**The Final Project**

**Professor: Marco Montes de Oca**

Submitted by: Tianyu Zhang, Min-Chi Tsai, Tuhena Sen

Northeastern University

Winter 2020 CPS Quarter

ALY6020 22757: Predictive Analytics

Friday, April 8

**Abstract:** In the whole course, we mainly studied a variety of prediction models, including regression model, nearest neighbours, tree model, neural networks, and naïve Bayes. We will learn the algorithms and concepts of these models and apply them to real life work. In this final project, we will apply these models to our dataset. Our main purpose is to establish a price model so that the insurance company can obtain the highest interest. The methodology mainly includes logistic regression, KNN, SVM, random forests and Naïve Bayes. By comparing the final accuracy and AUC of these models, we will choose the most suitable price model for our dataset.

**Introduction**

The dataset comes from AIcrowd and is provided by a large auto insurance provider in a European country and contains real historical auto insurance policies in the past four years. It consists of 26 columns and around 228k rows.

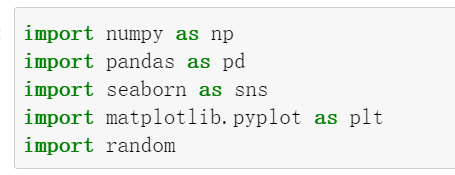
In the final project, we would like to analyze this dataset to identify the following questions: How to make the insurance company obtain the highest profit by establishing the pricing model?

We will use Python to do our following analysis and hope to answer the questions we address above. We will implement our final project from the following five aspects: EDA, methodology, data preparation and processing, model construction and conclusion.

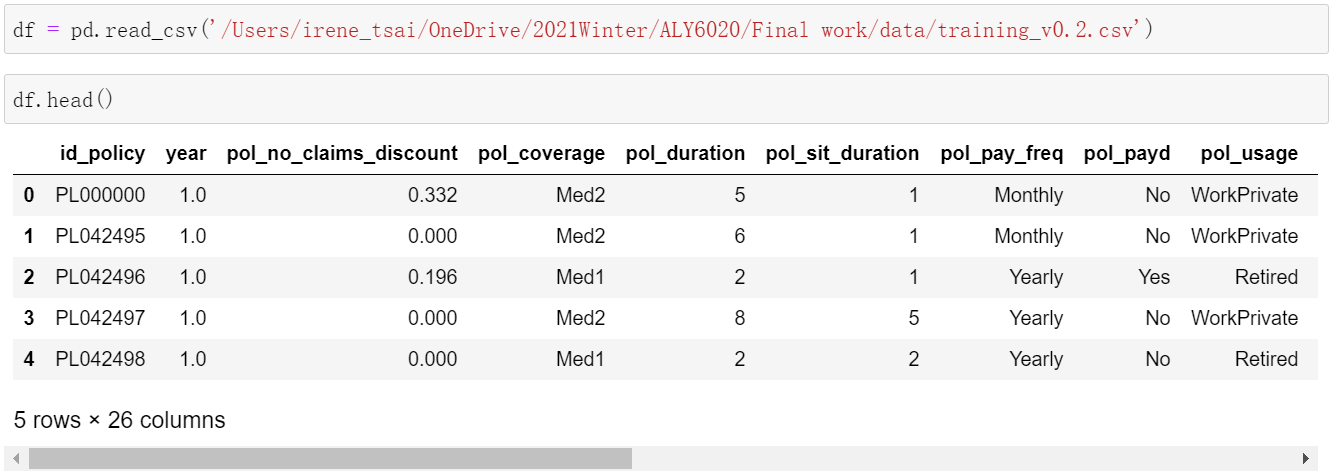
**Analysis and Results**

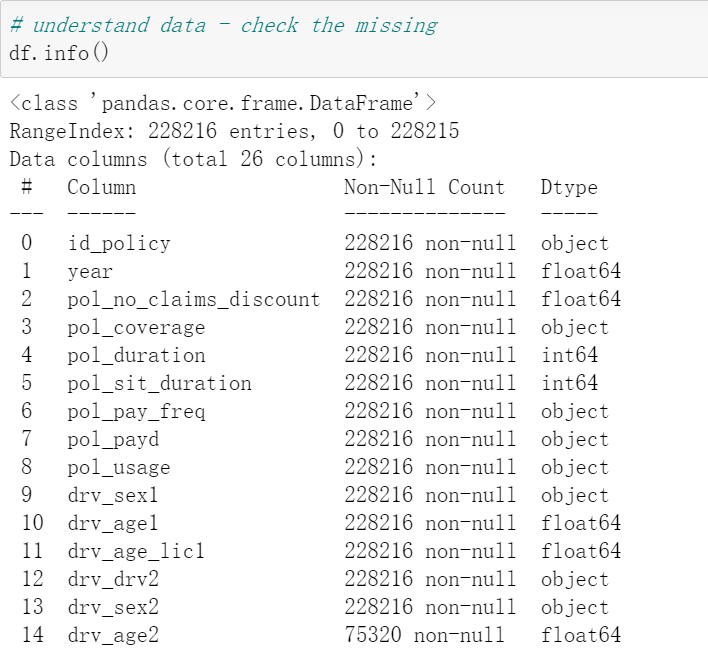
**Part 1: EDA**

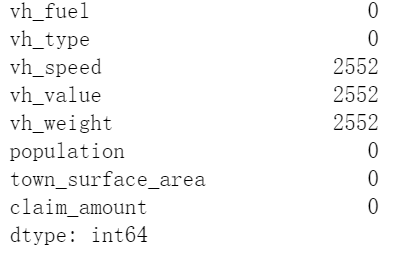
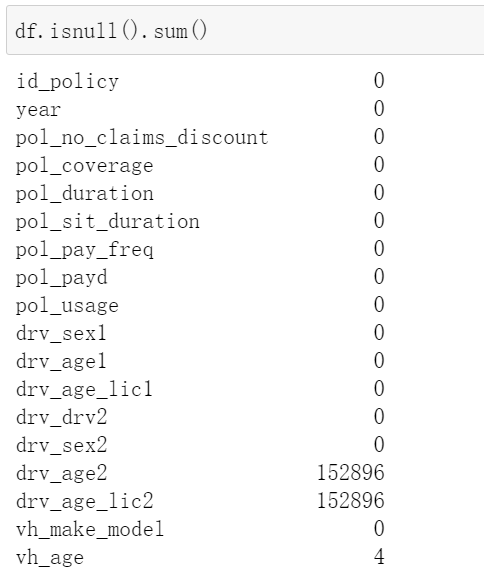
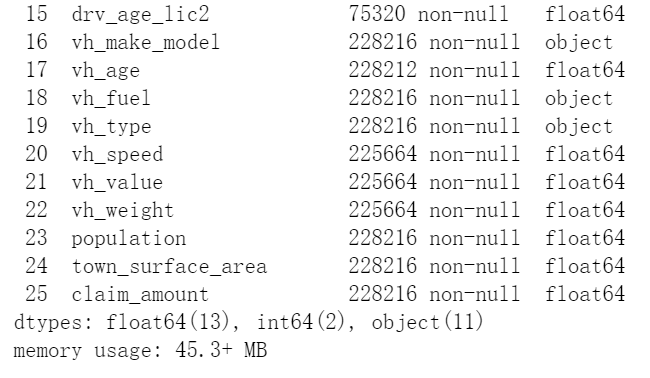
First, import the necessary libraries.



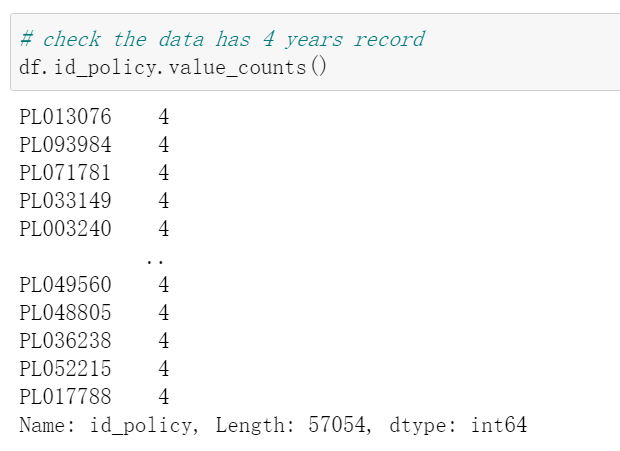
Then, import our dataset and display the information of our dataset. We can find that this data has 26 columns, in which ‘claim\_amount’ is the outcome variable we want to analyze, that is y.







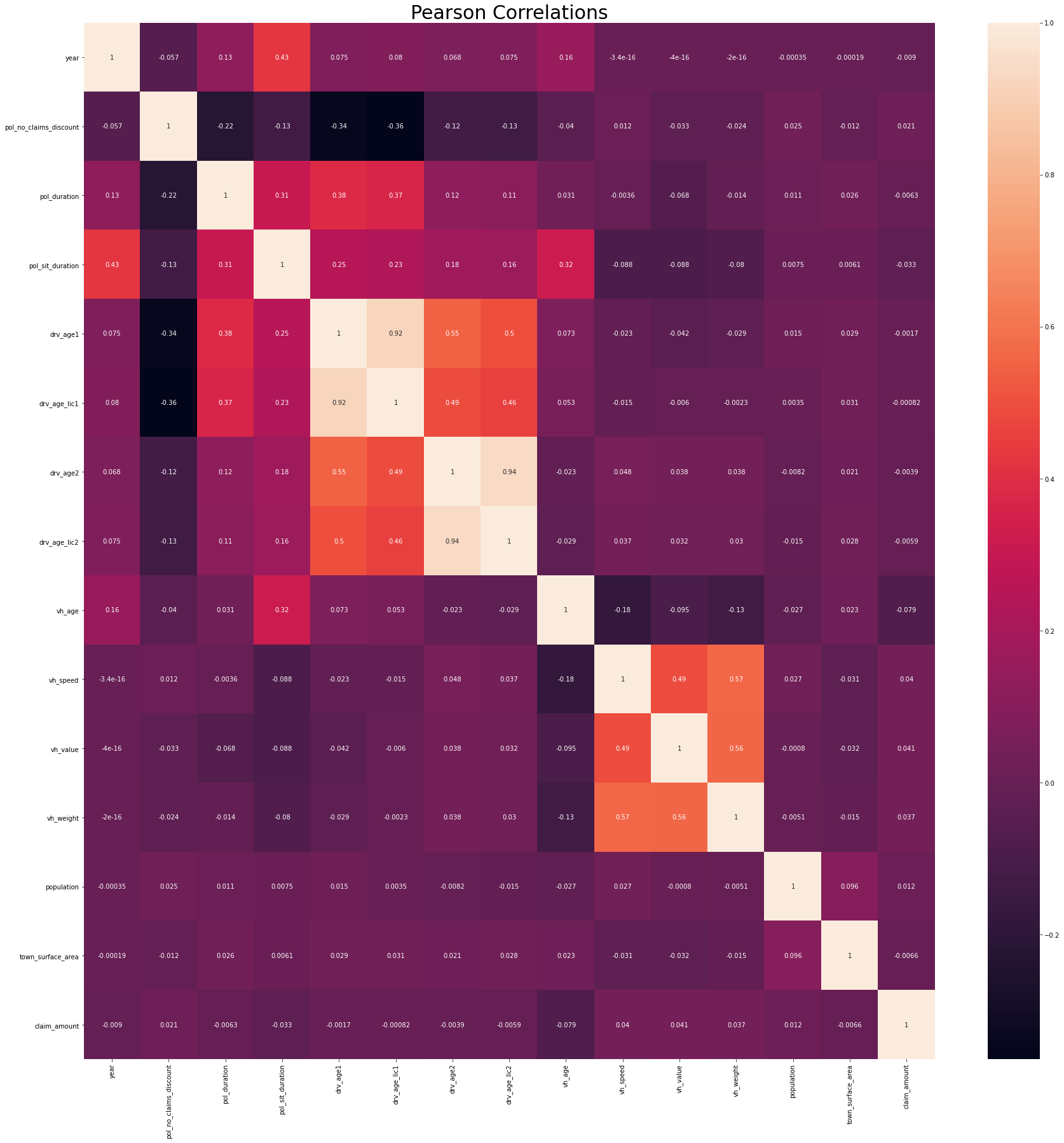
There are four years records, and we are going to predict the price model for the fifth year.

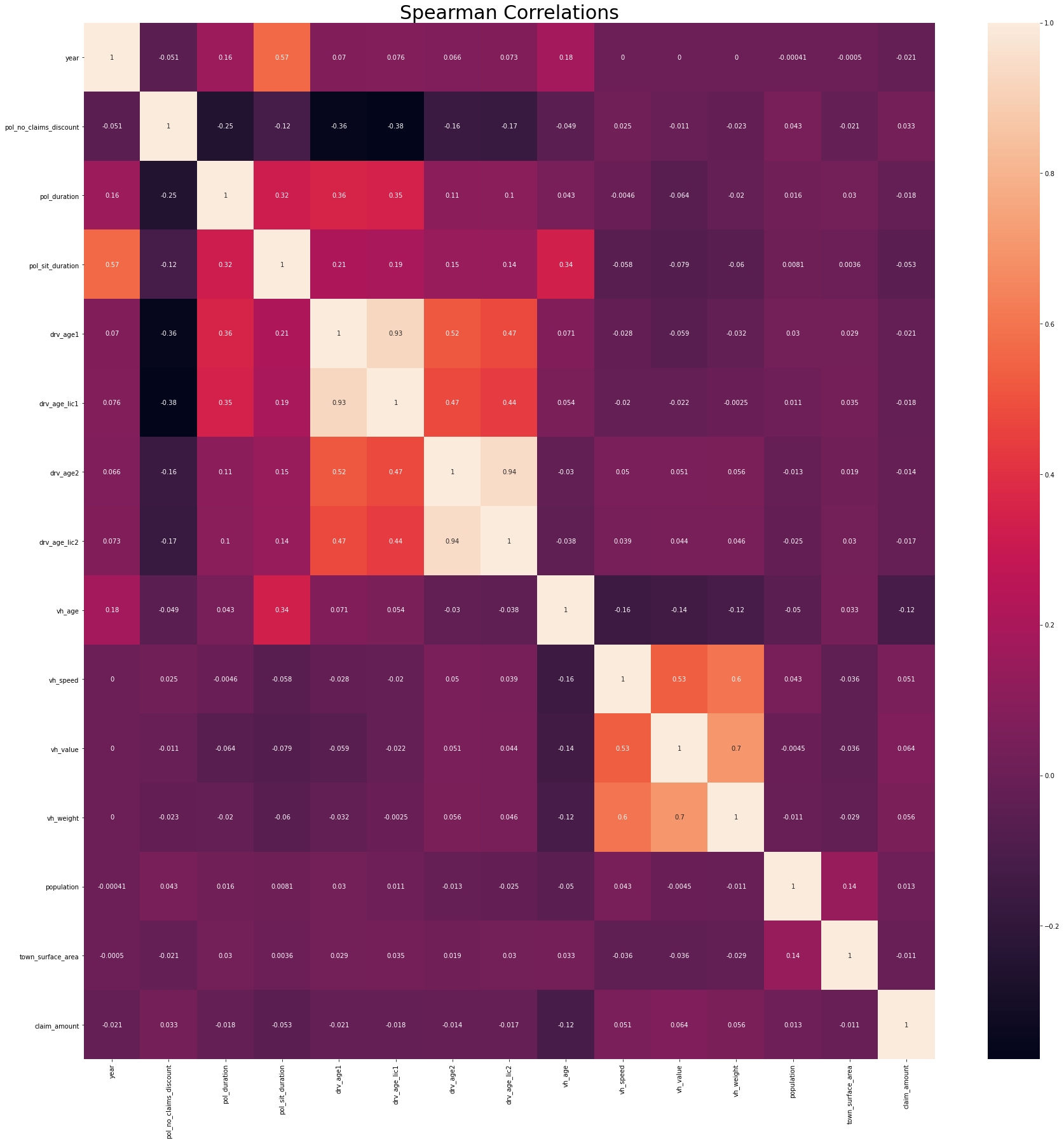


Next, we check the correlation of different variables. From the color change, we can find the correlation between the various variables. In order to better understand the different variables, we will explain some variables here.

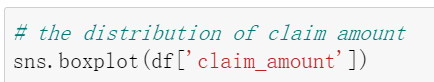
‘pol\_no\_claims\_discount’ indicates the claim value. The driver's claim coefficient increases by 0.2, ranging from 0 to 1. ‘pol\_coverage’ means different types of coverage. ‘pol\_usage’ shows how the driver uses the vehicle (4 types). ‘vh\_make\_model’ hash indicating the brand and model of the vehicle. The remaining variables will be easier to understand, so no explanation will be given.

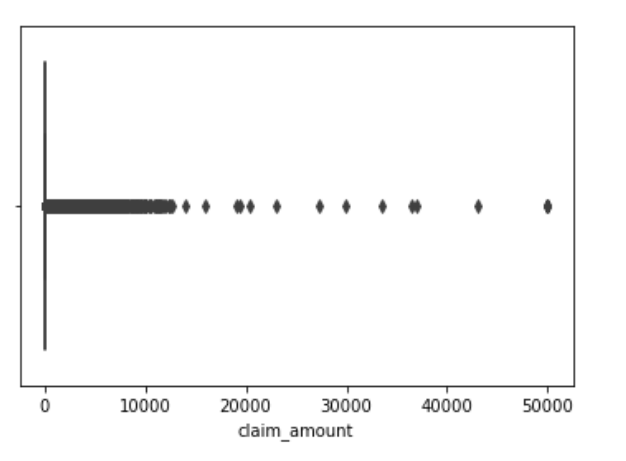
We use two different ways to check its correlation, which is Pearson method and Spearman method. We find that Spearman correlation is higher than Pearson correlation, which means that our final model can consider nonlinear model more.



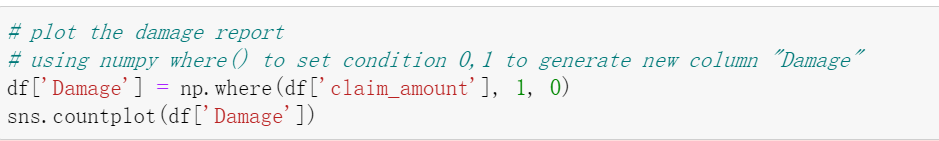


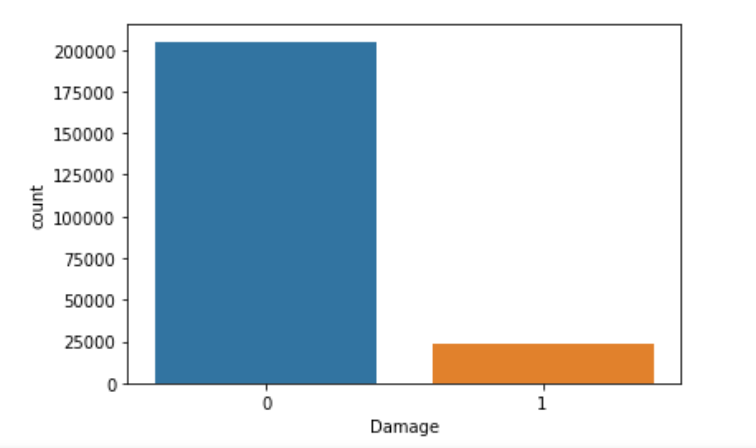
Then, we will conduct EDA according to the outcome variable (claim\_amount), so as to better understand the relationship between it and other variables. First of all, we display the amount range through using boxplot. We can find that most amount is under 10000 euros.



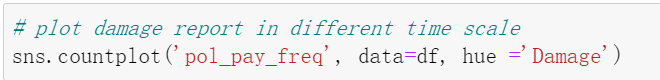


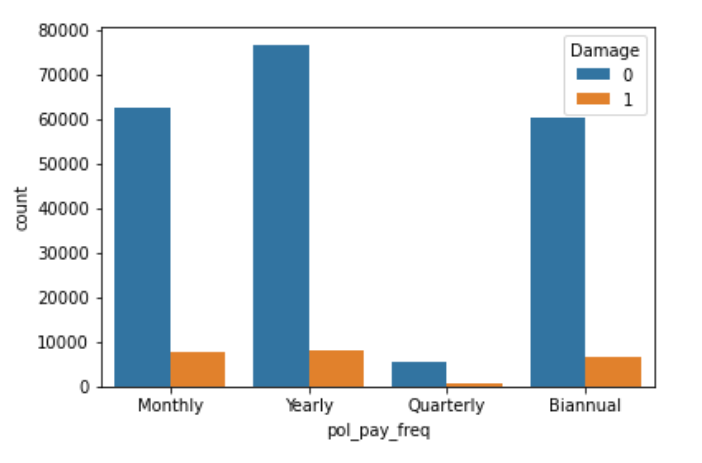
Then, we create a new column named ‘damage’, this is to be able to have a clearer understanding of whether the insurance company has compensation. If so, it will be 1. If not, it will be 0. We discover that most cases have no compensation, only part of them have insurance compensation.



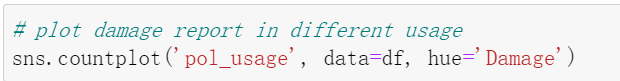


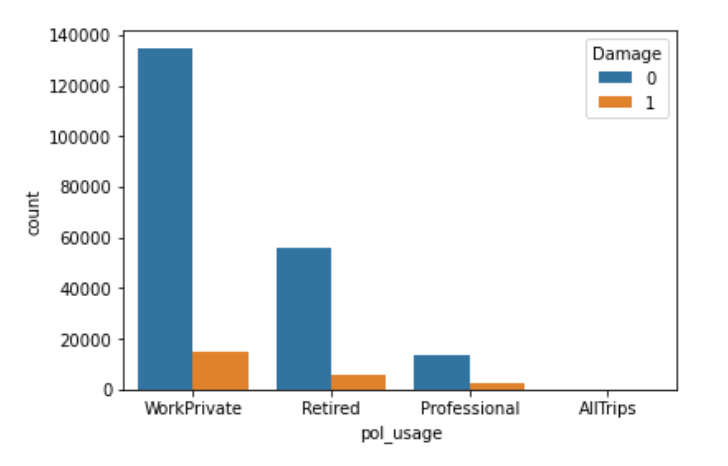
Next, let's look at the correlation between the compensation part of the case and other variables. The first is the relationship with compensation frequency. We can find that the frequency of quarterly compensation is very small, and the remaining three frequencies are relatively close.



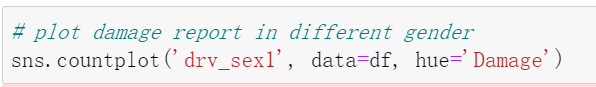


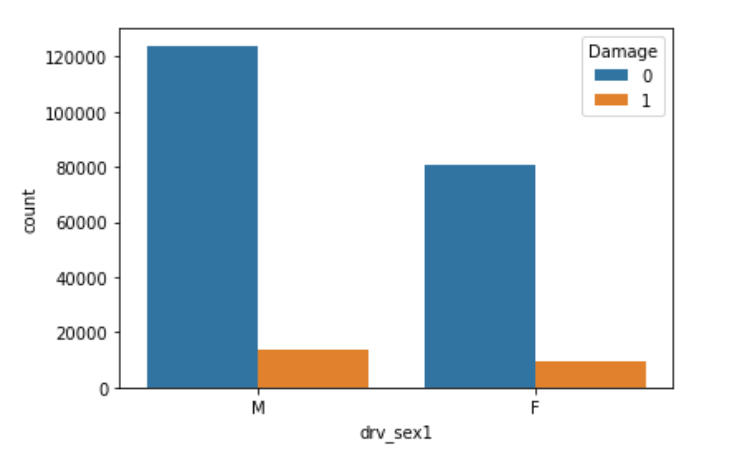
Second, we study the correlation between claims and vehicle use. We observe that most of the cars are used for work and private use, some are used by retirees, and a few are used for professional vehicles.





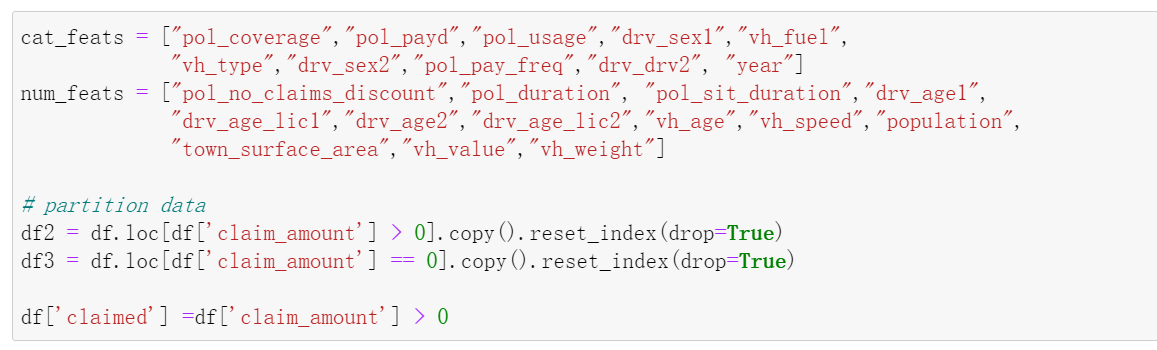
Third, we count the number of claims from different gender. We can see that most of the people who drive cars are still men.



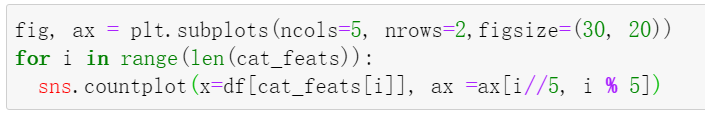


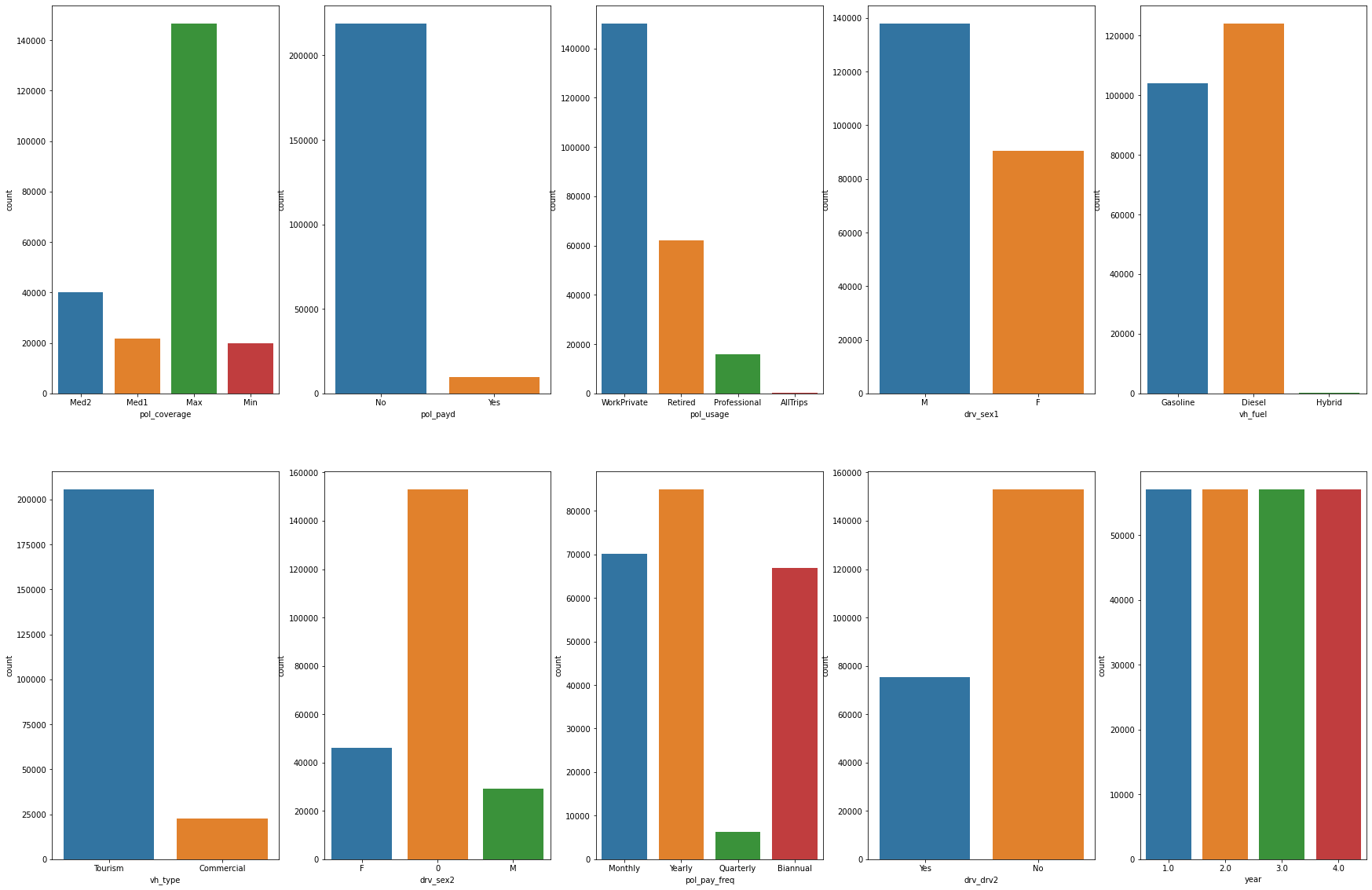
After looking at the relationship between several specific variables, let's look at the relationship between all variables and target.

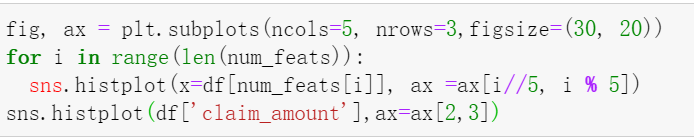
First of all, we classify different variables to facilitate the next step of plotting. We divide all variables into category and number classes according to their types. Then create a new dataset, which only contains insurance company claims.

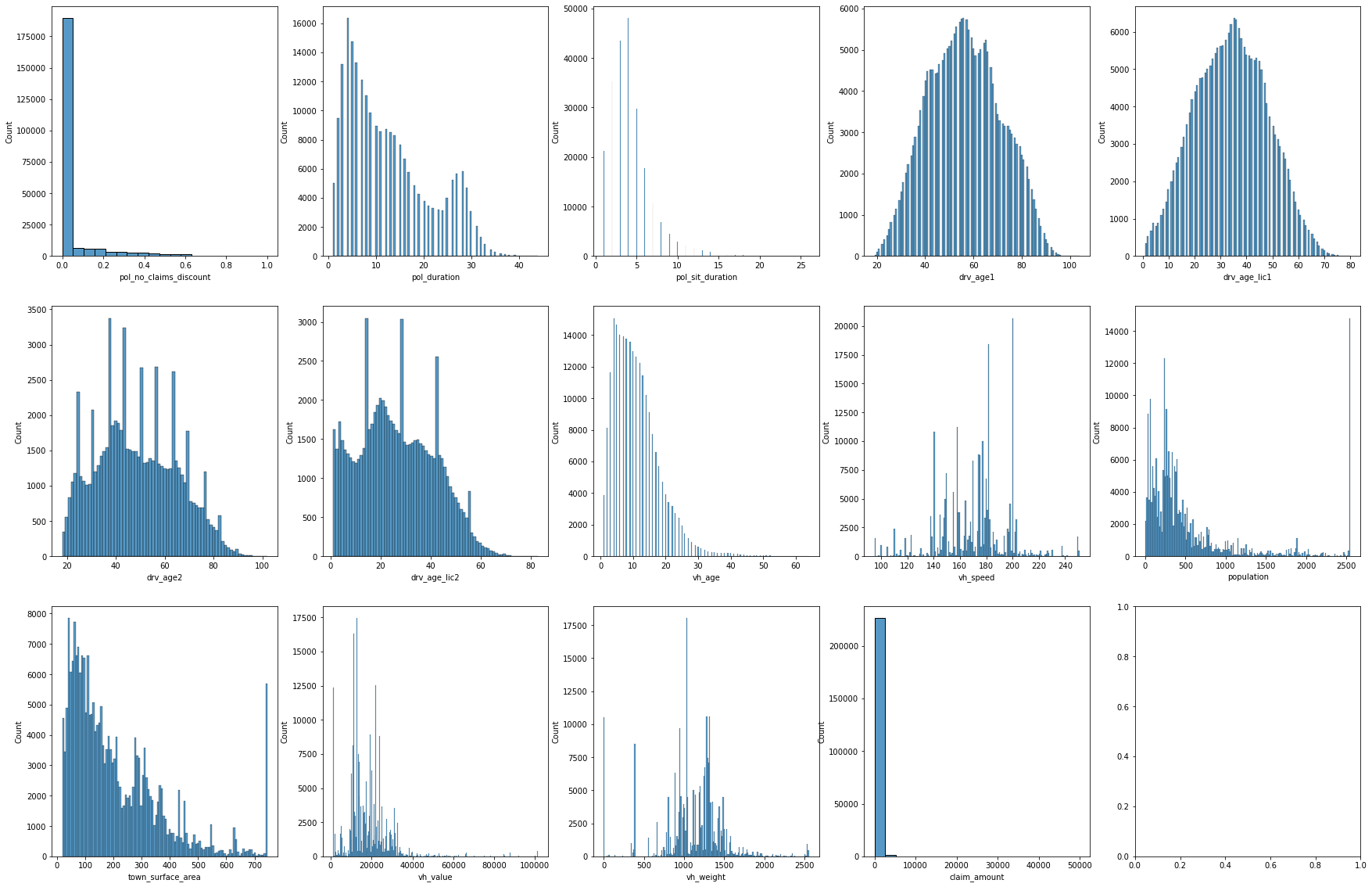


Plotting different figures according to target (claim amount).

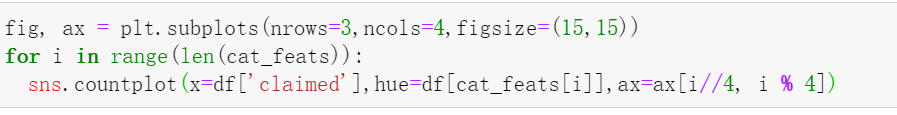


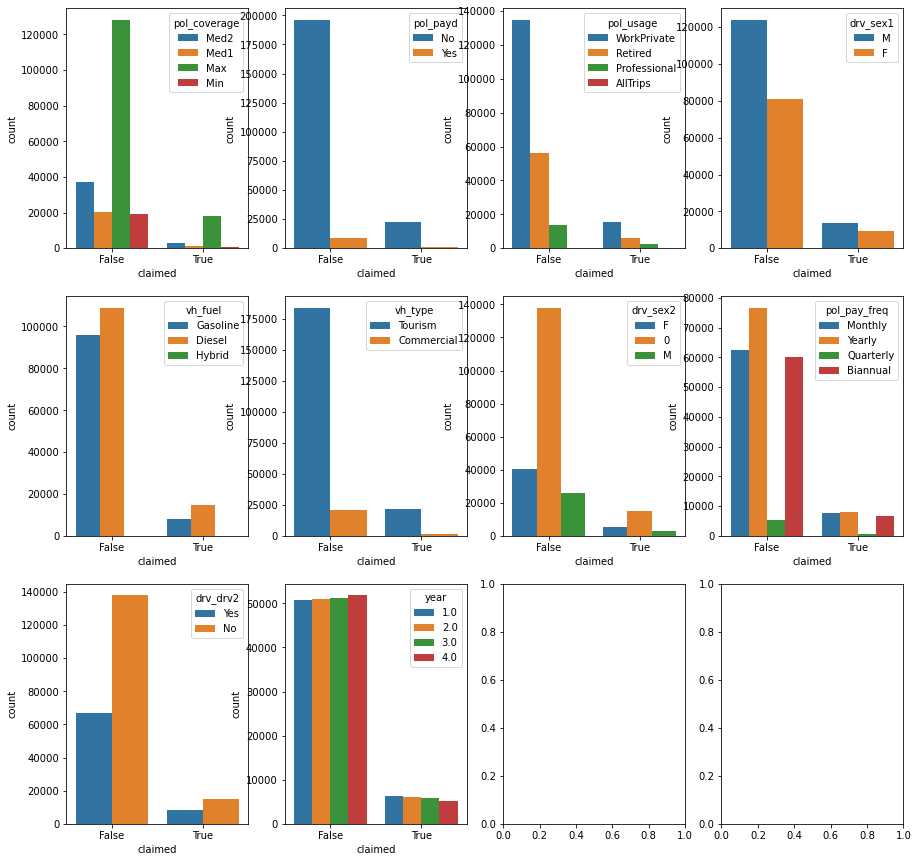






Then, we plotting the figures according to our new dataset which only contains the claim > 0.





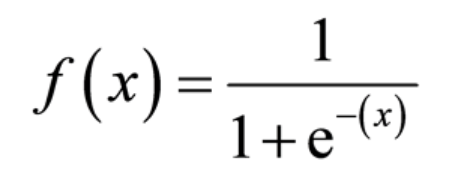
Through the above EDA, we have a certain understanding of our dataset, including the relationship between each variable and target. In the next part, we will introduce the method we use in our final project.

**Part 2: Methodology**

In this part, we will mainly introduce the prediction analysis methods we used in the project, so as to make better preparation for our models.

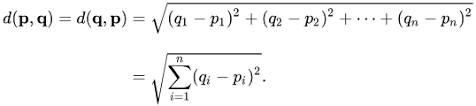
1. logistic regression algorithm

The first method we use is logistic regression algorithm. Logistic regression is a machine learning method used to solve the binary (0 or 1) problem, which is used to estimate the possibility of something. For our project, our target is a new ‘damage’ column based on claim amount, which is already a binary. The sigmoid function is:



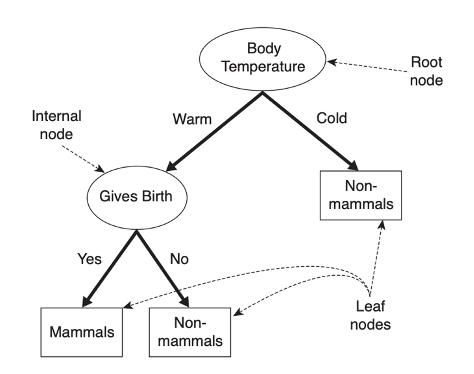
1. K-nearest neighbor

The second method we use is K-nearest neighbor. K-nearest neighbor method is a method based on classification and regression. The selection of K value, distance measurement and classification decision rules are the three basic elements of k-nearest neighbor. It is easy to use and has a fast calculation time, but the disadvantages are obvious, we must find the best K value. (Schott, 2019). The algorithm is:



1. Random forest

The random forest constructs multiple decision trees. When a sample needs to be predicted, the prediction results of each tree in the forest for the sample are counted, and then the final result is selected from these prediction results by voting method. Its ease of use and flexibility drive its adoption because it can handle classification and regression problems. (IBM, 2020). Random is reflected in two aspects, one is to take features at random, the other is to take samples at random, so that each tree in the forest has both similarities and differences.



*source* [*https://www-users.cs.umn.edu/~kumar001/dmbook/ch4.pdf*](https://www-users.cs.umn.edu/~kumar001/dmbook/ch4.pdf)

1. Naive Bayes

Naive Bayes is a classification technique based on Bayes theorem, which assumes that the predictive variables are independent. (Ray, 2017). Naive Bayes has stable classification efficiency, good performance for small-scale data, can handle multiple classification tasks, and is not sensitive to missing data, and the algorithm is relatively simple. However, it requires the assumption that the attributes are independent of each other, which is often not true in practical applications. The formula is: P (X, Y) = P(X)P(Y).

1. SVM

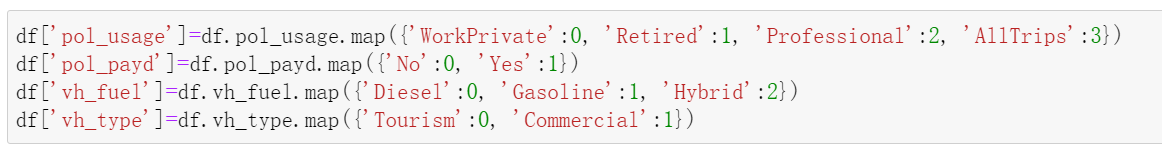
The last method we use is support vector machines. SVM is a binary classification model. Its basic model is the linear classifier with the largest interval defined in the feature space, which makes it different from perceptron. SVM also includes kernel techniques, which makes it a non-linear classifier in essence. The learning strategy of SVM is interval maximization. The basic idea of SVM learning is to solve the separation hyperplane which can divide training data set correctly and the geometric interval is the largest. Hyperplane is a decision boundary which is helpful to classify data points. (Gandhi, 2018).

In the next part, we will do the data preparation work and build a prediction model according to the correlation results, hoping to solve the goal we want to achieve in the final project.

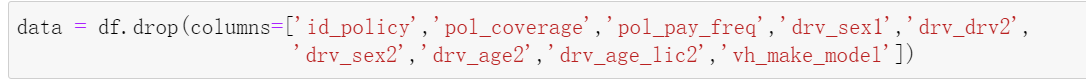
**Part 3: Data Preparation and Processing**

Before we build the model, we need to clean up the data and determine the variables.

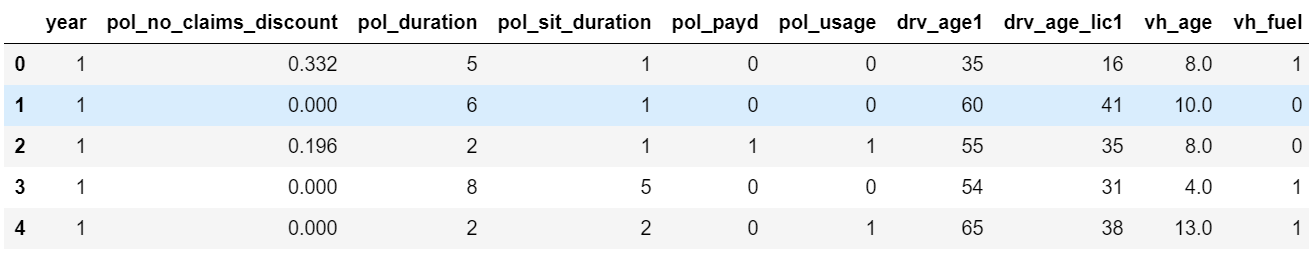
In the previous data information, we can also see that some variables are not numbers, but text. Therefore, we need to convert its format.



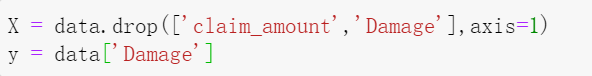
According to the previous correlation diagram, visual analysis and the definition given by each column, we filter the dataset. Finally, it is decided to discard the following variables.

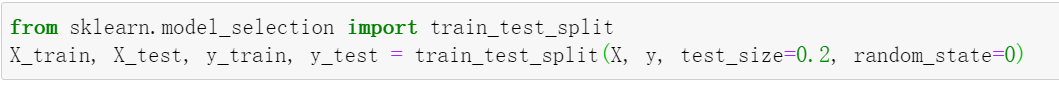


Then, we display the new dataset, and it is all numerical and all the variables are needed.



Next, we determine x, y variables. The target variable is ‘damage’, and the rest variables are our x variables except the claim\_amount. And we split our dataset into 8:2 train data and test data.

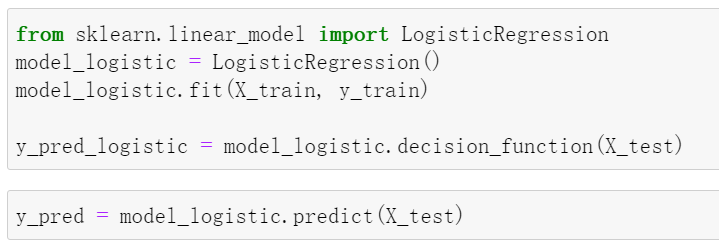




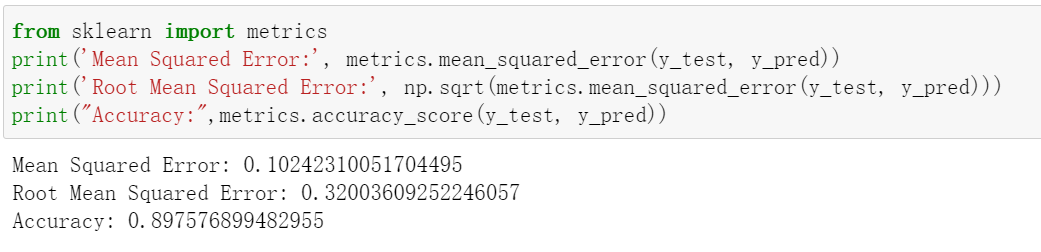
**Part 4: Model construction**

1. logistic model

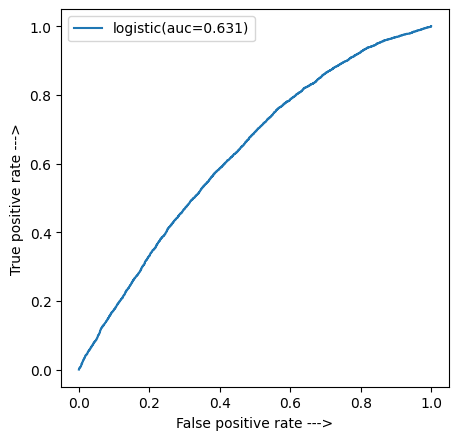
We use the following command to build the model and get the prediction.



We use the MSE, R-Square, accuracy to test our model. We obtain that although MSE is very small, R2 is not ideal. The accuracy of the model is 0.8976 which is relatively high.

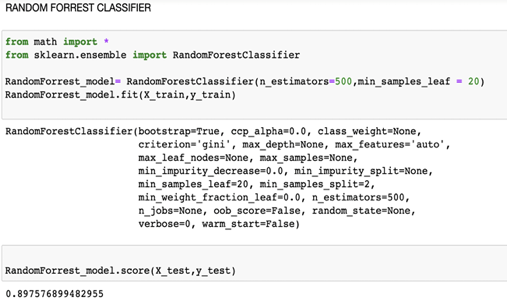


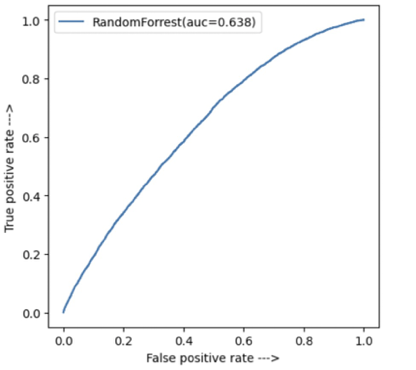
In order to better detect our model, we draw the ROC curve. And the AUC is 0.631 which is not good.



1. Random Forest

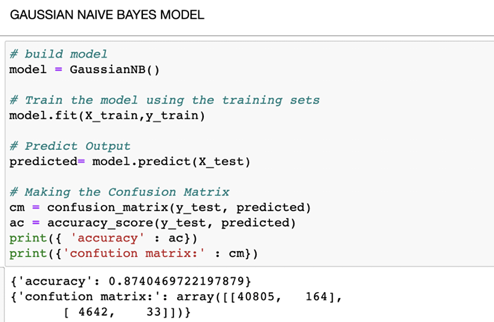
We used RandomForestClassifier to build a random forest model from the sklearn library. We tried couples of models and compared the accuracy. Finally, the model has 500 nodes and 20 minimum sample leaves. The accuracy of this model is 0.897 with auc equal to 0.64.

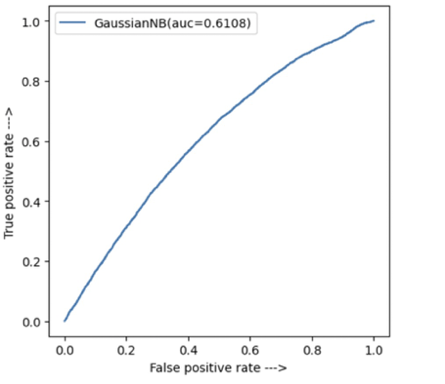




1. Gaussian Naive Bayes

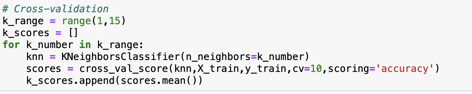
We used the Gaussian Naive Bayes classifier from Scikit Learn package. Normally, Gaussian Naive Bayes classifier is widely used in sentiment analysis. In this case, we want to compare the results of different models so we perform this model. True Negative equal to 40805, while True Positive equal to 33. The accuracy is 0.874, and auc equal to 0.61.

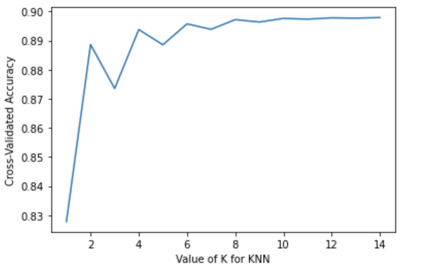




1. KNN classifier

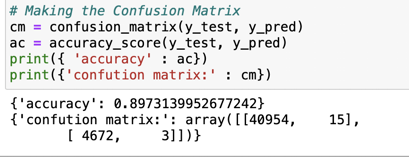
In our case, the first step of KNN classifier model building is cross-validation. Firstly, we set the range from 1 to 10, and we found out the trend still goes up. Thus, we expand the range to 15 and get the following picture. It is clear that the trend turns flat when k over 11.

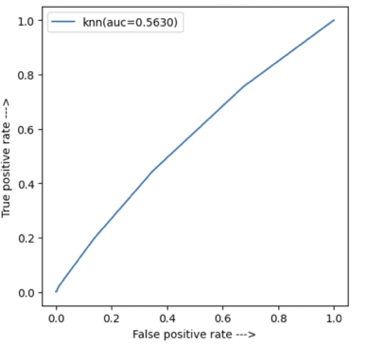




We train the model with k equal to 12. The accuracy is close to 0.9 and the auc equal to 0.56. . True Negative equal to 40954, True Positive equal to 3.

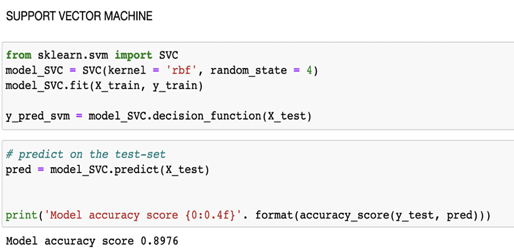


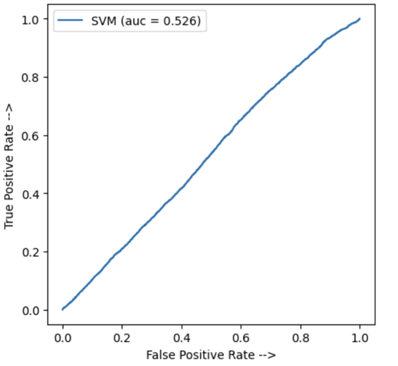




1. Support Vector Machine (SVM)

SVM is a popular supervised machine learning algorithm, which could help capture complicated relationships between data points . The accuracy of SVM is close 0.9 and the auc equal to 0.52.





**Conclusion**

1. We have chosen accuracy score and roc-auc score as our model selection parameter.

2. In our dataset we got high accuracy but low roc auc score because for a given threshold we can calculate accuracy which is proportion of true positives and negetives whereas auc computes how true positive rate and false positive trade off.

3. To select the best model out of all 5, we have checked that out of accuracy and roc curve parameter Roc-auc score is a better parameter

4. We chose RandomForrest as the best model for predicting the insurance pricing of the motor company based on below table.

|  |  |  |
| --- | --- | --- |
| **MODEL** | **ACCURACY** | **ROC -AUC SCORE** |
| LOGISTIC | 89.7 | 0.631 |
| RANDOM FORREST | 89.7 | 0.638 |
| GAUSSIAN NB | 87.4 | 0.61 |
| KNN | 89.7 | 0.563 |
| SVM | 89.7 | 0.526 |

Reference

1.Gandhi, R. (2018). Support Vector Machine — Introduction to Machine Learning Algorithms. Retrieved from: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>.

2.IBM Cloud Education. (2020). Random Forest. Retrieved from: <https://www.ibm.com/cloud/learn/random-forest>.

3.Motor insurance market simulation. (2020). Retrieved from: <https://www.aicrowd.com/challenges/insurance-pricing-game#dataset>.

4.Schott, M. (2019). K-Nearest Neighbors (KNN) Algorithm for Machine Learning. Retrieved from: <https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26>.

5.Ray, S. (2017). 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R. Retrieved from: <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/#:~:text=Naive%20Bayes%20Model-,What%20is%20Naive%20Bayes%20algorithm%3F,presence%20of%20any%20other%20feature>.