Data Science - Capstone: Cars Insurance Claim. An application of machine learning techniques to know if a client will claim their loan

Paula González Mataix

7/28/2021

Introduction

This present work is part of the capstone project of the ninth course, called Capstone, of the Professional Certificate in Data Science at Harvard University and EdX. Its goal is to demonstrate that the student acquired skills and knowledge during the previous eight courses of the program.

The task consists of analyzing a data set called "Cars Insurance Data" that contains information from an insurer about claims that its customers have had in their cars with a series of descriptive characteristics to develop a prediction algorithm that allows guessing if a customer will claim their loan or not.

To do so, the first thing that will be done is to become familiar with the data, then a classification tree and random forest model will be applied, the results will be discussed, and finally some conclusions will be drawn.

Data overview

The first step is to know the data. The database that will be used in this work is called the Cars Insurance Data data set.

The chosen database includes a sample of 10,000 claims that have an outcome variable that indicates whether or not clients claim the loan. In addition, 18 other variables are collected that will be discussed later.

There is no single, clear or even less simple rule to distinguish whether the client will claim the loan or not from a few features, and that's where machine learning techniques to recognize more complex hidden rules will come in.

The variables present in the database are divided between categorical and ordinal and their description is as follows:

ID -> identifies each claim with a numeric code

AGE -> indicates the age of the insured divided into ranges that are: 16-25, 26-39, 40-64 or 65+

GENDER -> indicates the sex of the insured divided as male or female

RACE -> indicates the race of the insured divided as majority or minority

DRIVING_EXPERIENCE -> indicates the years of experience that the insured has driving divided into the ranges: 0-9y, 10-19y, 20-29y o 30y+

EDUCATION -> indicates the level of completed education that the insured has divided as: high school, none or university

INCOME -> indicates the class to which the insured belongs according to their income divided into: middle class, poverty, upper class or working class

CREDIT_SCORE -> indicates the credit score received by the insured

VEHICLE_OWNERSHIP -> indicates if the insured is the owner of the vehicle or not indicated with 0 or 1

VEHICLE_YEAR -> indicates the vehicle registration year divided as before 2015 or after 2015

MARRIED -> indicates if the insured is married or not with values 0 or 1

CHILDREN -> indicates if the insured has children or not with values 0 or 1
POSTAL_CODE -> indicates the postal code of the insured differentiated between the following values: 10238, 21217, 32765 or 92101

ANNUAL_MILEAGE -> indicates the annual mileage that the insured realize on his vehicle

VEHICLE_TYPE -> indicates the type of vehicle that the insured has differentiated into sedan or sports car

SPEEDING_VIOLATIONS -> indicates speeding limit violations committed by the insured

DUIS -> indicates the amount of fines the insured has for driving with alcohol

PAST_ACCIDENTS -> indicates the number of accidents the insured has had previously

OUTCOME -> indicates 1 if a customer has claimed his/her loan else 0

More information about the data can be obtained in the following link:

https://www.kaggle.com/sagnik1511/car-insurance-data

The first step, after converting the variables into factors to be processed correctly, is to make a first general observation of the content of all the variables.

```
##        ID          AGE          GENDER           RACE      DRIVING_EXPERIENCE
##   100153 :   1   16-25:2016   female:5010   majority:9012   0-9y  :3530
##   100198 :   1   26-39:3063   male  :4990   minority: 988   10-19y:3299
##   1003   :   1   40-64:2931                                 20-29y:2119
##   100513 :   1   65+  :1990                                 30y+  :1052
##   100624 :   1
##   100647 :   1
##   (Other):9994
##        EDUCATION              INCOME                   CREDIT_SCORE
##   high school:4157   middle class :2138                       : 982
##   none       :1915   poverty      :1814   0.053357545462743516:   1
##   university :3928   upper class  :4336   0.060866616311608834:   1
##                      working class:1712   0.06481035331020586 :   1
##                                           0.09538709369408786 :   1
##                                           0.0972109997290097  :   1
##                                           (Other)             :9013
##   VEHICLE_OWNERSHIP      VEHICLE_YEAR  MARRIED    CHILDREN   POSTAL_CODE
##   0.0:3030          after 2015 :3033   0.0:5018   0.0:3112   10238:6940
##   1.0:6970          before 2015:6967   1.0:4982   1.0:6888   21217: 120
##                                                              32765:2456
##                                                              92101: 484
##
##
##
##   ANNUAL_MILEAGE     VEHICLE_TYPE  SPEEDING_VIOLATIONS DUIS     PAST_ACCIDENTS
##   11000.0:1253   sedan     :9523   0      :5028        0:8118   0      :5584
##   12000.0:1218   sports car: 477   1      :1544        1:1470   1      :1783
##  13000.0:1137                     2      :1161         2: 331   2      :1104
##  10000.0:1071                     3      : 830         3:  68   3      : 646
##         : 957                     4      : 530         4:  10   4      : 366
##  14000.0: 894                     5      : 319         5:   2   5      : 232
##  (Other):3470                     (Other): 588         6:   1   (Other): 285
##  OUTCOME
##  0.0:6867
##  1.0:3133
##
##
##
##
##
```

The first thing that emerges from this is that the ID variable has a different value for each record and is only an identifier of the claim, therefore, it does not provide any information and will be eliminated from now on.

The first thing that comes out of this is that the ID variable has a different value for each record and is only an identifier of the claim, therefore it does not provide any information and will be removed from now on.

Other important data regarding the records that can be found in the sample are: that the majority are between 26 and 39 years old, that there are more women than men, almost all the insured are of the majority race, most of them are between 0 and 19 years of experience driving, they have passed high school, they are upper class, owners of the crashed vehicle, whose registration date is after 2015, without marrying, with children, their zip code is 10238, almost everyone has a sedan and few a car sports, most have no speeding violations, no DUIS, no accidents in the past, and no loan his claim. In addition, we can see that the annual mileage and credit score vary a lot, so they will be analyzed differently.

Data Analysis

In the first previous analysis, it was seen that the variables CREDIT_SCORE, ANNUAL_MILEAGE, SPEEDING_VIOLATIONS and PAST_ACCIDENTS have more than six different values, and in fact the variable CREDIT_SCORE has a different value for each record so this value is continuous and will be treated in a way different, but the other variables will be formatted so that they can be viewed in table format and decide how to proceed.

```
## SPEEDING_VIOLATIONS
##    0    1   10   11   12   13   14   15   16   17   18   19    2   22    3    4
## 5028 1544   50   30   20   12    5    8    4    3    1    2 1161    1  830  530
##    5    6    7    8    9
##  319  188  140   75   49
```

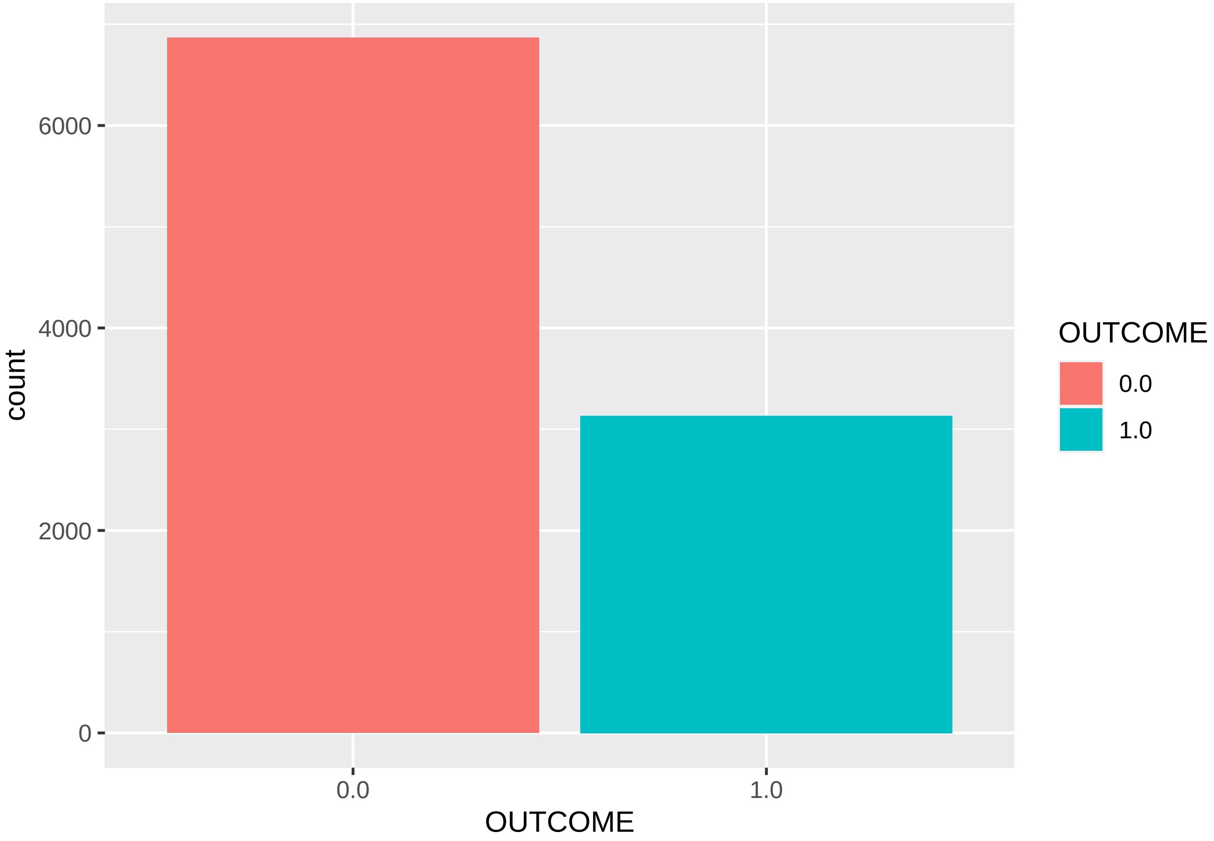
```
## PAST_ACCIDENTS
##    0    1   10   11   12   14   15    2    3    4    5    6    7    8    9
## 5584 1783    9    7    2    1    1 1104  646  366  232  144   61   41   19
```

```
## ANNUAL_MILEAGE
##         10000.0 11000.0 12000.0 13000.0 14000.0 15000.0 16000.0 17000.0 18000.0
##     957    1071    1253    1218    1137     894     632     419     246     103
## 19000.0  2000.0 20000.0 21000.0 22000.0  3000.0  4000.0  5000.0  6000.0  7000.0
##      48       2      13       3       2      10      27      65     165     322
##  8000.0  9000.0
##     557     856
```

As can be seen in the previous tables, these three variables do not have a large number of different values, so they are converted to numeric so that they appear ordered numerically and not alphabetically, but they will be treated like the rest of the variables with discrete values.

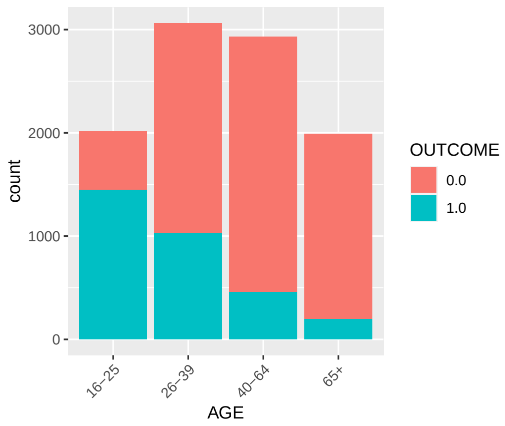
The CREDIT_SCORE variable is also passed to numeric, but it will be treated differently because it is continuous.

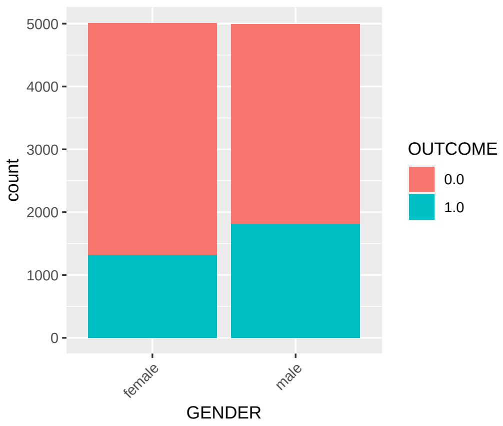
After this little treatment of the data, the variables continue to be analyzed. An important fact is to observe what proportion of claims in the sample have a value of 1 (the client has claimed his loan) and what proportion have a value of 0 (the client has not claimed his loan).

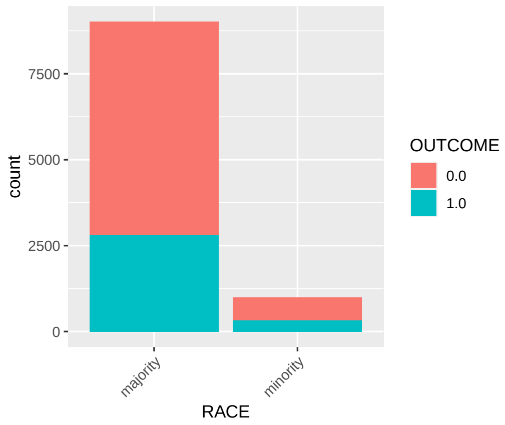


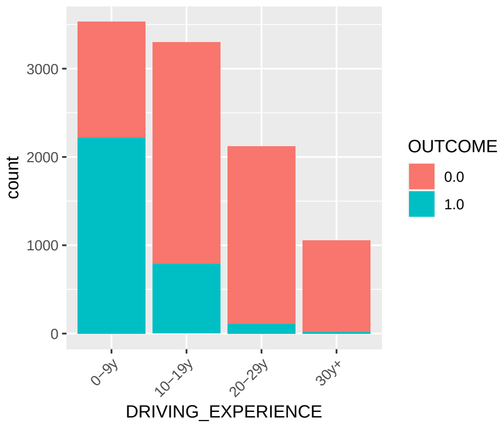
In the previous graph, it can be seen that the number of loans that are not claimed is double that of the number that are, so there may be low prevalence problems.

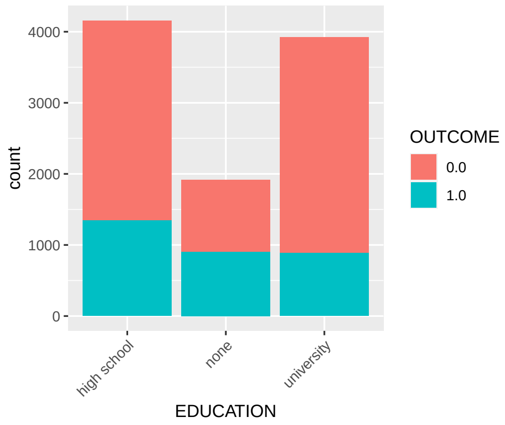
It is also logical to graphically observe the distribution according to the rest of the variables, distinguishing between claimed loan or unclaimed loan.

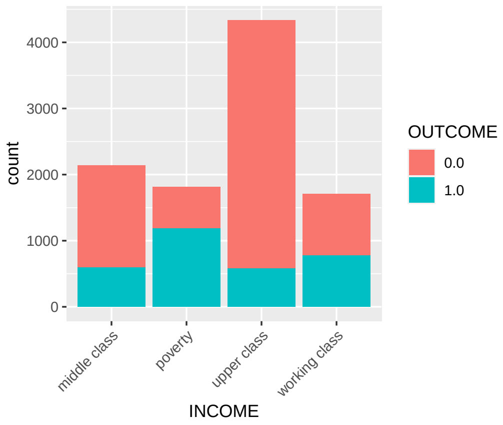

3000
2000
1000
0
count
16–25
26–39
40–64
65+
AGE
OUTCOME
🟥 0.0
🟦 1.0

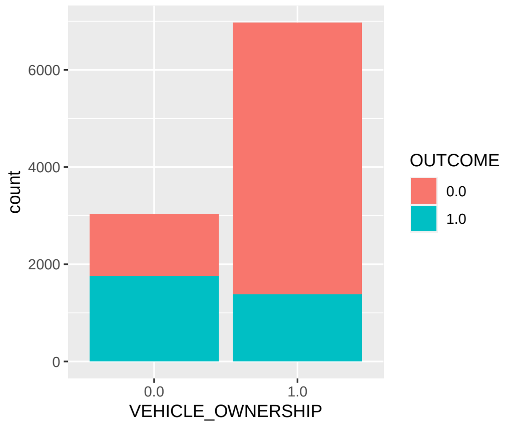

5000
4000
3000
2000
1000
0
count
female
male
GENDER
OUTCOME
🟥 0.0
🟦 1.0

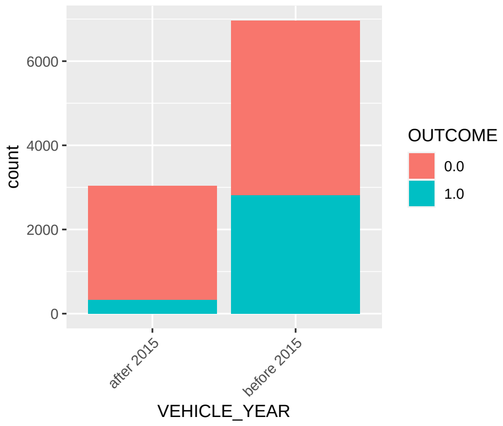

7500
5000
2500
0
count
majority
minority
RACE
OUTCOME
🟥 0.0
🟦 1.0

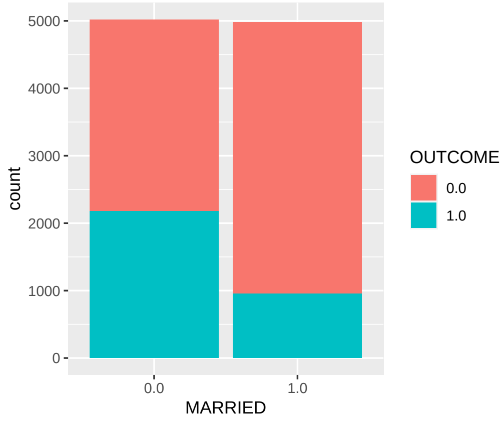

3000
2000
1000
0
count
0–9y
10–19y
20–29y
30y+
DRIVING_EXPERIENCE
OUTCOME
🟥 0.0
🟦 1.0

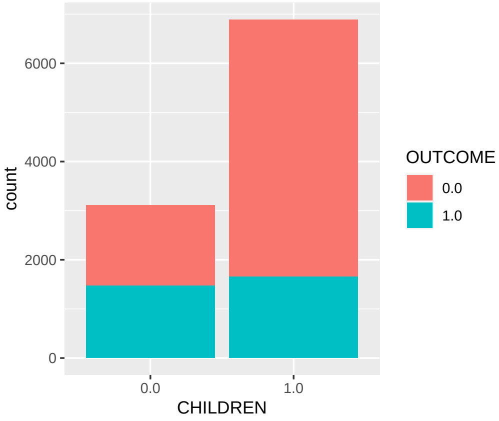

4000
3000
2000
1000
0
count
high school
none
university
EDUCATION
OUTCOME
🟥 0.0
🟦 1.0

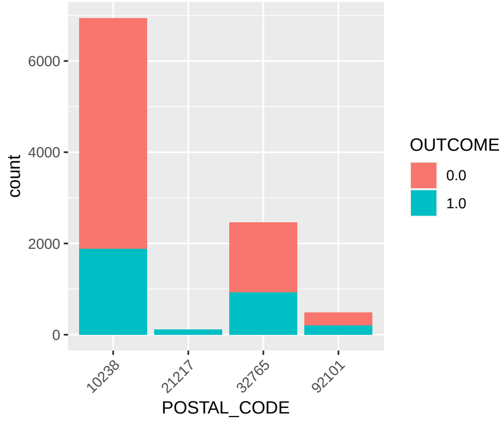

4000
3000
2000
1000
0
count
middle class
poverty
upper class
working class
INCOME
OUTCOME
🟥 0.0
🟦 1.0

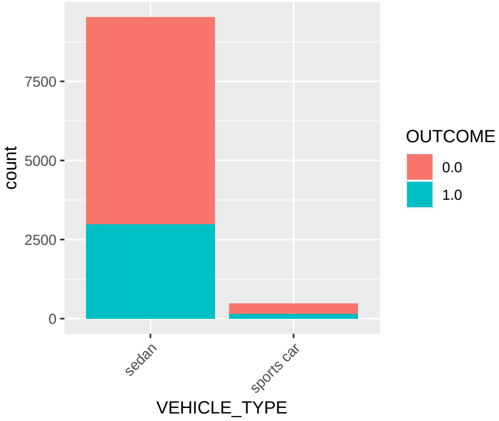

6000
4000
2000
0
count
0.0
1.0
VEHICLE_OWNERSHIP
OUTCOME
🟥 0.0
🟦 1.0

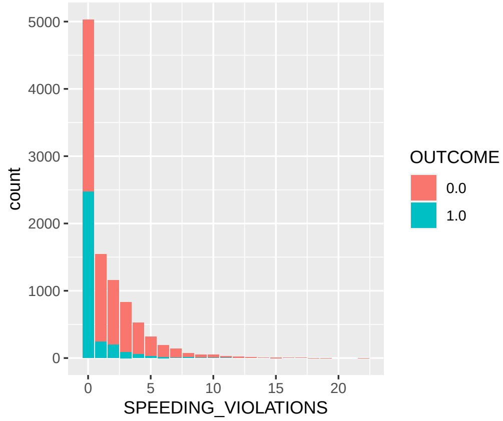

6000
4000
2000
0
count
after 2015
before 2015
VEHICLE_YEAR
OUTCOME
🟥 0.0
🟦 1.0

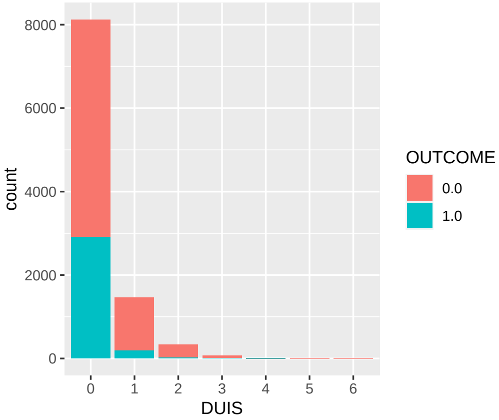

count
5000
4000
3000
2000
1000
0
0.0
1.0
MARRIED
OUTCOME
0.0
1.0

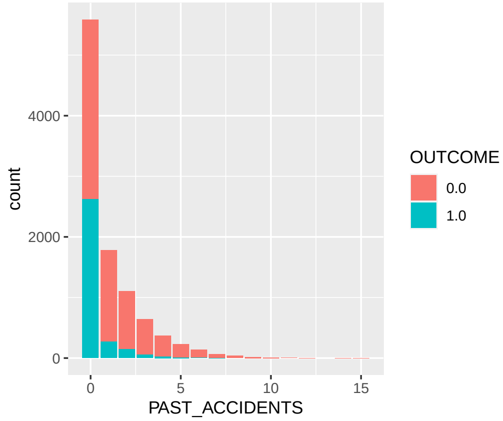

count
6000
4000
2000
0
0.0
1.0
CHILDREN
OUTCOME
0.0
1.0

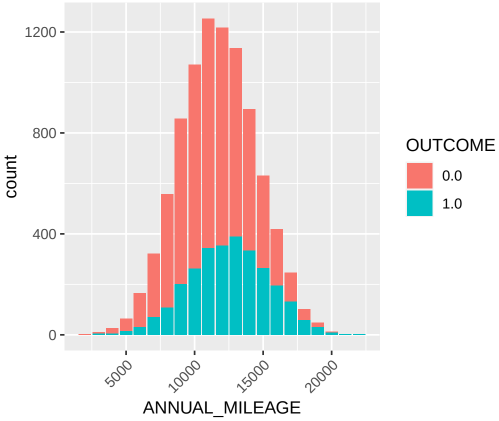

count
6000
4000
2000
0
10238
21217
32765
92101
POSTAL_CODE
OUTCOME
0.0
1.0


count
7500
5000
2500
0
sedan
sports car
VEHICLE_TYPE
OUTCOME
0.0
1.0


count
5000
4000
3000
2000
1000
0
0
5
10
15
20
SPEEDING_VIOLATIONS
OUTCOME
0.0
1.0

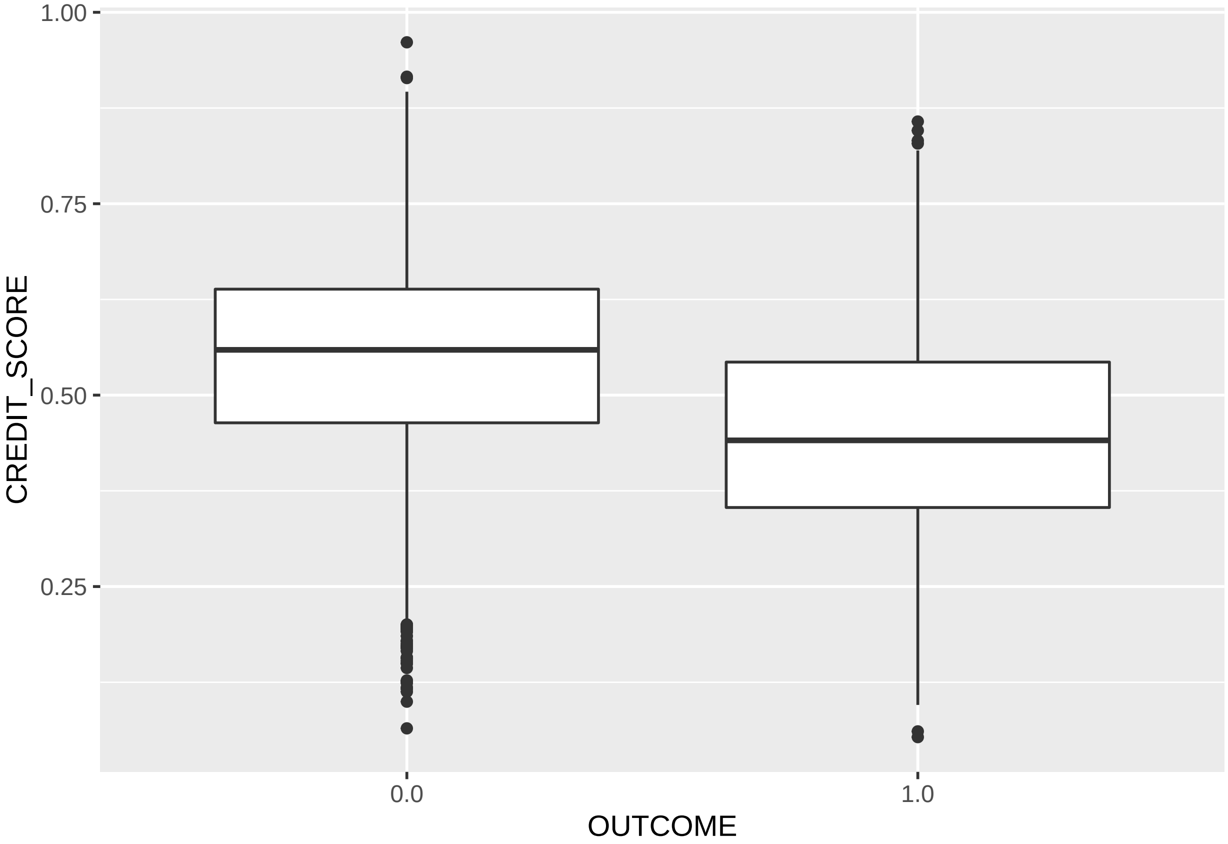

count
8000
6000
4000
2000
0
0
1
2
3
4
5
6
DUIS
OUTCOME
0.0
1.0


count
4000
2000
0
0
5
10
15
PAST_ACCIDENTS
OUTCOME
0.0
1.0


count
1200
800
400
0
5000
10000
15000
20000
ANNUAL_MILEAGE
OUTCOME
0.0
1.0

The reader can appreciate that there are some differences. For example, most clients who claim their loan are between the ages of 16 and 15, while those over 65 tend not to claim it; this also occurs with driving experience, where the majority of clients who claim their loan have between 0 and 9 years of experience, while those with more than 30 years of experience usually do not claim it; men demand it more than women; upper-class clients tend not to claim loans while poverty-class clients do; there is a higher proportion of owners who do not claim than those who do; there are more clients who are not married claim their loan than those who are; the proportion of claimed when they do not have or do have children is similar, but there are many more who have children and do not claim than those who do; those who do not violate speed limits, do not have DUIS and have not had accidents in the past have a similar distribution between those who claim and those who do not and finally, while everyone in the 21217 zip code has claimed the loan, the majority of the 10238 have not they have done.

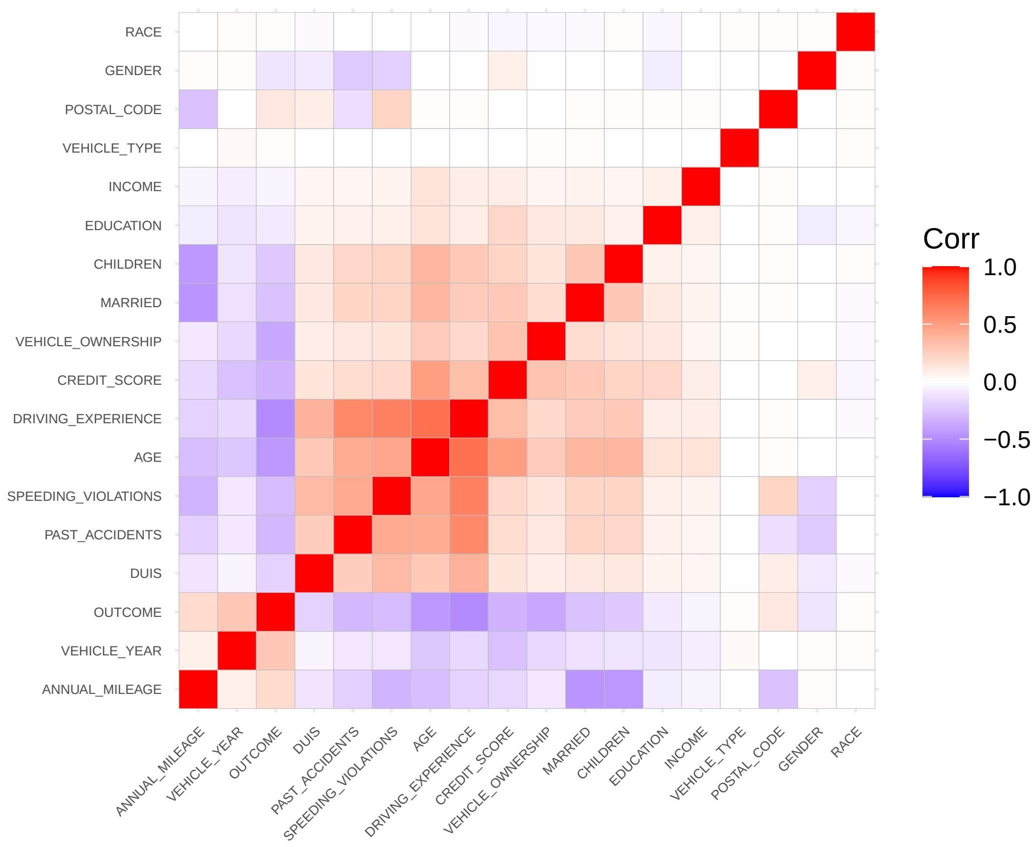
Finally, it is observed how the credit score is distributed according to the outcome value.



As can be seen in the previous boxplot, those who do not claim their loan have a higher credit score, because they have a higher median and higher values and those who do claim it have a lower credit score, in general.

These characteristics are the ones that will help the models to classify the outcome.

Now it will be checked whether there are also correlations between some of these features. Do they usually appear together? Separately? Or is it completely random? This can be analyzed through a correlation matrix, after making some conversions, since some data are categorical, and not numerical.



Indeed, it can be seen that there are elements that relate to others, either in a positive or a negative way.

However, these variables, even when they do not explain the same thing, are related since they indirectly explain the same information or are logical relationships. Example: AGE - DRIVING EXPERIENCE.

Methods

This is a classification problem based on categorical and numerical variables, for which it has been considered that classification trees and random forest would be appropriate techniques. To do this, the database is first broken down into a training set with 80% of the observations and a test set with the remaining 20%. After removing the rows with NA values, the training dataset has 6513 rows and the test dataset 1636.

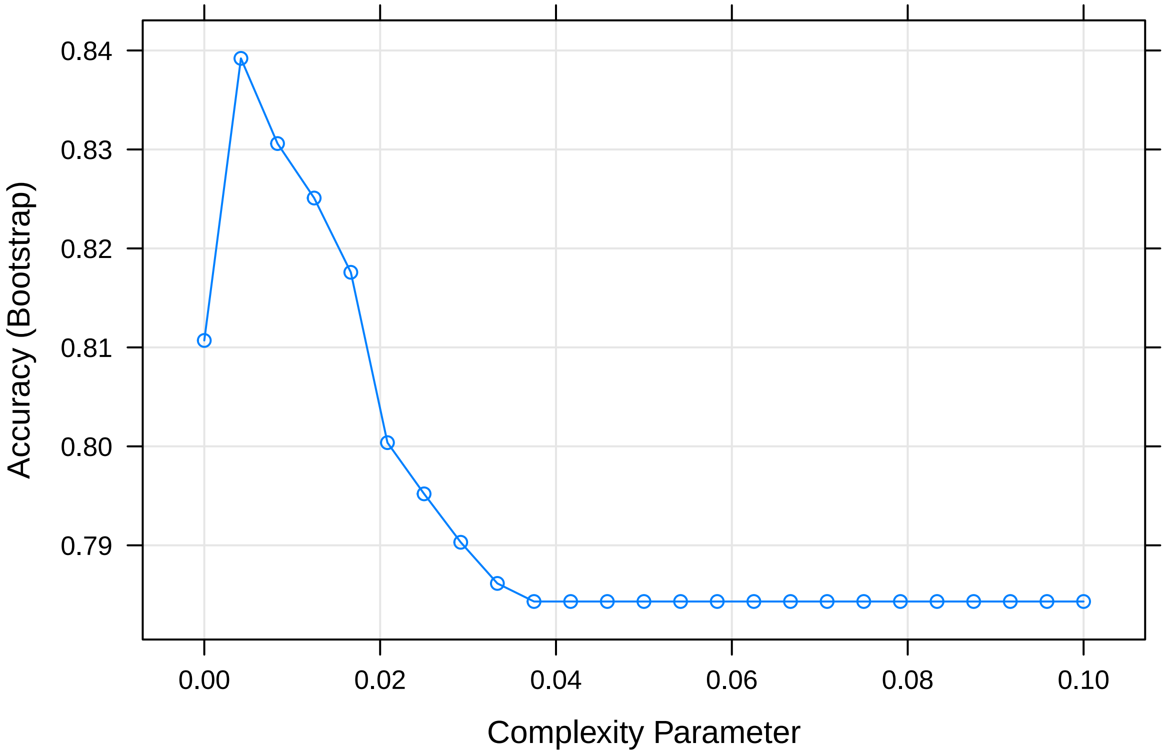
Two different models will be tested: classification tree and random forest.

Classification Tree

The first approach considered is that of the decision tree, which is basically a flow chart or yes or no questions by partitioning the predictors. The predictions at the ends are referred as nodes. Since the outcomes in this case are categorical, we call this a classification tree.

To apply this method, 25 complexity parameters are tested, from 0 to 0.05. The one that gets a better result in terms of precision is 0.0041666678. With lower and higher values, the accuracy decreases (as can be seen in the next plot), so this is the selected parameter.

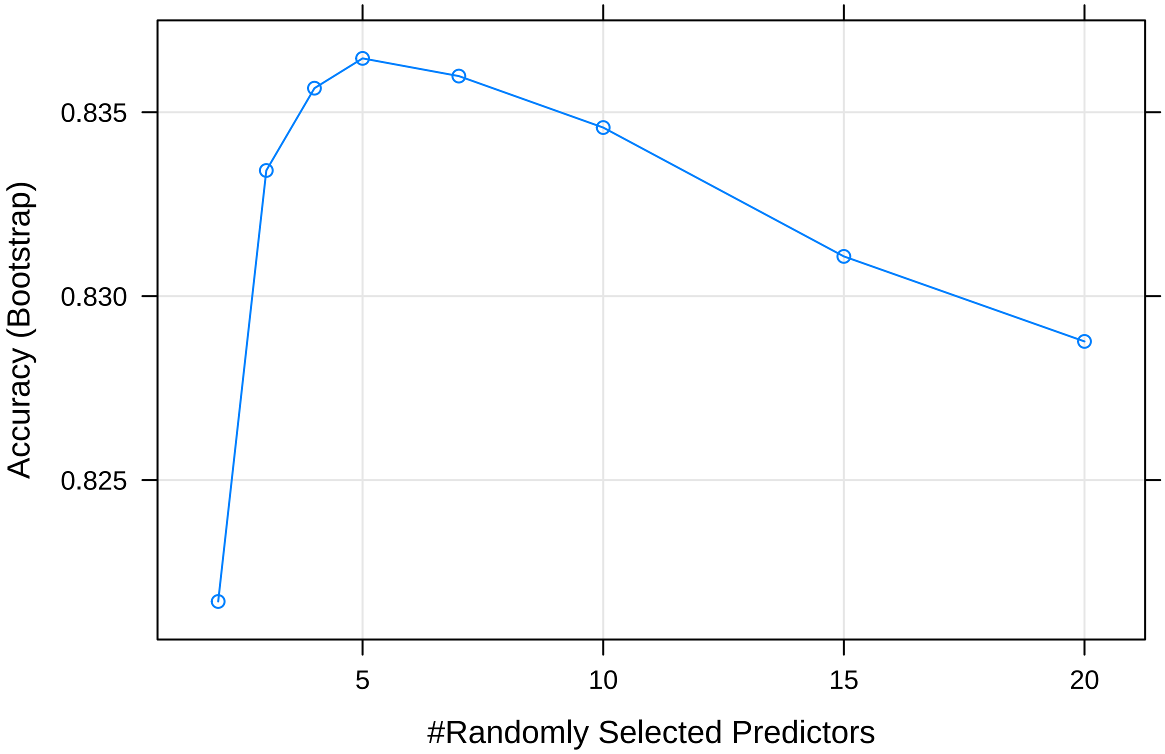
```
## [1] 0.004166667
```



Random Forest

Now, a random forest model is applied. Since there are only two possible outcomes and both have a noticeable presence in the data, there is no point in tuning the minimum node size. This is not the case for the number of selected predictors. Options from 2 to 20 are tested. The best fit in terms of precision occurs with 5 predictors.

```
##   predFixed minNode
## 4         5       2
```

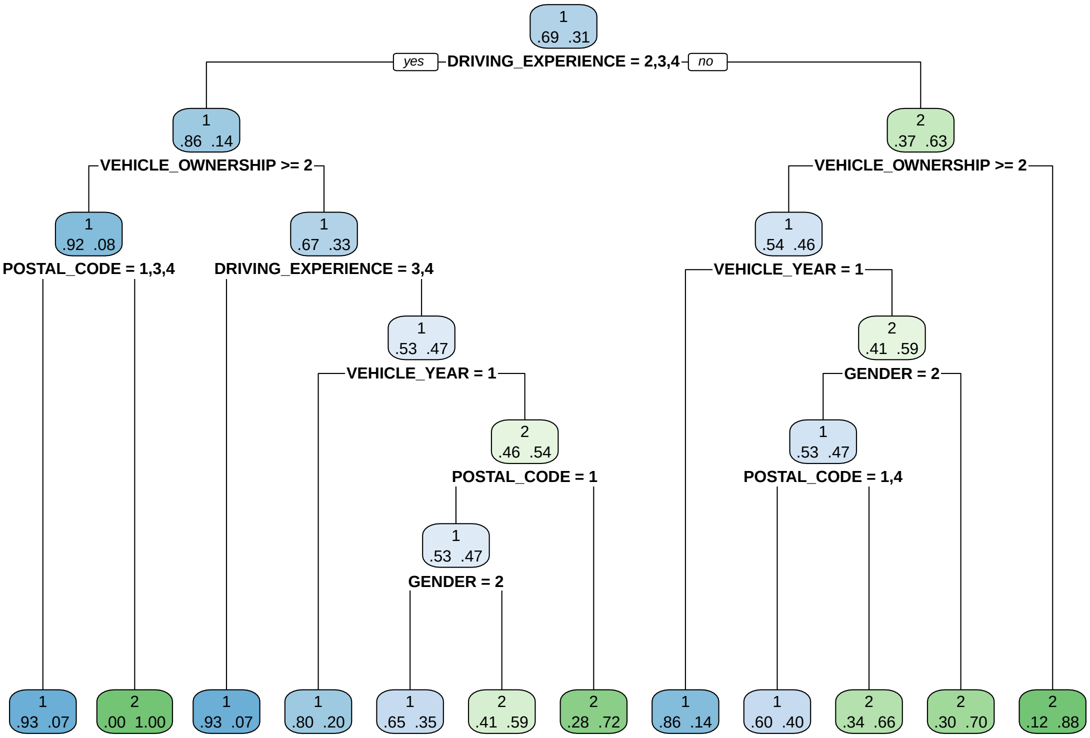


Results

The models have not yet been applied in the test set, so they may be over-adjusted. To check their actual validity they should be tested on a data sample that has not been used for training.

Classification Tree

Before testing it, it is convenient to observe the classification tree that has been generated in the previous section. It is a tree with 12 end nodes. The = or >= of the plot should be interpreted as a question. If this characteristic (variable name + value) is fulfilled, the value is 1; if it is not fulfilled, it is 0. With this information, we answer each question and go to the right if the answer is no, and to the left if the answer is yes. For example, in the first division rule, the tree asks if the customer's driving experience is more than 9 years. If you do not have it (negative answer), continue to the right, if it is over 9 years old (positive answer), continue to the left. One of the main advantages of decision trees is that they are easy to interpret.



This critical point obtained previously will be used for the adjustment of the final model and against which the final evaluation of its effectiveness will be made.

Now it is time to validate the model with the test set. The results of the confusion matrix are shown below.

```
##
## testPredRpart    1    2
##             1 1000  140
##             2  118  378
```

```
## [1] 84.22983
```

It can be seen that, although the global precision achieved is sufficiently high (84.23%), there are 140 classified as claimed. In this case, sensitivity is much more important than specificity.

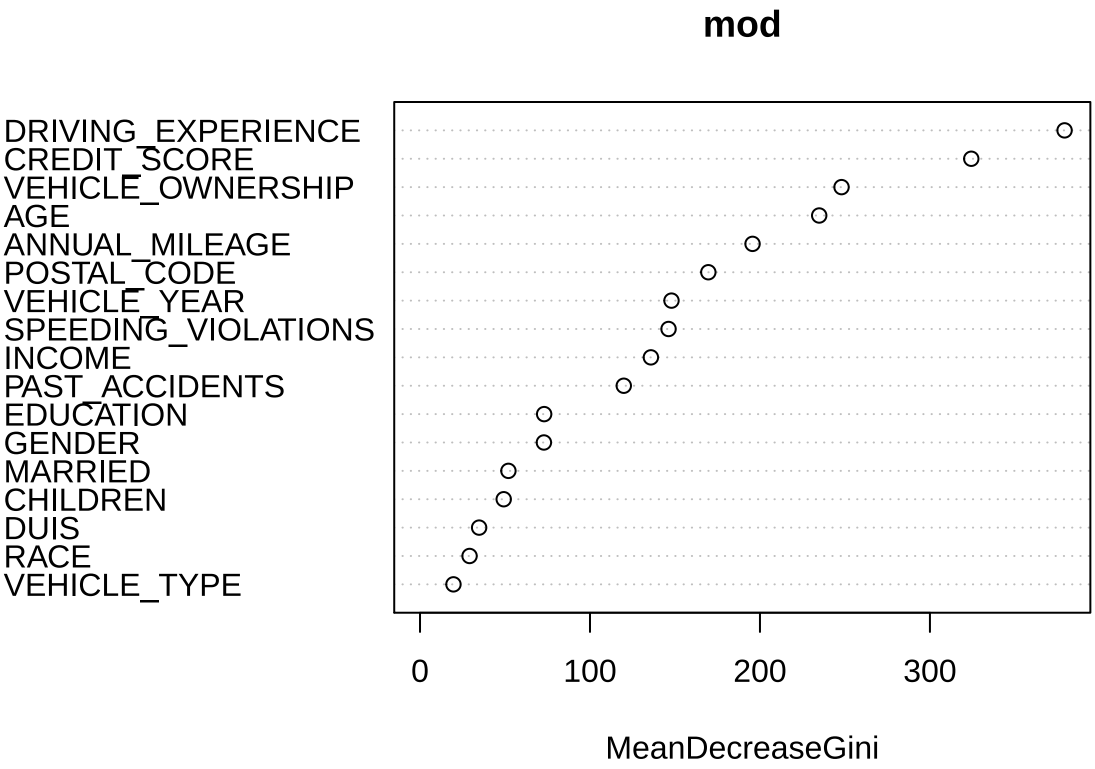
The random forest method is designed to correct some problems of overfitting and instability of decision trees. That is why this has been the next step: will the random forest model be able to correct these errors?

Random Forest

As seen in the previous section, the final random forest model with 5 randomly selected predictors achieves 84% accuracy, although it still needs to be validated in the test set.

Random forest models are not as easily interpreted as decision trees. However, it is possible to know that the most important variables for the classification made by the random forest are the following:

```
##        VEHICLE_TYPE                RACE                DUIS            CHILDREN
##            19.67029            29.15037            34.77743            49.25018
##             MARRIED              GENDER           EDUCATION      PAST_ACCIDENTS
##            52.01296            72.88680            73.03533           119.87882
##              INCOME SPEEDING_VIOLATIONS        VEHICLE_YEAR         POSTAL_CODE
##           135.83462           146.22880           147.95314           169.63767
##      ANNUAL_MILEAGE                 AGE   VEHICLE_OWNERSHIP        CREDIT_SCORE
##           195.64779           234.87026           248.00300           324.33176
##  DRIVING_EXPERIENCE
##           379.22436
```



The importance of the variables in the model shows a clear bias in the variables credit_score and vehicle_year.

Finally, it is validated that the model is the test set. The confusion matrix is now provided:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    2
##          1 1011  150
##          2  107  368
##
##                Accuracy : 0.8429
##                  95% CI : (0.8244, 0.8602)
##     No Information Rate : 0.6834
##     P-Value [Acc > NIR] : < 2.2e-16
```

```
##
##                    Kappa : 0.6287
##
##  Mcnemar’s Test P-Value : 0.008796
##
##             Sensitivity : 0.9043
##             Specificity : 0.7104
##          Pos Pred Value : 0.8708
##          Neg Pred Value : 0.7747
##              Prevalence : 0.6834
##          Detection Rate : 0.6180
##    Detection Prevalence : 0.7097
##       Balanced Accuracy : 0.8074
##
##        ’Positive’ Class : 1
##
```

As can be seen in the confusion matrix above, the Random Forest does not improve the classification of the Classification Tree much since the accuracy is 84.29 and before it was 84.23, but we know that the sensitivity is 0.9043 and the specificity 0.7104.

Conclusion

In this work, two classification algorithms have been developed to distinguish if a client claims their loan or not based on a series of characteristics. Specifically, a classification tree and a random forest have been applied, both with about 84% accuracy. The final model is, therefore, improvable.

As both algorithms have an accuracy of 84%, it is difficult to decide on one or the other algorithm, although the advantage offered by the classification tree is its easy and intuitive interpretation, so that by following the different division rules it would be possible to reach a conclusion for ourselves and when deciding between one or the other, this could help us choose.

Some observations must be taken into account. In this problem in which we find ourselves, sensitivity is more important than specificity since it is important to see the outcome correctly marked as claimed than the possibility that a claim is marked as claimed and has not been.

As the model has room for improvement, future research will have to improve the model, not from the general precision, but looking for the highest possible sensitivity. Also in this case, a better fit of the model should be sought by exploring other techniques and combining them together.