LDAT 2310 : Data science for finance and insurance

Project

Patrick Guerin

Novembre 2018

Université Catholique de Louvain

‘

Contents

’

1. Introduction

This project aims to determine the determinants of the claim frequency. To identify those factors we will first understand the dataset with descriptive statistics before modeling the claim frequency with different models(GLM,decision tree,random forest and gradient boosting).

We have at our disposal a training set of 60.000 observations and a test set of 20.000 observations.

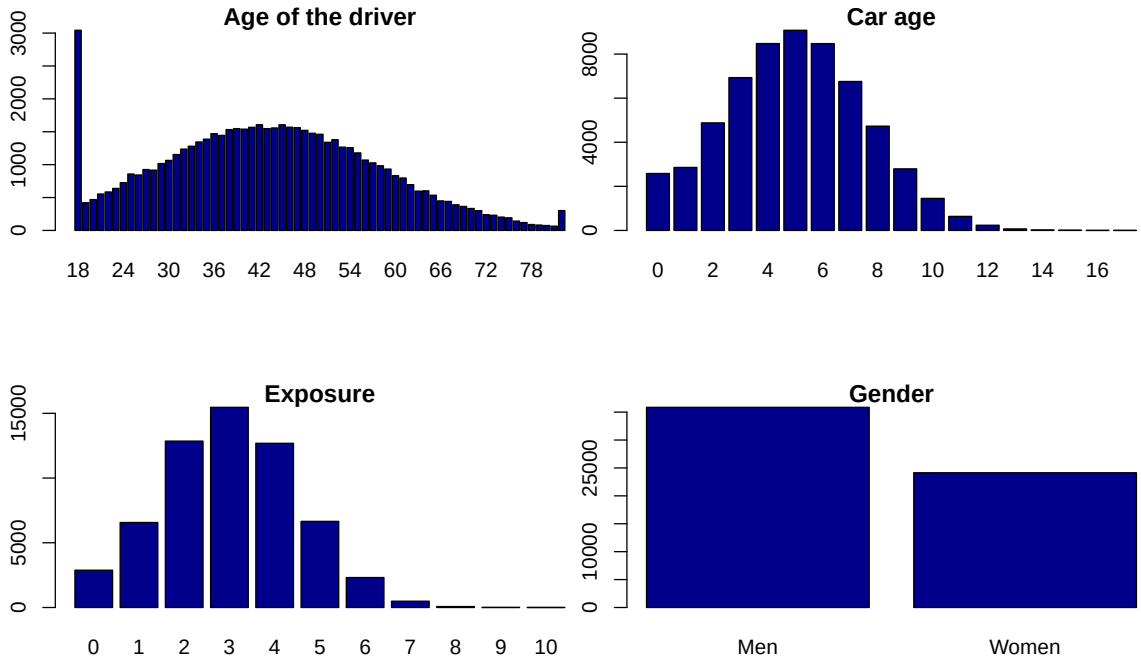
The claim frequency is not initially present in the training set and we construct the it as the ratio of the number of Claims by the exposure, and add it to the dataset. $ClaimFreq = \frac{Nbclaims}{Exposure}$

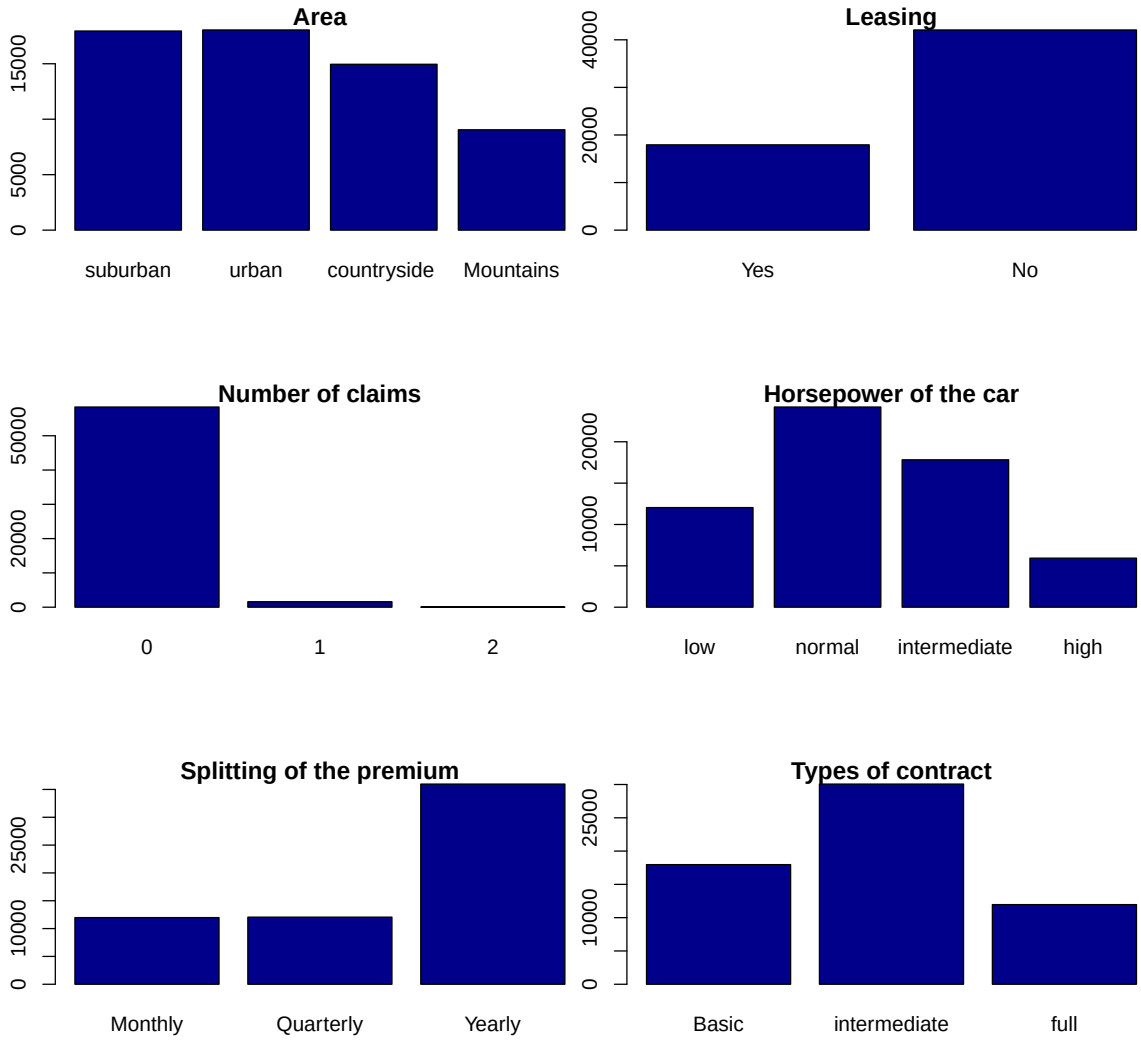
2. Exploratory analysis

To understand the data, start by looking at the distribution of the variables. Barplots are used for the categorical variables.

Table 1: Descriptive statistics of the continuous variables

	DriverAge	CarAge	Exposure
Mean	43.3	4.99	3.02
Median	43.0	5.00	2.99
Standard deviation	14.3	2.55	1.47
Q25	33.0	3.00	1.99
Q75	53.0	7.00	4.02
Min	18.0	0.00	0.08
Max	82.0	17.00	9.54

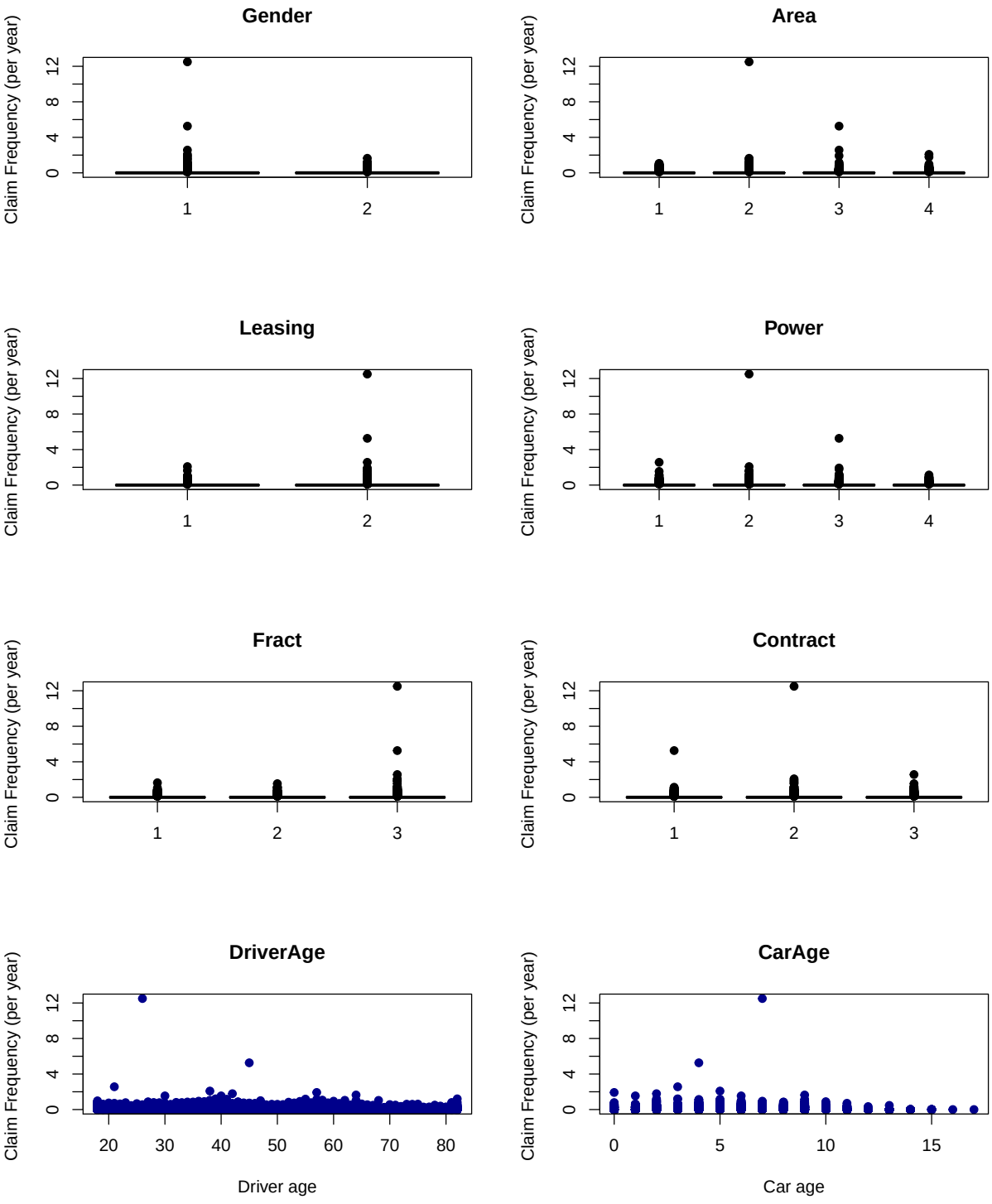


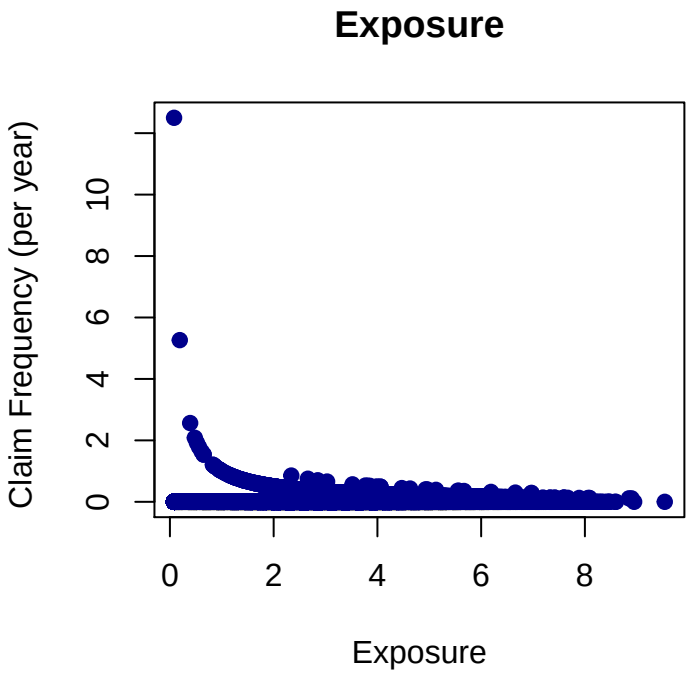


The typical insured person is a man of 43 years living in city owning a 5 year old car with a Horsepower in the normal category. He is unsured by an intermediate type of contract, pays his prime yearly, has no leasing or claims and has been insured for 3 years.

Besides this general picture, We can notice that: - The drivers of 18 years old are over-represented. 97.3% of the individuals never had any claims.

After first exploration, we plot the explanative variables against the claim frequency to get an hint about the variables associated with the claim frequency, although interactions between variables are not taken into account yet.


Gender
Claim Frequency (per year)
0
4
8
12
1
2
Area
Claim Frequency (per year)
0
4
8
12
1
2
3
4
Leasing
Claim Frequency (per year)
0
4
8
12
1
2
Power
Claim Frequency (per year)
0
4
8
12
1
2
3
4
Fract
Claim Frequency (per year)
0
4
8
12
1
2
3
Contract
Claim Frequency (per year)
0
4
8
12
1
2
3
DriverAge
Claim Frequency (per year)
0
4
8
12
20
30
40
50
60
70
80
Driver age
CarAge
Claim Frequency (per year)
0
4
8
12
0
5
10
15
Car age



We can observe that:

-the highest Claim frequencies are achieved by men.

-Customers living in urban areas tend to have a higher claim frequency than others.

-The leasing does not seems to be a major factor, considering the mean of the two groups.

-Customers owning a low horsepower car have a tend ot have a lower claim frequency,

-There is more outliers in the high horsepower category, indicating that a small number of individuals with powerful cars have a large number of claims.

-The splitting of the premium does not seems to differentiate the customers.

Finally, in average, customer who beneficiate of a full contract have higher claim frequencies.

For more details, means and standard deviations by categories are available in annexe.

It is also important to check if the training and tests variables have similar distributions, and if it wasn't the case some adjustements would be necessary to obtain the best possible fit. After comparing the distributions of the continuous variables and analysing the bar plots of the categorical variables, the two datasets seems identical from a statistical point a view. Test set statistics can be found in annexe.

3.Modeling

3.1 Generalized linear model

Our first model is a generalized linear model, and more specifically a poisson regression with the variable *Exposure* as offset. The non significative variables are eliminated using a stepwise backward selection, taking AIC as criterion.

Table 2: Poisson regression parameters

	Estimate	Std. Error	z value	Pr(> abs(z))
(Intercept)	-4.7038246	0.1061161	-44.3271696	0.0000000
Gender2	-0.1023488	0.0508688	-2.0120166	0.0442182
CarAge	-0.0156141	0.0097101	-1.6080174	0.1078314
Area2	0.1124069	0.0614721	1.8285845	0.0674619
Area3	-0.1951566	0.0700847	-2.7845822	0.0053597
Area4	-0.0535724	0.0787307	-0.6804511	0.4962188
Power2	0.1275257	0.0707329	1.8029174	0.0714012
Power3	0.2031573	0.0734678	2.7652556	0.0056878
Power4	0.2010116	0.0960970	2.0917581	0.0364602
Fract2	-0.0818502	0.0758169	-1.0795767	0.2803307
Fract3	-0.1353190	0.0618092	-2.1893036	0.0285748
Contract2	0.0510031	0.0592473	0.8608511	0.3893200
Contract3	0.3234622	0.0683874	4.7298504	0.0000022

Table 3: Deviance indicators

Null deviance	Residual deviance
11426.95	11360.25

Two variables are dismissed: The driver age and the leasing.

The Residual deviance and null deviance are close to each other, however the residual deviance is smaller,indicating a better fit. Moreover this closeness is not surprising considering the number of zeros in the response variable.

At at risk of 5% one can observe that:

- The women have a claim frequency 10.2% lower than men
- The customers residing in Area3(Countryside) makes 19.5% less claims for a given duration, while difference between others areas is not significative.
- There is a positive relationship between the claim frequency and the Horsepower of the car and owning a car with high horsepower raise the claim frequency of 20.1% compared to low horsepower owners.
- Customers with a yearly splitting of the premium have a claims frequency reduced of 13.5% compared to others.
- full contract subscripter make 32.3% more claim than other for a given duration. It is a clear case of *adverse selection*.

Finally, we can see below that there is no overdispersion phenomenon.

the number of insurance claims within a population for a certain type of risk would be zero-inflated by those people who have not taken out insurance against the risk and thus are unable to claim

One property of the poisson distribution is the equality of the mean and observed variance. If this property is not respected, for instance in the case of overdispersion, it can impede the model performance

Table 4: Mean and variance of the claim frequency

Mean	Variance
0.0090024	0.0067439

We can see there is no overdispersion phenomenon.Another phenomenon that could cause problem is the quantity of zeros from the response variable. We attempted to adress it with a zero-inflated model thanks to the package **pscl**, but it did not produced valuable results.

3.2 Poisson regression tree

The second model is a poisson regression tree

At each level of the tree the split point is chosen to minimize difference between the deviance of the parent node and the deviance of its two child nodes

The deviances are computed assuming a poisson distribution of the response variable.

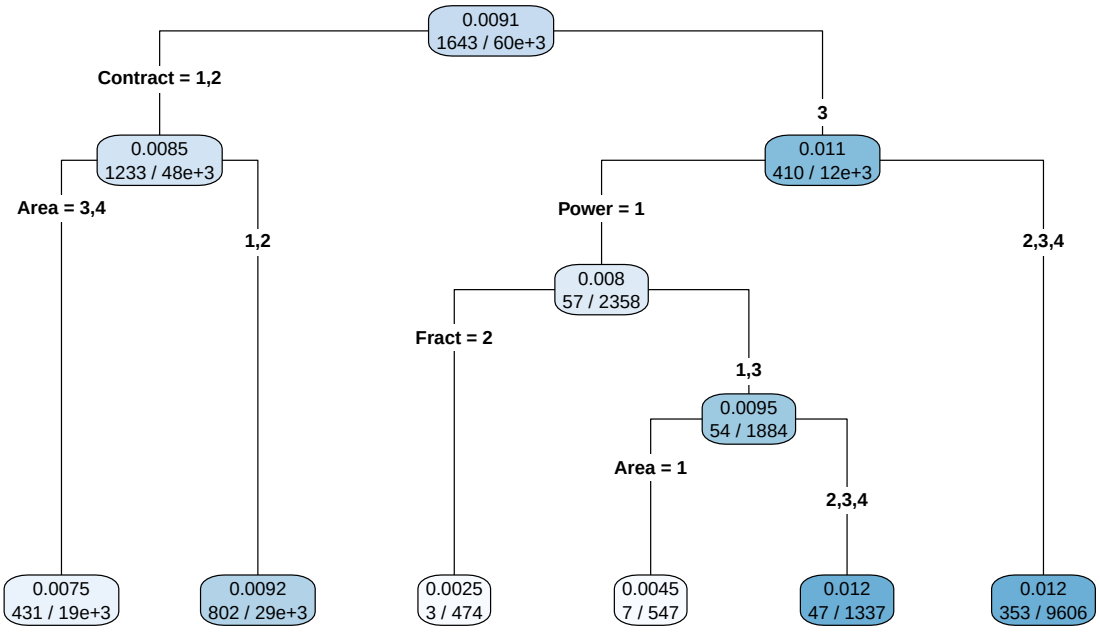
A large tree is first grown, and a 10-fold cross-validation is used to estimate the generalization error of the models for different depths.

Table 5: Results

CP	nsplit	rel error	xerror	xstd
0.0021715	0	1.0000000	1.0001690	0.0179979
0.0011206	1	0.9978285	0.9990126	0.0179972
0.0008948	2	0.9967078	0.9994934	0.0180307
0.0007147	4	0.9949183	0.9988050	0.0180194
0.0006625	5	0.9942037	1.0003832	0.0180725
0.0006622	7	0.9928787	1.0006558	0.0180975
0.0006087	8	0.9922164	1.0007138	0.0181014
0.0005820	9	0.9916077	1.0031788	0.0181916
0.0005759	11	0.9904437	1.0051419	0.0182480
0.0005324	12	0.9898678	1.0066779	0.0183274
0.0005300	13	0.9893354	1.0107295	0.0184715

The tree starts overfitting after the fourth split,therefore we set the depth of the tree to 4.

The graph of this tree is plotted below.



We can see that the most important factors is the type of contract.This observation can be linked with the analysis of the Poisson regression coefficients which revealed that those customer makes 32% more claims for a given duration compared to customers with other contracts.

-If the customer has a full contract and own a categorized as having a horsepower of category 1, his estimated claim frequency is 0.008, to compare with 0.012 if he has more horsepower.

-Countryside and mountains areas (3,4) are associated with lower claim frequencies if the customers have a basic or intermediate contract.

3.4 Poisson Random Forest

In this section we average the results of 100 Poisson Regression trees to reduce the variance of the estimator and improve our prediction. For each tree, a set of variables are sampled and used to grow the tree, and each tree is fitted on a different dataset generated by sampling with replacement from the real data.

This two randomizations foster statistical independence between trees, and eventually reduce the variance of the final estimator.The trees are ideal candidate for ensembling given their high variance and low bias.

To get a better understanding of the factor influencing the claim frequency, we segment the dataset into several customers profils:

7 axis are explored:

1 - The gender

2- The category

3- The Car age

4- The Area

6- The Driver age

7- The type of contract

We use the following methodology: we fix some factors,for instance "gender"=1, "Power"=1 and "CarAge" < 3 years, then predict the claim frequency for each customer corresponding to this criterion. Finally we average the predictions, so that we can plot the curve of the average claim frequency for the male customer owning a less than 3 years old low horsepower car. This approach has the advantage of focusing on just a few factors, *ceteris paribus*.

3.4.1 CarAge,Horsepower and Gender variables

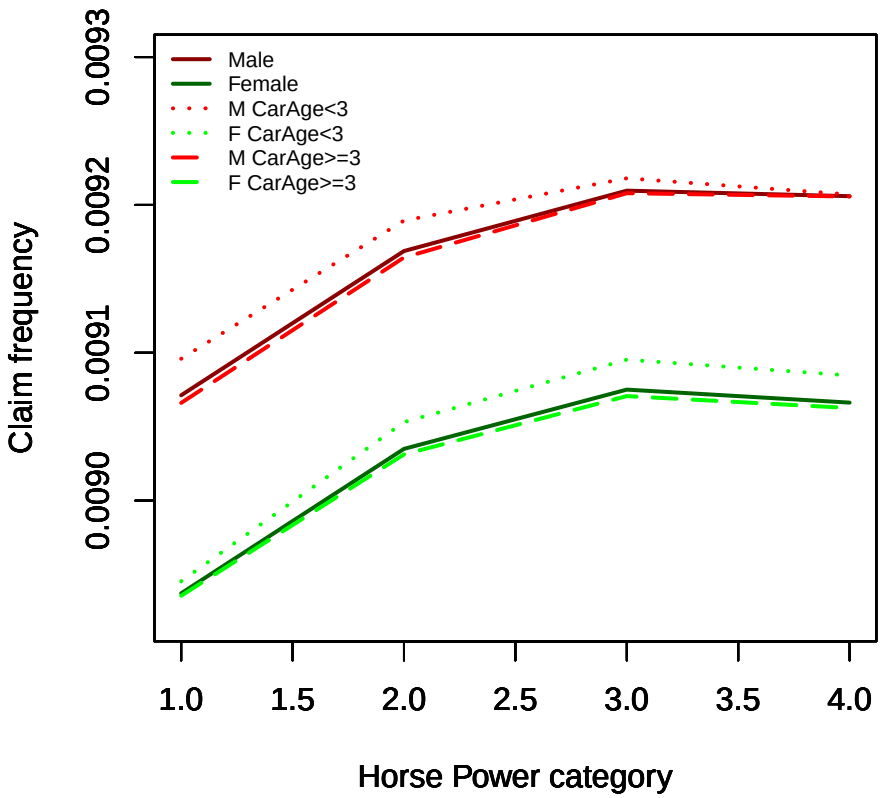
In this part we analyse the variable CarAge,Power and Gender.

Even if the car age was not a significative factor in the GLM analysis for a risk a 5%, it would have been considered at a risk of 10%. Hence we choose to include it in the analysis The "CarAge"" variable is discretized in 2 categories:

1- $CarAge < 3$

2- $CarAge >= 3$

You can find the results below.

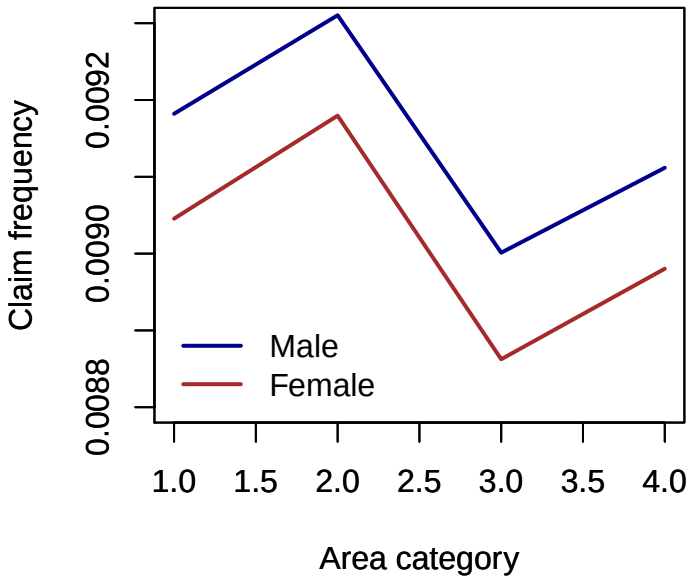


We can observe that the model associate the Male with higher claim frequencies. There is also a positive relation between the Horse power and the claim frequency. Those results have been found in the GLM analysis previously. This graph also shows the influence of the car age: customer with cars of less than 3 years have less claims than customers with older cars.

We can notice that compared to others factors, the influence of the car age is rather small, which can explain why it is not found significative by the GLM model at a risk of 5%. However one interesting thing to notice is that the difference in claim frequency between Male customers with a car age < 3 years and Male customers with older cars decrease for more powerful cars, while the opposite phenomenon occurs with female.

3.4.2 Gender and Area

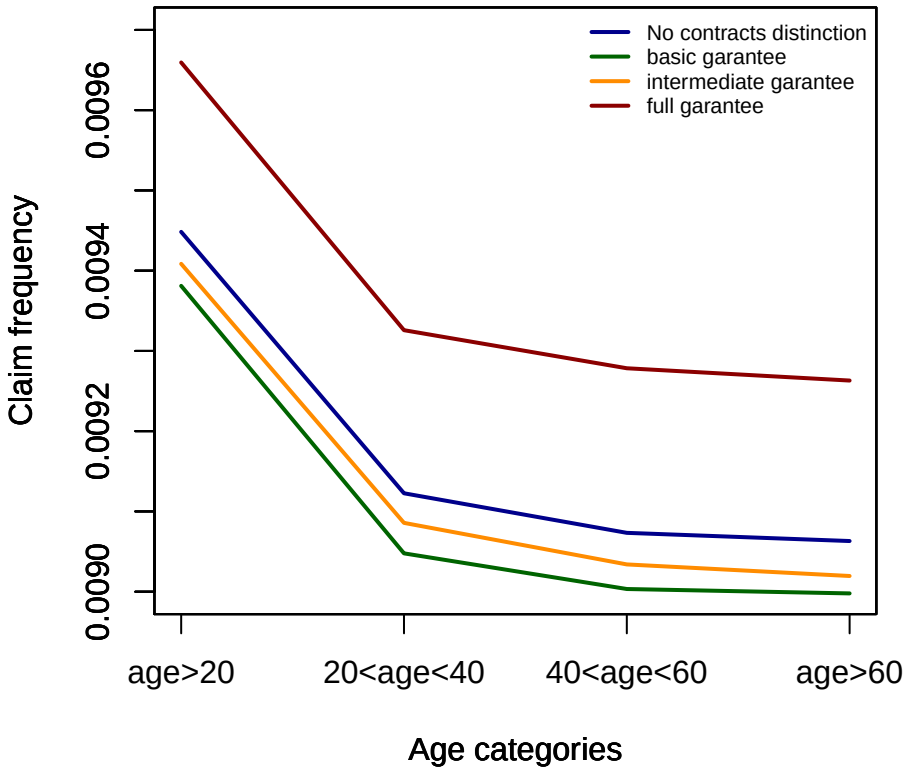
The results for the Area variable are plotted below.



The area 3 (countryside) is the region associated with the smallest claim frequency, similarly to the GLM analysis. The urban and mountain regions are almost at the same level, while the customers in suburban areas have the highest claim frequency.

3.4.3 DriverAge and Contract variables

Finally, we examine the influence of the age of the driver on the claim frequency, as well as the influence of the type of contract.



We can observe that both factors have a big influence on the probabilities of making a claim in the year. Younger customers makes more claims while higher customers with higher garantees will make more claims.

However the difference between having a basic garanty (contract 1) and an intermediate garantee is much lower than the difference between a full garantee and others types of contracts.

In the case of the age of the drivers, the youngest drivers (age<20) makes much more claims than others. There is little difference between the claim frequency of the customers in the categories 2020 0.0096594 0.0093811 0.0002784 2060 0.0092631 0.0089979 0.0002652

We can observe that for the 3 first age categories, the difference of claim frequency is very similar, while the difference is lower for customers more than 60 year old.

This could be explained by the fact that the older person have a higher *risk aversion*. This difference in risk aversion could for instance be explained by a change of mentality for this age class. A more detailed analysis of this phenomenon is outside the scope of the project.

3.5 Gradient boosting machine with poisson response

3.5.1 Introduction

The purpose of boosting is to sequentially apply a weak regression or classification algorithm to repeatedly modified versions of the data thereby producing a sequence of weak classifiers.[1].

[1] *The Elements of Statistical Learning*,Hastie,Tibshirani and Friedman,2009

Gradient Boosting [2] generalize the boosting methods by allowing the use of any differenciable loss function.

The model can be understood as a form of basis expansion, where the basis functions $b_m, m = 1; ..., M$ are the weak learners.Gradient boosting is typically used with decision trees as base learners.

For a poisson response variable the model is: $Log(\frac{Nbclaims}{Exposure}) = \sum_{m=1}^{M} v\beta_m b(x, \gamma_m)$

Where the β are weights, γ_m the parameters relative to the tree itself and v a regularizer called learning rate.

Ideally, we would like to find the solutions of

$$\min_{\beta_m, \gamma_m, v, M} \sum_{i=1}^{N} L\Big(y_i, \sum_{m=1}^{M} v\beta_m b(x_i, \gamma_m)\Big)$$

But this problem is too computationally intensive to be solved directly.

Instead we opt for a greedy algorithm, the Gradient Boosting algorithm, which we won't describe here.

In our case, the Poisson deviance is used as loss function.

We actually use a stochastic gradient boosting algorithm. Similarly to the random forest, each tree is fitted on a subsample of the training set drawn at random without replacement. This trick has been proved to substantially improve the performance of the model.[3]

3.5.2 Tuning

The 3 main tuning parameters of a gradient boosting machine model are:

- **The number of iterations M**: The number of trees to add.
- **The complexity of the tree (size or interaction depth)**: determine the order of the interactions taken into account. If J = 2 only main effects will be taken into account.
- **The learning rate v (shrinkage)** : Scale the value the contribution of each trees. A smaller learning rate will require a high number of iterations for a given training error.

This 3 parameters control the complexity of the model and help to prevent overfitting.

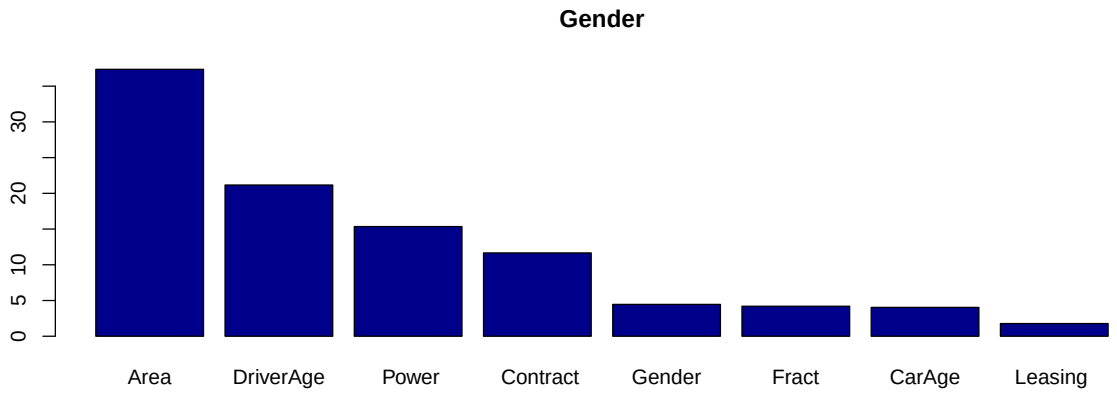
The number of trees is determined by first growing a large number of trees (4000) and observing the evolution of the performance.

The others parameters are tuned using a grid search.

The relative importance plot of the corresponding model is plotted below.

[2] Greedy Function Approximation: A Gradient Boosting Machine, Friedman,1999

[3] "Stochastic Gradient Boosting",Friedman, 2002



We can observe in the relative importance plot that the driver age,area, contract and powerhorse are the variables who have the most impact in the contruction of the trees.

4. Conclusions

The objective of this project was to determine the most relevant predictors of the claim frequency. After describing the variable and modeling the claims frequency, some variables appears more important than others.

Indeed, The type of contracts,age of the drivers and area are the most important predictors and should be primarily used for the segmentation of the customers, although the GLM model did not select the age variable.

Compared to those variables, the gender variable seems to have a slightly lower impact on the claim frequency. However it is also an important predictors which have been used by all the model.

The splitting of the premium and the car Age have a rather minor influence on the claim frequency, and should be used only for a finer segmentation.

Finally, The leasing does not appear to be an important factor and should not be an important criterion for segmentation.

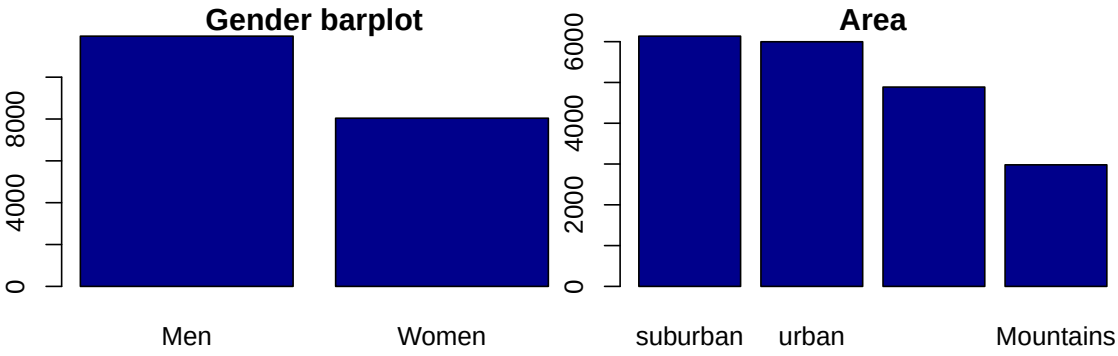
In conclusion, we are now able to have a segmentation of the customers based on their claim frequency, however an analysis of the claim severity will have to be done before being able to determine a grid of customized premiums.

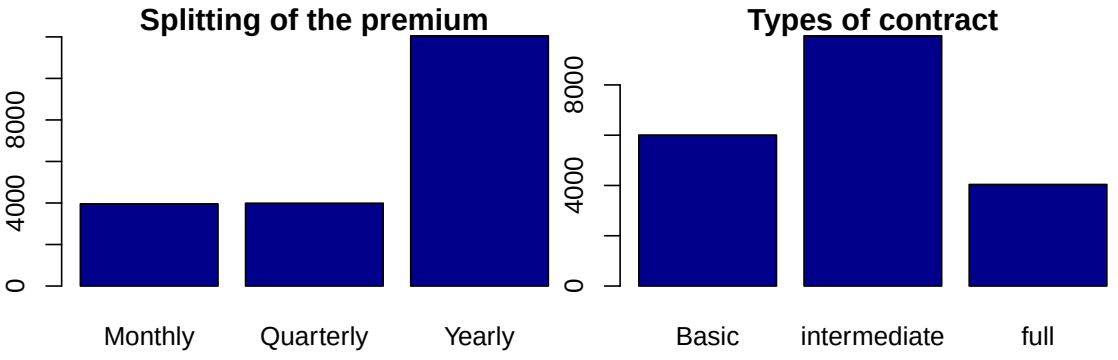
5.Annexe

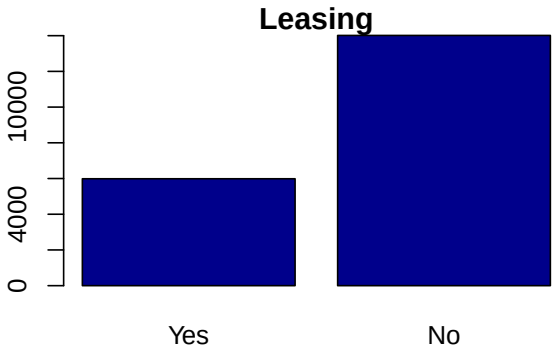
5.1 Descriptive statistics of the test set

Table 6: Descriptive statistics of the test set

	DriverAge	CarAge	Exposure
Mean	43.35	5.04	3.01
Median	43.00	5.00	2.98
Standard deviation	14.30	2.57	1.47
Q25	33.00	3.00	1.98
Q75	53.00	7.00	4.01
Min	18.00	0.00	0.08
Max	82.00	16.00	8.88







5.2 Mean and standard deviation of categorical variables, by categories (training set)

Table 7: Claim Frequency by Gender

	Mean	Standard deviation
Men	0.0096	0.0953
Women	0.0081	0.0571

Table 8: Claim Frequency by Area

	Mean	Standard deviation
suburban	0.0085	0.0561
urban	0.0109	0.1148
countryside	0.0080	0.0731
Mountains	0.0080	0.0579

Table 9: Claim Frequency by Leasing

	Mean	Standard deviation
Yes	0.0092	0.0623
No	0.0089	0.0892

Table 10: Claim Frequency by Power

	Mean	Standard deviation
low	0.0074	0.0569
normal	0.0092	0.1003
intermediate	0.0098	0.0754
high	0.0091	0.0590

Table 11: Claim Frequency by Fract

	Mean	Standard deviation
Monthly	0.0091	0.0586
Quarterly	0.0089	0.0604
Yearly	0.0090	0.0942

Table 12: Claim Frequency by Contract

	Mean	Standard deviation
Basic	0.0080	0.0672
intermediate	0.0088	0.0938
full	0.0111	0.0703

5.3 Extreme Gradient Boosting

The Extreme Gradient Boosting (XGBoost) is an implementation of Gradient Boosting. XGBoost is sometimes more than 10 time faster than others gradient boosting implementations, and generally have a better predictive performance.

The main features of XGBoost are listed below:

1- Instead of having a tree learn the negative gradient at each stage, the second order Taylor expansion of the loss function is learnt.

2- A regularization term is added to the Taylor expansion so that the new tree minimize under constraint the loss function, which reduces overfitting.

3- Introduce the concept of structure score as a measures how good a tree structure is, when growing a tree the structure score is used as a splitting criterion.

4- Columns sub-sampling: in addition to take the observations on which each tree is trained, a number of features is sampled, similarly to random forests.

5- From the engineering side, a lot of tricks are used to speed up computations (Histogram-based method, sparsity-aware algorithm, better use of multiprocessing and more)