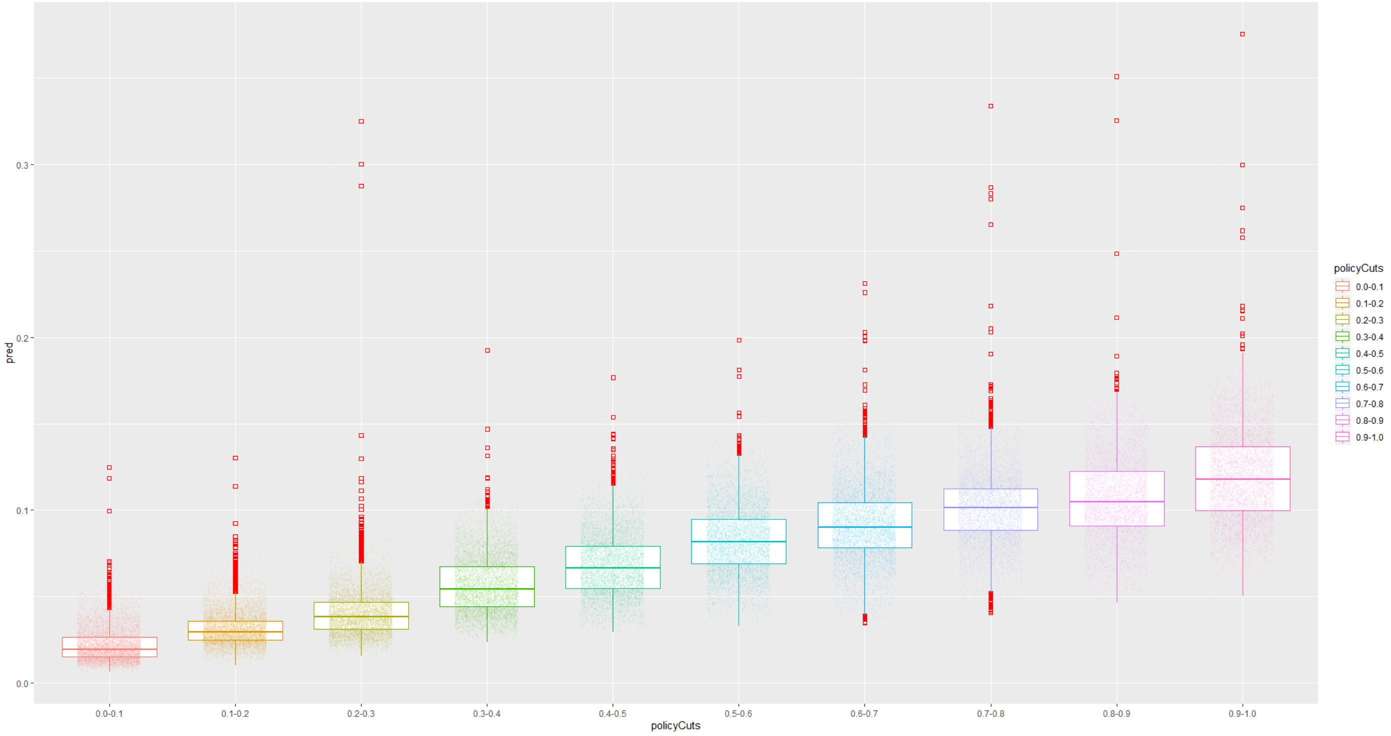
3. Risk Analysis

In this final section I will look at the risk distribution for each variable. As the risk probabilities consider all other variables. I will present this section in order of importance.

3.1 Exposure

Exposure exhibits a clear linear trend whereby risk increases with policy length. There are no surprises here.

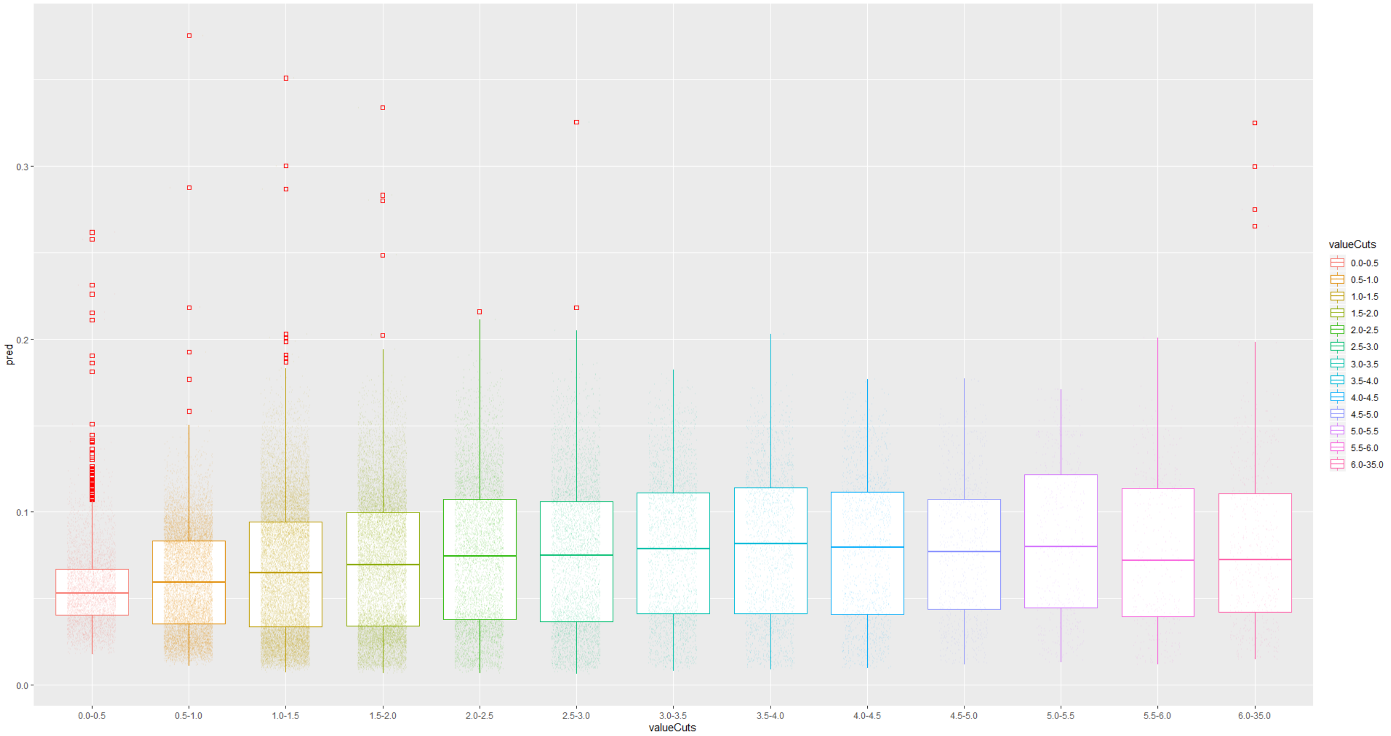
Figure 14. Exposure risk levels



3.2 Vehicle Value

The vehicle value being the second most important variable came as a surprise to me as I expected drivers age to be more important. That being said, the non-linear trend shown in Figure 14 follows intuition. Cheaper cars are not worth claiming, this is especially true for cars of value 0-0.5 as denoted by the tightness of the box plot, and expensive cars have a slightly lower claim rate than mid-level cars, likely due to the behaviour of the drivers.

Figure 15. Vehicle value risk levels



3.3 Vehicle Body

Once again, vehicle body is a variable I expected to be further down the importance ranking. Figure 16 shows a fair amount of variation between vehicle types. As noted earlier, caution should be taken with the variable as a lot of the categories have a weaker sample size which can be identified through the density of the jitter plotted on top of the box plots. A positive to note from this plot is that station wagons appear to represent more risky customers.

3.4 Driver Age

Driver age is the variable I expected the random forest to find more important. However, a positive to note is that Figure 17 clearly presents some sort of patterns to be seen. Younger drivers are in fact riskier, followed by workers with elder people less risky. It may be the fact that driver age is further down the list as it is highly correlated with vehicle value, i.e. older people have more expensive, safer vehicles.

Figure 16. Vehicle body risk levels

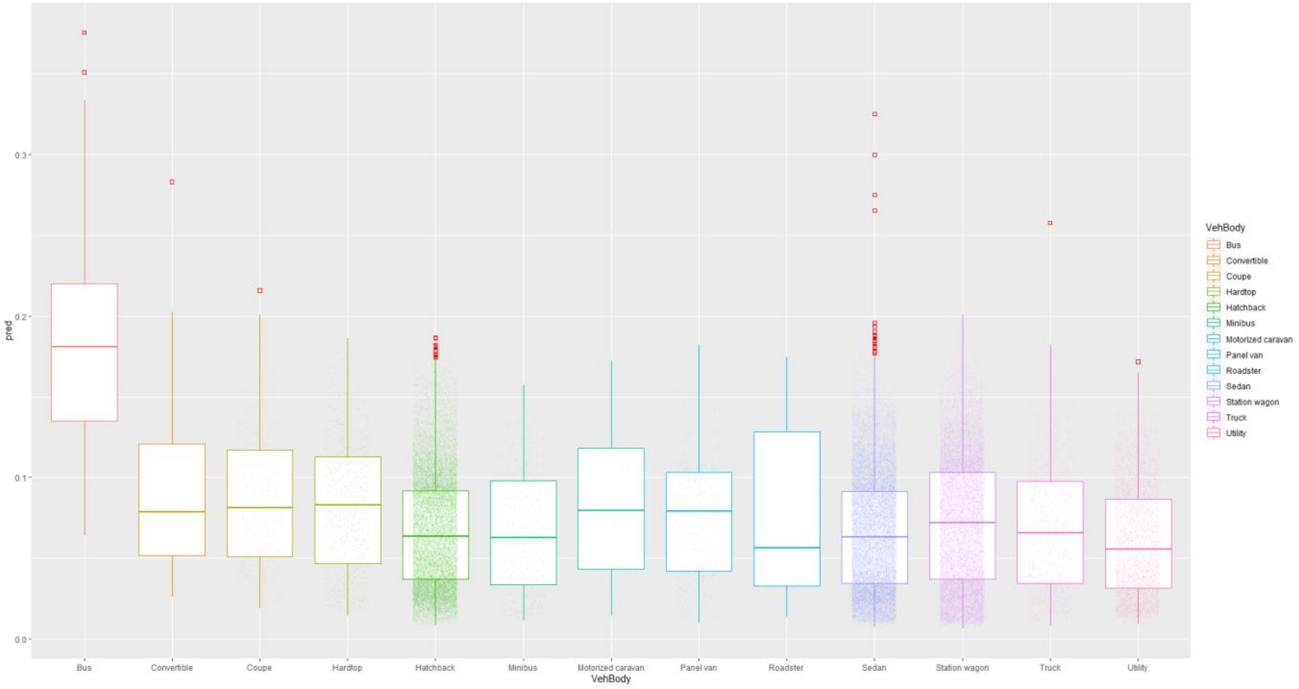
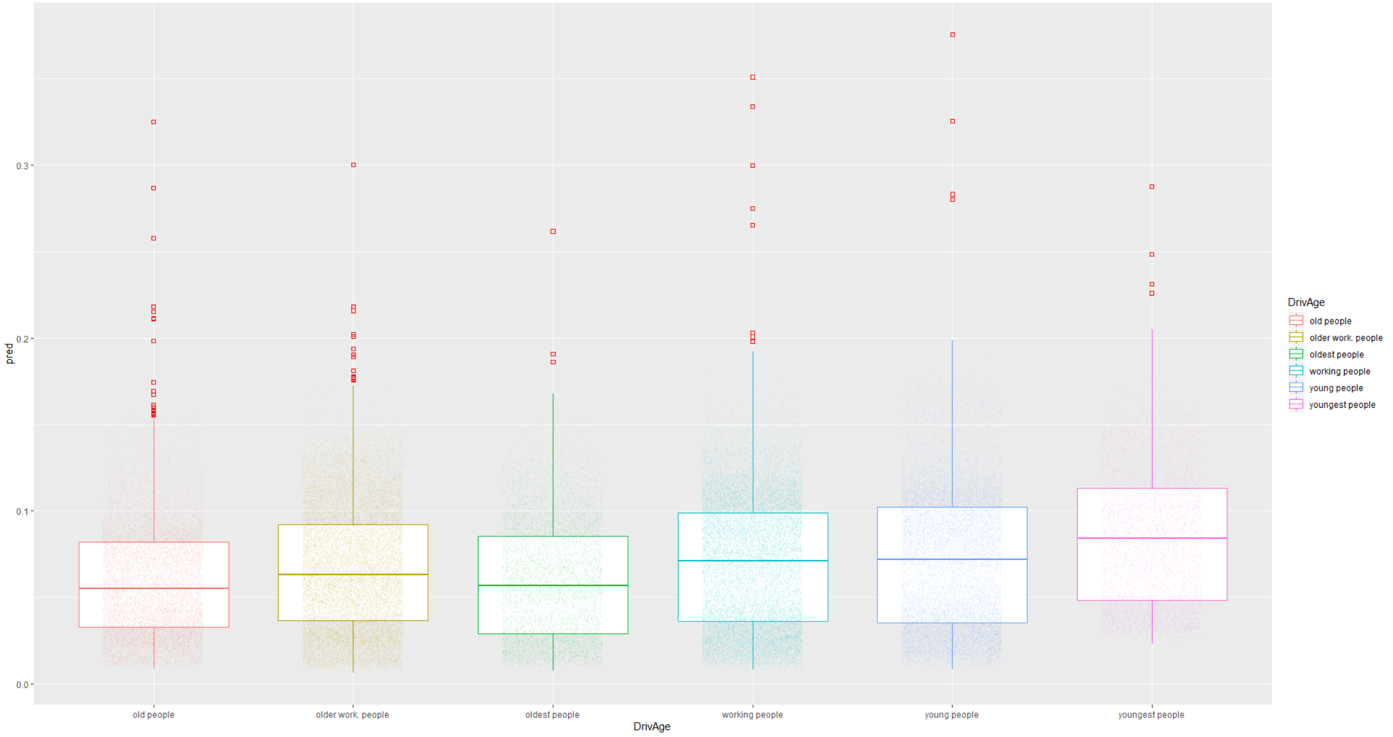


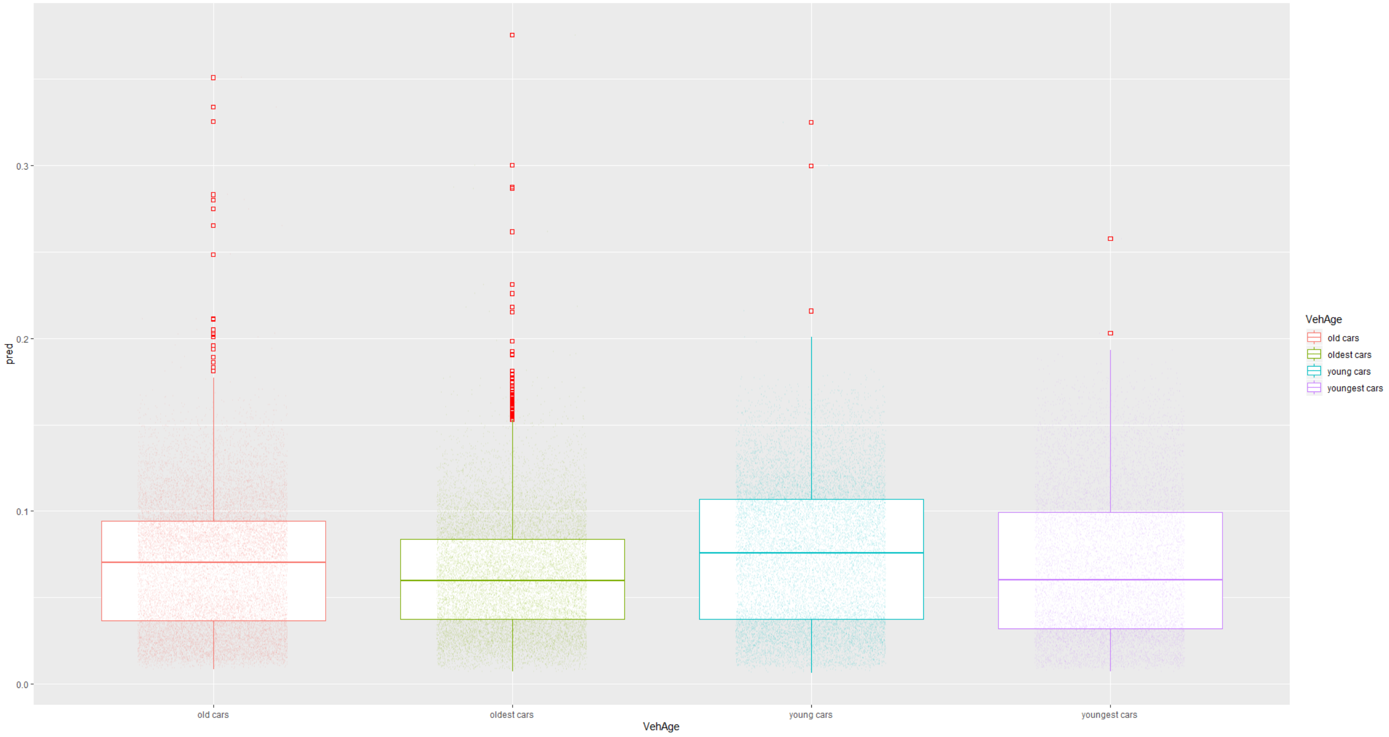
Figure 17. Driver age risk levels



3.5 Vehicle Age

Vehicle age risk follows suit as expected with weaker rates of the oldest and youngest cars. This likely lacked importance due to the importance of value. Additionally, this variable was dichotomized so likely lost a lot of valuable information.

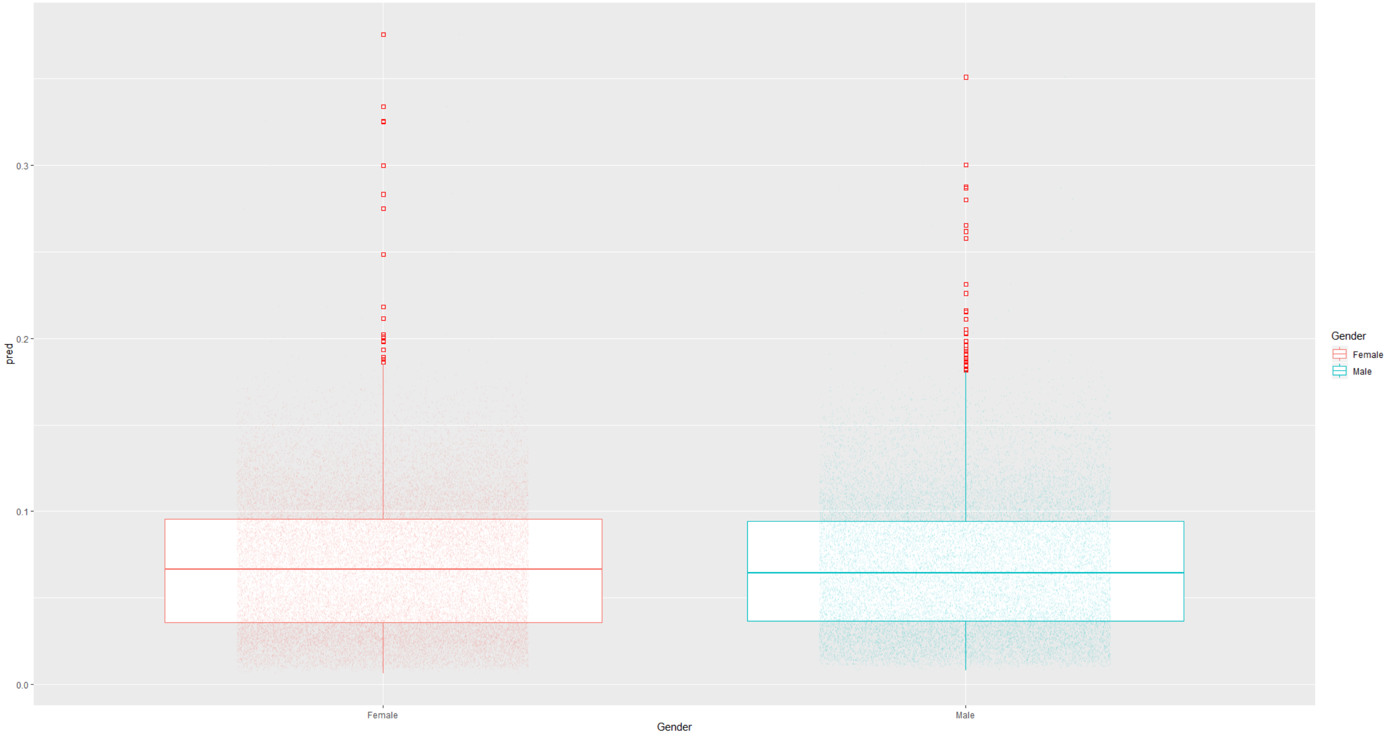
Figure 18. Vehicle age risk levels



3.6 Gender

Finally, gender was the weakest variable as predicted as held no discernible trend between categories.

Figure 19. Gender risk levels



Conclusion

In this project I presented a methodology to determine and visualize the possible risk levels of certain customers. The information created through this method can be used as the basis for further investigation into risk patterns and anomalies which in turn can be used to support pricing strategies. It is important to note that continuous variables performed better than discrete variables and it is likely that the age variables were too widely dichotomized. Furthermore, an interesting development on this methodology would be to segment the data to determine risk factors within categories. A good area to test this on would be gender which in isolation appears to hold no difference between male and female. Finally, this methodology could be improved by testing multiple algorithms to see if they can handle the data better.