Application of machine learning methods for claim prediction in car insurance data

Report for Workshop I - "Actuarial science"
Master of Data Science, University of Luxembourg

Barbara Gendron-Audebert

December 22, 2022

1 Data and related insurance problem

1.1 Dataset description

This dataset contains exactly 10000 samples, and 19 columns.

The credit score reflect the ability for a policyholder to pay for his debts. The higher the score, the more creditworthy the policyholder is. In car insurance, it is has been observed that this parameter has a significant influence on the insurance rate.
`DUIS` refers to DUI, which stands for *driving under the influence* (whether it be alcohol or drugs).

Variable	Type	Value ranges (if meaningful)
VEHICLE OWNERSHIP	Binary	0 or 1
MARRIED	Binary	0 or 1
CHILDREN	Binary	0 or 1
OUTCOME	Binary	0 or 1
AGE	Category	16-25, 26-39, 40-64, 65+
GENDER	Category	female, male
RACE	Category	majority, minority
DRIVING EXPERIENCE	Category	0-9y, 10-19y, 20-29y, 30y+
EDUCATION	Category	high school, none, university
INCOME	Category	middle class, poverty, upper class, working class
VEHICLE TYPE	Category	sedan, sports car
VEHICLE YEAR	Category	after 2015, before 2015
CREDIT SCORE	Float	From 0.0534 to 0.9608
ID	Integer	–
POSTAL CODE	Integer	–
ANNUAL MILEAGE	Integer	From 2000 to 22000
SPEEDING VIOLATIONS	Integer	From 0 to 22
DUIS	Integer	From 0 to 6
PAST ACCIDENTS	Integer	From 0 to 15

Table 1: A short description of the covariates, along with some insights about categorical variables.

1.2 The insurance context

2 Preliminary analysis of the data

2.1 Exploratory data analysis

In this part I will conduce an exploratory data analysis focusing on the following points:

- distribution of variables for numerical ones
- distribution of numerical features depending on the `OUTCOME` value
- classes distributions depending on the `OUTCOME` value for categorical features

2.1.1 Distribution of variables and classes balance

2.1.2 Distribution of numerical features depending on the `OUTCOME` value

The following figure aims at showing how different is the distribution of a certain numerical feature with respect to the `OUTCOME` value. For this purpose, I used violin plots, which are basically boxplots enhanced by showing the shape of the distribution.

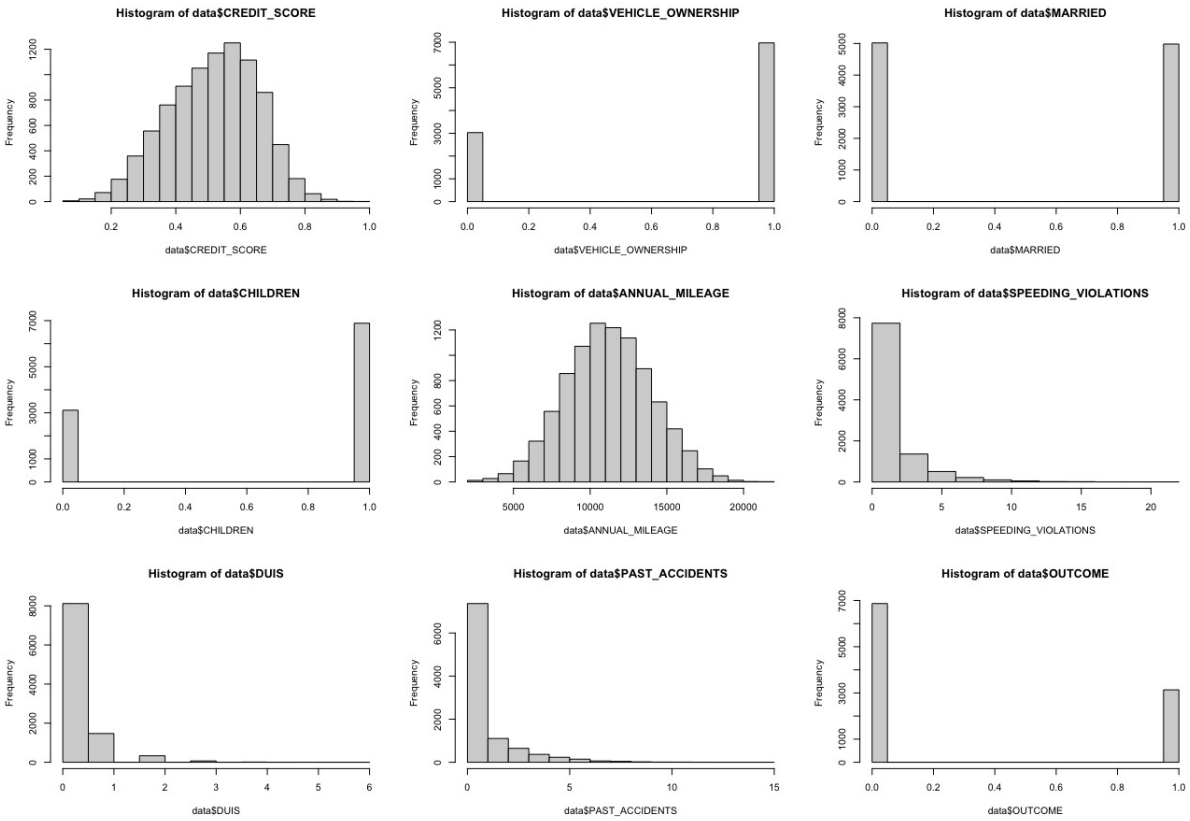

Figure 1: Histograms of the numerical variables.

2.1.3 Classes distributions depending on the `OUTCOME` value for categorical features

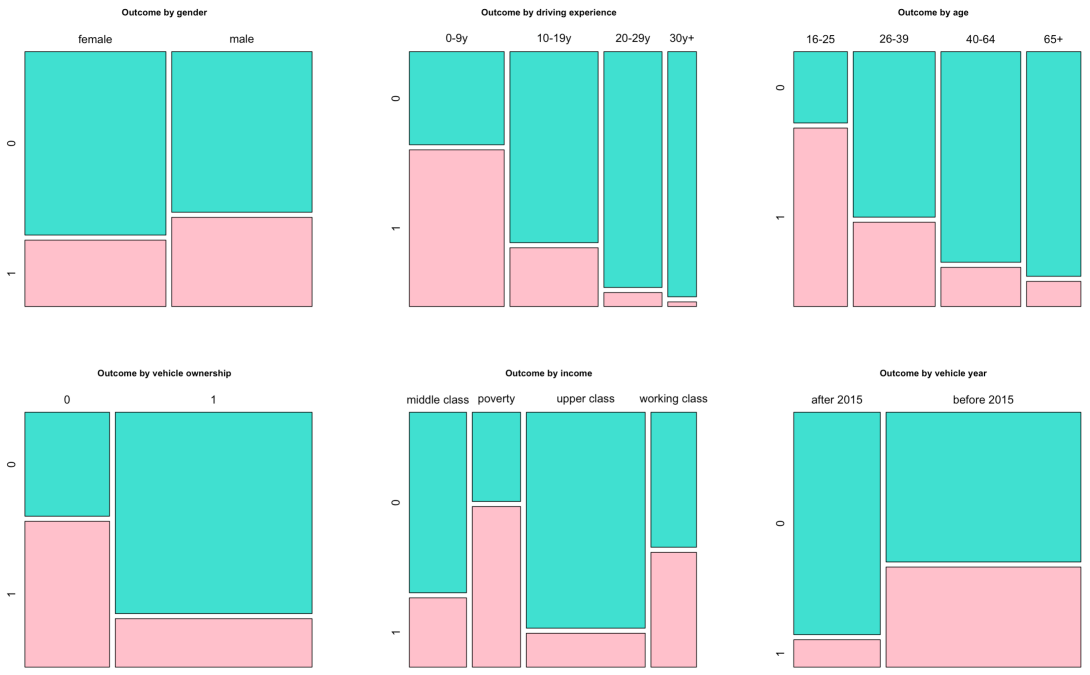

Figure 2: Value counts of categorical features with respect to the `OUTCOME` value.

2.1.4 Dataset summary and completeness of the dataset

Using the `summary()` R function, we observe that the variables `CREDIT SCORE` and `ANNUAL MILEAGE` respectively have 982 and 957 NA's values. This represents approximately 1% of the whole data for each variable, that's why we can consider simply delete them. Thus, the remaining dataset contains 8149 rows.

3 Brief description of the models used

4 Analysis of the results and conclusion

Model	Accuracy	Sensitivity	Specificity	PPV	NPV
Logistic regression	0.8245	0.8629	0.7334	0.8850	0.6923
Decision Tree Classifier	0.8164	0.8534	0.7252	0.8844	0.6675
Random Forest Classifier	0.8245	0.8719	0.7206	0.8725	0.7197
XGBoost	0.8164	0.8844	0.6675	0.8534	0.7252

Table 2: A sum-up table of the classification metrics for each model. PPV stands for Predicted Positive Values and NPV stands for Negative Predicted Values.