Classification: The Probability of Customers Claiming Loans on Auto Insurance

Research Paper

Rahul Narvekar

November 17th, 2021

Table of Contents

Introduction

Loan or Insurance Gap coverage is something that many people keep in the event that their car is involved in an accident. If the car is deemed to be totaled, loan coverage allows customers to pay off the residual value of their car payment after the customer has been compensated for damages incurred in the accident. In such a scenario it would be useful for insurance companies to determine the general likelihood of a customer claiming their loan coverage. Furthermore, customers who finance their vehicles are required to maintain full coverage car insurance. In such a case understanding the likelihood on which loans will have to be claimed can help car insurance companies gauge the various tiers at which to offer their products and associated protection. This research paper intends to look into this subject and explore models that will help insurance companies predict the likelihood of a customer claiming car insurance claim residuals.

Data

Data Description

This data set on Kaggle was obtained from a car insurance company and contains information regarding customer demographics as it pertains to whether customers claimed their auto insurance loans. After accounting for null values, the data contains more than 8,000 records with the following features:

<u>Age:</u> values of 16-25, 26-39, 40-64, 65+

<u>Gender:</u> Male or Female

<u>Race:</u> Minority or Majority

Driving Experience: values of 0-9y, 10-19y, 20-29y, 30y+

Income: poverty, working class, middle class, upper class

Credit Score: normalized metric between 0 and 1

Vehicle Ownership: if the customer has owned a vehicle

Vehicle Year: if the vehicle was made before or after 2015

Married: if the customer is married

Children: if the customer has children

Postal Code: the customers mailing zip code

Annual Mileage: the estimated annual mileage that the customer drives

Vehicle Type: sedan, sport car

Speeding Violations: number of accumulated speeding violations

DUIs: number of accumulated DUI's

Past Accidents: number of accumulated past accidents

Outcome: whether the customer claimed their insurance loan or not, this is what we are classifying

Data Cleaning and Processing

Missing Values: We first removed all missing data for rows that contained null or missing values. This reduced the size of the dataset by about 18%, with 8149 observations.

Data Processing: We converted our classification variable, outcome, to a factor. We did the same for Vehicle Ownership, Children, and Married. This is useful as the models will automatically be accounted for degrees of freedom and many features in this dataset only take on finite values.

For each of the various data modeling methods, different forms of data processing were done to fit the model. Depending on the model some variables were converted to numeric values to match model parameters.

Methods

In this data set we are trying to classify the outcome variable. An outcome of 1 representing that the customer has claimed their insurance loan, while 0 represents that they haven't. Thus a classification based approach would be the most advantageous. We ran Logistic Regression, LDA/QDA, kNN, Classification Trees, Random Forests, Boosting, and SVM. For each model a training and testing set was used. The training set was obtained from a random sample of 70% of the data, while the remaining observations were used for testing. This left 5704 observations in the training set and 2445 observations in the testing set. We also used the caret library to interpret our trained model, and used the confusion matrix function to give us useful metrics on model accuracy.

Logistic Regression

To run the regression we used the glm function with the “binomial” option for family. Below is the output.

```
Call:
glm(formula = OUTCOME ~ ., family = "binomial", data = InsuranceClaims,
    subset = trainingSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9634  -0.5114  -0.1773   0.4286   3.4917

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.294e+00  3.977e-01  -3.253  0.00114 **
ID                       -6.210e-08  1.386e-07  -0.448  0.65401
AGE26-39                 -2.116e-01  1.359e-01  -1.557  0.11955
AGE40-64                 -2.059e-01  1.598e-01  -1.288  0.19764
AGE65+                   -1.366e-01  2.009e-01  -0.680  0.49662
GENDERmale                9.406e-01  8.755e-02  10.743  < 2e-16 ***
RACEminority             -6.094e-02  1.294e-01  -0.471  0.63782
DRIVING_EXPERIENCE10-19y -1.986e+00  1.318e-01 -15.061  < 2e-16 ***
DRIVING_EXPERIENCE20-29y -3.626e+00  2.273e-01 -15.952  < 2e-16 ***
DRIVING_EXPERIENCE30y+   -4.576e+00  4.329e-01 -10.572  < 2e-16 ***
EDUCATIONnone            -2.148e-02  1.138e-01  -0.189  0.85036
EDUCATIONuniversity      -9.167e-03  1.004e-01  -0.091  0.92726
INCOMEpoverty             1.114e-01  1.606e-01   0.693  0.48800
INCOMEupper class        -2.838e-02  1.345e-01  -0.211  0.83294
INCOMEworking class       2.037e-01  1.305e-01   1.561  0.11857
CREDIT_SCORE              5.008e-01  4.453e-01   1.125  0.26069
VEHICLE_OWNERSHIP1       -1.853e+00  9.245e-02 -20.046  < 2e-16 ***
VEHICLE_YEARbefore 2015   1.739e+00  1.118e-01  15.556  < 2e-16 ***
MARRIED1                 -3.788e-01  9.559e-02  -3.963 7.41e-05 ***
CHILDREN1                -6.874e-02  9.615e-02  -0.715  0.47466
POSTAL_CODE               2.210e-05  2.252e-06   9.814  < 2e-16 ***
ANNUAL_MILEAGE            7.675e-05  1.820e-05   4.217 2.47e-05 ***
VEHICLE_TYPEsports car    2.325e-01  1.893e-01   1.228  0.21946
SPEEDING_VIOLATIONS       7.026e-02  3.387e-02   2.074  0.03806 *
DUIS                      1.602e-01  9.847e-02   1.627  0.10378
PAST_ACCIDENTS           -1.395e-01  4.880e-02  -2.859  0.00425 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7078.2  on 5703  degrees of freedom
Residual deviance: 3927.0  on 5678  degrees of freedom
AIC: 3979

Number of Fisher Scoring iterations: 7

[1] "Testing Error:  0.170961145194274"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1499  230
         1  188  528

               Accuracy : 0.829
                 95% CI : (0.8135, 0.8438)
    No Information Rate : 0.69
    P-Value [Acc > NIR] : < 2e-16
```

The output shows that the dummy variables for being male, driving experience, having owned a vehicle, the vehicle being older than 2015, being married and the numeric variable annual mileage were statistically significant at all standard significance levels. The numeric variable for number of past accidents and the dummy variable for being a minority were also significant but to a lesser extent. We then used the model to predict outcomes in our testing set. Any p-value greater than 0.5 was classified as an outcome of 1. We found that the testing error was 17.1%. However the number of false negatives were 230, this gives logistic regression a high false negative rate at 30.3%. The model also has a false positive rate of 11.1%. Logistic regression underestimates the true number of customers who claim their loan amount but offers good interpretability.

LDA/QDA

```
Call:
qda(OUTCOME ~ ., data = training)

Prior probabilities of groups:
        0         1 
0.6865358 0.3134642 

Group means:
   AGE26-39  AGE40-64     AGE65+ GENDERmale RACEminority DRIVING_EXPERIENCE10-19y
0 0.2936670 0.3615935 0.26072523  0.4675689   0.09805924                0.3572523
1 0.3310962 0.1448546 0.06487696  0.5822148   0.10346756                0.2494407
  DRIVING_EXPERIENCE20-29y DRIVING_EXPERIENCE30y+ EDUCATIONnone EDUCATIONuniversity INCOMEpoverty
0               0.29749745            0.152196118     0.1506639           0.4420327    0.09627171
1               0.03579418            0.008389262     0.2790828           0.2841163    0.37192394
  INCOMEupper class INCOMEworking class CREDIT_SCORE VEHICLE_OWNERSHIP1 VEHICLE_YEARbefore 2015
0         0.5418795           0.1322778    0.5459436          0.8245659               0.6052094
1         0.1862416           0.2460850    0.4483424          0.4496644               0.8903803
   MARRIED1 CHILDREN1 POSTAL_CODE ANNUAL_MILEAGE VEHICLE_TYPEsports car SPEEDING_VIOLATIONS
0 0.5845250 0.7581716    18661.14       11352.15             0.04647600           1.8981103
1 0.3154362 0.5341163    22095.26       12416.67             0.05089485           0.5268456
        DUIS PAST_ACCIDENTS
0 0.31435138      1.4116445
1 0.08277405      0.2964206
        Length Class  Mode     
prior      2   -none- numeric  
counts     2   -none- numeric  
means     48   -none- numeric  
scaling 1152   -none- numeric  
ldet       2   -none- numeric  
lev        2   -none- character
N          1   -none- numeric  
call       3   -none- call     
terms      3   terms  call     
xlevels   11   -none- list     
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1223   99
         1  474  649
                                          
               Accuracy : 0.7656          
                 95% CI : (0.7483, 0.7823)
    No Information Rate : 0.6941          
    P-Value [Acc > NIR] : 2.09e-15        
                                          
                  Kappa : 0.516           
```

```
Call:
lda(OUTCOME ~ ., data = training)

Prior probabilities of groups:
        0         1 
0.6931978 0.3068022 

Group means:
   AGE26-39  AGE40-64     AGE65+ GENDERmale RACEminority DRIVING_EXPERIENCE10-19y DRIVING_EXPERIENCE20-29y DRIVING_EXPERIENCE30y+ EDUCATIONnone
0 0.2875569 0.3641882 0.26757714  0.4699039   0.09711684                0.3583713               0.29312089            0.158067779     0.1469398
1 0.3377143 0.1417143 0.06685714  0.5697143   0.11200000                0.2565714               0.04171429            0.007428571     0.2822857
  EDUCATIONuniversity INCOMEpoverty INCOMEupper class INCOMEworking class CREDIT_SCORE VEHICLE_OWNERSHIP1 VEHICLE_YEARbefore 2015  MARRIED1 CHILDREN1
0           0.4390491     0.0958523         0.5533637           0.1277188    0.5448467          0.8156297               0.6052099 0.5950936 0.7610015
1           0.2845714     0.3742857         0.1925714           0.2508571    0.4502893          0.4251429               0.8937143 0.3148571 0.5371429
  POSTAL_CODE ANNUAL_MILEAGE VEHICLE_TYPEsports car SPEEDING_VIOLATIONS       DUIS PAST_ACCIDENTS
0    18642.86       11339.40             0.04906424           1.9476480 0.31208902      1.4516945
1    22716.85       12445.14             0.05200000           0.5382857 0.08971429      0.3148571

Coefficients of linear discriminants:
                                   LD1
AGE26-39                 -3.068028e-01
AGE40-64                 -5.039808e-01
AGE65+                   -4.240745e-01
GENDERmale                4.864073e-01
RACEminority             -2.001782e-02
DRIVING_EXPERIENCE10-19y -1.202829e+00
DRIVING_EXPERIENCE20-29y -1.574439e+00
DRIVING_EXPERIENCE30y+   -1.548317e+00
EDUCATIONnone             3.278492e-03
EDUCATIONuniversity       7.882396e-03
INCOMEpoverty             1.722170e-01
INCOMEupper class         4.456756e-02
INCOMEworking class       1.656725e-01
CREDIT_SCORE              1.519453e-01
VEHICLE_OWNERSHIP1       -1.096580e+00
VEHICLE_YEARbefore 2015   7.496329e-01
MARRIED1                 -1.879890e-01
CHILDREN1                -6.134791e-02
POSTAL_CODE               1.073175e-05
ANNUAL_MILEAGE            2.758149e-05
VEHICLE_TYPEsports car    5.335695e-02
SPEEDING_VIOLATIONS      -1.127820e-02
DUIS                      1.968533e-02
PAST_ACCIDENTS           -4.538815e-02
        Length Class  Mode     
prior    2     -none- numeric  
counts   2     -none- numeric  
means   48     -none- numeric  
scaling 24     -none- numeric  
lev      2     -none- character
svd      1     -none- numeric  
N        1     -none- numeric  
call     3     -none- call     
terms    3     terms  call     
xlevels 11     -none- list     
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1490  226
         1  169  560
                                          
               Accuracy : 0.8384          
                 95% CI : (0.8232, 0.8528)
```

We used the MASS library's functions for QDA and LDA. The output shows that the prior probability for group 0 is 0.6932 and the probability for group 1 is 0.3068, meaning that in the whole dataset, 69.32% of people did not file a claim, and 30.68% of people did claim their insurance loan. We used all predictors, as removing any would result in a higher misclassification error rate. Looking at the LDA output, we see the dummy variables for age, gender, children, education, vehicle type and accidents seem to be highly influential. This is because they have the largest coefficients, which will result in having a large impact on the classification chosen. Our misclassification rate for LDA can be found by 1-Accuracy, so 16.16%. The false positive rate is 10.19% and false negative rate is 28.75%. For QDA, we have a misclassification rate of 23.44%, with a false negative rate of 27.98% and a false positive rate of 27.93%. LDA has a much larger false negative rate than LDA, but has an overall higher accuracy.

kNN

For kNN we primarily used the Class library in R. kNN works by looking at the k nearest data points and if more than 50% of them accept the outcome of 1, then the outcome is classified as 1. We ran the kNN algorithm for various values of k up to 25, and found that the value of k = 7 consistently yielded the lowest test error in repeated trials. Summary shown below.

```
                   [1] 7
         "Test error:  0.0486707566462168"

              Reference
     Prediction    0    1
              0 1636   92
              1   27  690

                 Accuracy : 0.9513
                   95% CI : (0.942, 0.9595)
      No Information Rate : 0.6802
      P-Value [Acc > NIR] : < 2.2e-16

                    Kappa : 0.8856
```

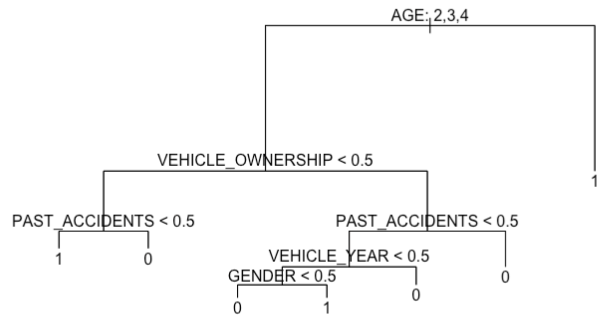
So far this is the lowest test error, having a value of just 4.86%. Again the False Negative rate of 11.5 % is larger than the false positive rate of 2%. However, compared to other models, kNN has produced the most accurate results against the test set, but lacks interpretability.

Classification Trees

```
          Reference
Prediction    0    1
         0 1384  183
         1  335  543

               Accuracy : 0.7881
                 95% CI : (0.7714, 0.8042)
    No Information Rate : 0.7031
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5215
```



For classification trees, we used the trees library. Looking at the output, we can see that the number of misclassifications is lowered with either 6 or 2 predictors. We have a misclassification rate of 20.98% alongside a false positive rate of 3.68% and a false negative rate of 57.51%. Overall, this model has a similar amount of false positives and false negatives. The decision tree and output shows us that age is a highly influential factor. When pruning the tree, we find no real benefit to accuracy but it does simplify the model.

Random Forests

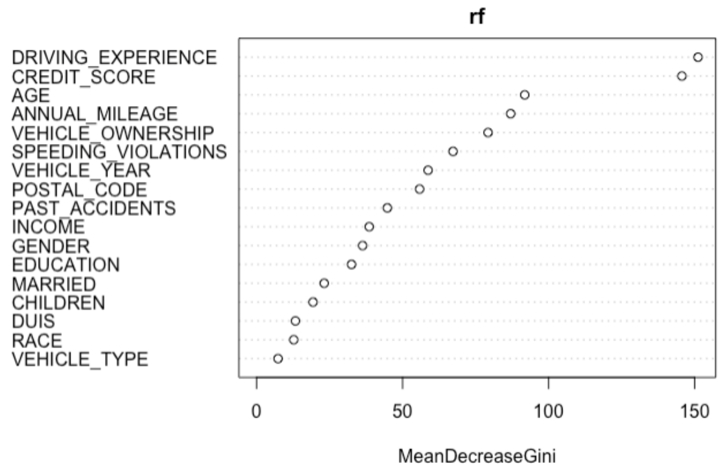
For Random Forest we used the randomForest library. Every variable was used with the exception of ID. This returned an OOB estimate error of 17.78% and a testing error of 1.64%

```
No. of variables tried at each split: 4

        OOB estimate of  error rate: 17.79%
Confusion matrix:
     0   1 class.error
0 1485 183   0.1097122
1  252 525   0.3243243
[1] "test error:  0.016359918200409"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1630    2
         1   38  775

               Accuracy : 0.9836
                 95% CI : (0.9778, 0.9883)
    No Information Rate : 0.6822
    P-Value [Acc > NIR] : < 2.2e-16
```



Random forests rated driver experience, credit score, age, annual mileage, speeding violations , and vehicle ownership as the more important variables. Furthermore, the model has low inaccuracy as its false negative rate is 0.02% and a false positive rate of 2.27%. Random Forest does an overall good job of predicting when people truly claim their car insurance loans and it is easy to interpret which features are most important. This model outperforms knn, with a test error of just 1.64%.

Boosting

	var <chr>	rel.inf <dbl>
CREDIT_SCORE	CREDIT_SCORE	26.7086435
AGE	AGE	18.4766253
VEHICLE_OWNERSHIP	VEHICLE_OWNERSHIP	16.3573403
ANNUAL_MILEAGE	ANNUAL_MILEAGE	9.5683583
PAST_ACCIDENTS	PAST_ACCIDENTS	7.8435884
SPEEDING_VIOLATIONS	SPEEDING_VIOLATIONS	6.9431942
INCOME	INCOME	6.1773352
GENDER	GENDER	4.0656237
MARRIED	MARRIED	1.3933456
RACE	RACE	0.8602272

```
gbm(formula = as.integer(OUTCOME) - 1 ~ ., distribution = "bernoulli",
    data = training, n.trees = 2500, cv.folds = 3)
A gradient boosted model with bernoulli loss function.
2500 iterations were performed.
The best cross-validation iteration was 188.
There were 12 predictors of which 12 had non-zero influence.
Using 188 trees...

Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1517  313
         1  158  457

               Accuracy : 0.8074
                 95% CI : (0.7912, 0.8228)
    No Information Rate : 0.6851
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5279
```

For boosting, we used the gbm library. Initially, we run gbm and initially use 2500 trees; this seems to have the highest accuracy. Increasing it any further will overfit and lower accuracy. Furthermore, we can easily see what predictors matter the most. Credit score, age, and vehicle ownership are the most influential predictors, especially credit score. Afterwards, we perform 3-fold cross validation and we can see that our accuracy goes up to 80.74%, test error rate of 19.26% with a false positive rate of 9.43% and a false negative rate of 40.65%.

SVM

For SVM we used the e1071 library. Due to how large our dataset was, it became unfeasible to use 70% of the dataset for training. Therefore we reduced the training set size to a random sample of 30% of the data. While this took a few minutes to run it allowed us to run the model in a reasonable time frame. Below are the results.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 epsilon cost
       0  128

- best performance: 0.2422616


     Parameters:
        SVM-Type:  C-classification
      SVM-Kernel:  radial
            cost:  128

     Number of Support Vectors:  2153

      ( 1147 1006 )


     Number of Classes:  2

     Levels:
      0 1

      "test error:  0.22159509202454"

          Reference
Prediction    0    1
         0 2509  532
         1  371  663

               Accuracy : 0.7784
                 95% CI : (0.7653, 0.7911)
    No Information Rate : 0.7067
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4435
```

We used the tune function to find an optimal value for cost. After obtaining the best model from tune, we found that the radial kernel with a cost of 128 had the best performance. Again like other models the false negative rate is very high: 44% in fact. While the false positive rate of 12.8% is lower. In general SVM does a poor job of predicting car insurance loan claims, as it underestimates the true number of loans claimed.

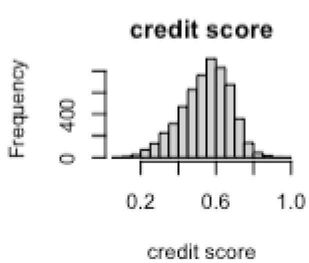
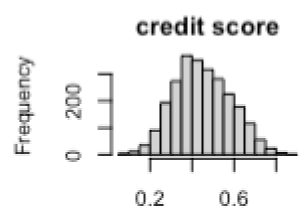
Conclusion

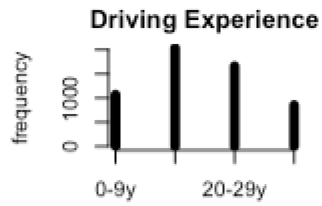
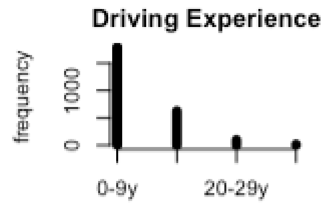
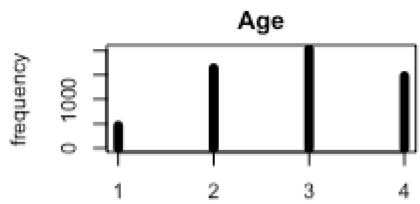
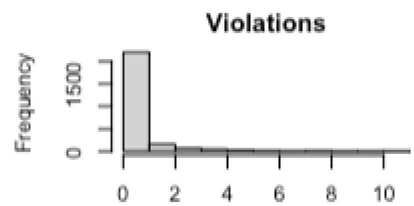
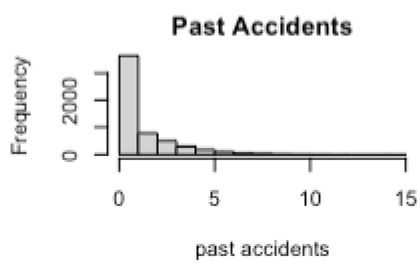
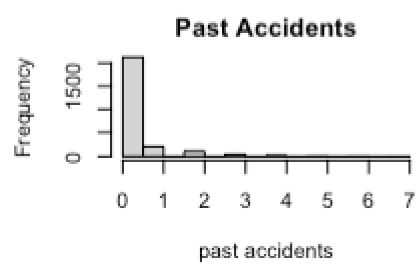
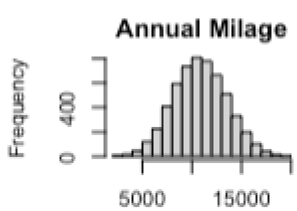
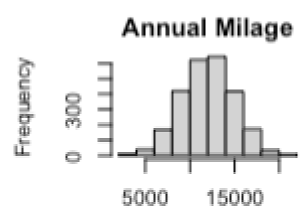
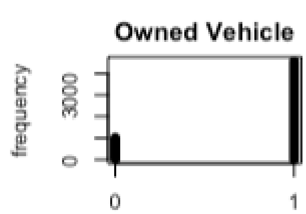
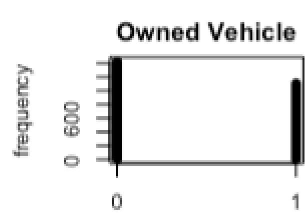
After conducting all our models. We find that random forest has the lowest test error as well as the best performance. Random forest seems to produce the most accurate results when trying to predict auto loan claims. Moreover, random forest's methodology makes it easy to interpret which features were the most impactful. kNN was a close second runner up having marginally higher test error. However, kNN does a poor job of interpretation. In terms of features that contributed towards the loan decisions, from the variety of tests conducted it can be determined that the following predictors are most useful: driving experience, credit score,vehicle ownership, annual mileage, age, speeding violations, and past accidents.

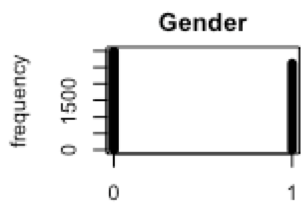
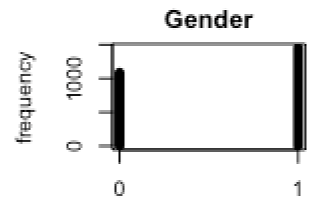
Model	**Test Error**	**False Positive Rate**	**False Negative Rate**
Logistic Regression	17.1%	11.1%	30.3%
LDA	16.16%	10.19%	28.75%
QDA	23.44%	27.93%	27.98%
kNN	**4.87%**	**2%**	**11.5%**
Classification Tree	20.98%	3.68%	57.51%
Random Forest	**1.64%**	**2.27%**	**0.02%**
Boosting	19.26%	9.43%	40.65%
SVM	22.15%	12.8%	44%

We decided to go ahead and plot the most important features as determined by the models to visualize the disparity between loan outcomes.

Visualizing Important Features

Feature	**Outcome 0**	**Outcome 1**
Credit Score		

<table>
<tr><td>Driving Experience</td><td>
</td><td>
</td></tr>
<tr><td>Age</td><td>
</td><td>
</td></tr>
<tr><td>Speeding Violations</td><td>
</td><td>
</td></tr>
<tr><td>Past Accidents</td><td>
</td><td>
</td></tr>
<tr><td>Annual Mileage</td><td>
</td><td>
</td></tr>
<tr><td>Vehicle Ownership</td><td>
</td><td>
</td></tr>
</table>

Gender		

Takeaways

After plotting our most important features we gathered the following insights

- Credit Score: Those who did not file a claim had credit scores that seemed more higher, while those that did file claims tended to have a lower score for credit. As seen from the visualization, those who did not file claims had credit scores that skewed higher, while those that did had credit scores that skewed lower.
- Driving Experience: Those who did not file a claim had more driving experience, compared to those that did file a claim
- Vehicle Ownership: Those who were vehicle owners were less likely to file a claim compared to those who were not vehicle owners. This implies that those who own a vehicle practice safer driving habits, in turn not needing to file for a loan
- Age: among those that did file for claims, younger drivers were more prevalent

Unsurprisingly, the number of speeding violations and past accidents were found to be important factors in our models. These features are often measures of driver safety in the real world. Moreover, features that indicate how experienced a driver is or how long they have been driving also played a big role in the outcome. Surprisingly, we found that those who did own their vehicle were less likely to file claims. Similarly, those with better credit scores were not as prevalent in filed claims. The metrics of vehicle ownership and credit scores can be regarded as metrics of higher financial responsibility and associated safer driving practices. This may be a reason they were less likely to file loans. In some of the models gender was also found to be an important factor. Our visualization shows that being male is also associated with loan claims.