FUNIX
Data Science Course

Final Project Report

VEHICLE INSURANCE CROSS-SELL PREDICTION

Presented by Duc Ha Minh
A *Data Science Course* student at Funix

Vehicle insurance cross-sell prediction

Business Understanding

Data Understanding

Exploratory Data Analysis

Modelling & Model Evaluation

Conclusion

Business Understanding

Business Understanding

Problem:
The client is an Insurance company that has provided Health Insurance to its customers. Now, the firm needs a model to predict whether the ex-policyholders will still be interested in Vehicle Insurance provided by them.

Goal:
Identify customers' insights in order to set a marketing and communication plan to approach potential customers. The firm's business model and its revenue will hopefully be optimized.

Objective:
Predict whether groups of ex-customers would be interested in Vehicle Insurance based on their demographics, tyles of vehicles, etc.

Criteria for a successful or useful outcome to the project:
The precision of the algorithm and the number of useful insights.

2 Data Understanding

Data Understanding

- These datasets were downloaded from Kaggle.

- There are two datasets included ‘train.csv’ and ‘test.csv’. These files have no null value in the dataset.

- The ‘train.csv’ had 12 columns and 381,109 rows. This file had 5 categorical columns (Gender, Driving_License, Previously_Insured, Vehicle_Damaged and Response) and 7 numerical columns.

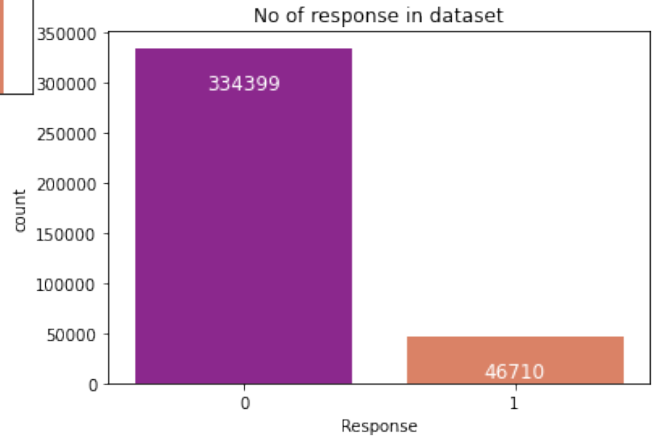
- The ‘test.csv’ had 127,037 rows and 11 columns which were similar column with the ‘train.csv’. However the ‘test.csv’ did not have ‘Response’ column which is the prediction of output.

	count	mean	std	min	25%	50%	75%	max
id	381109.00	190555.000	110016.83620	1.000000	95278.000000	190555.000000	285832.00000	381109.00000
Age	381109.00	38.822584	15.511611	20.000000	25.000000	36.000000	49.00000	85.00000
Driving_License	381109.00	0.997869	0.046110	0.000000	1.000000	1.000000	1.00000	1.00000
Region_Code	381109.00	26.388807	13.229888	0.000000	15.000000	28.000000	35.00000	52.00000
Previously_Insured	381109.00	0.458210	0.498251	0.000000	0.000000	0.000000	1.00000	1.00000
Annual_Premium	381109.00	30564.3895	17213.155057	2630.000000	24405.000000	31669.000000	39400.00000	540165.00000
Policy_Sales_Channel	381109.00	112.034295	54.203995	1.000000	29.000000	133.000000	152.00000	163.00000
Vintage	381109.00	154.347397	83.671304	10.000000	82.000000	154.000000	227.00000	299.00000
Response	381109.00	0.122563	0.327936	0.000000	0.000000	0.000000	0.00000	1.00000

Data Understanding

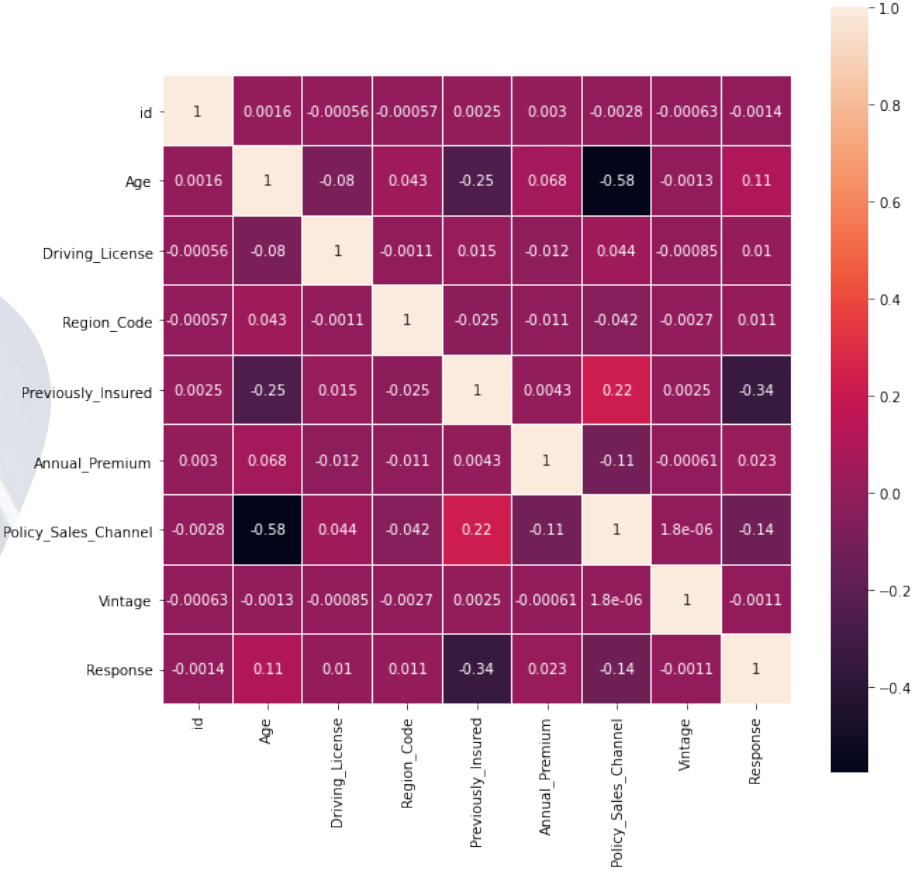
The “train.csv” had:

- The min of Annual Premium is 2,630$ while the max is 540,165$.
- Dataset has 260,089 male and 175,020 female.
- There is a big gap of customers’ response when 46,710 of them are willing to pay for insurance.



3 Exploratory Data Analysis

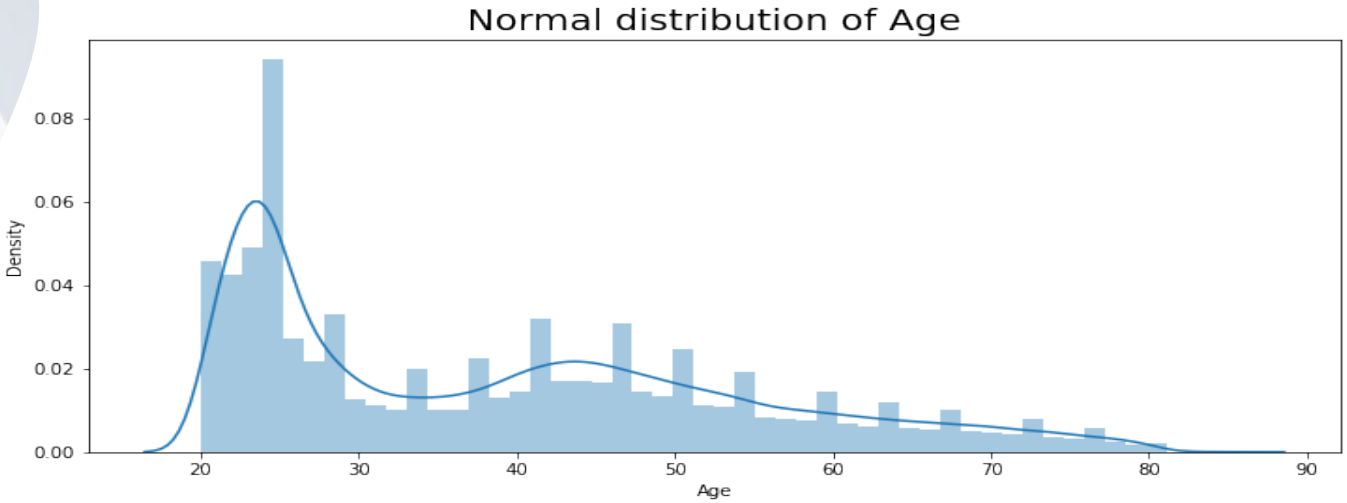
Exploratory Data Analysis



Almost of variables do not have strong correlation with ‘Response’, so the algorithm applied in this dataset must have high accuracy.

Exploratory Data Analysis

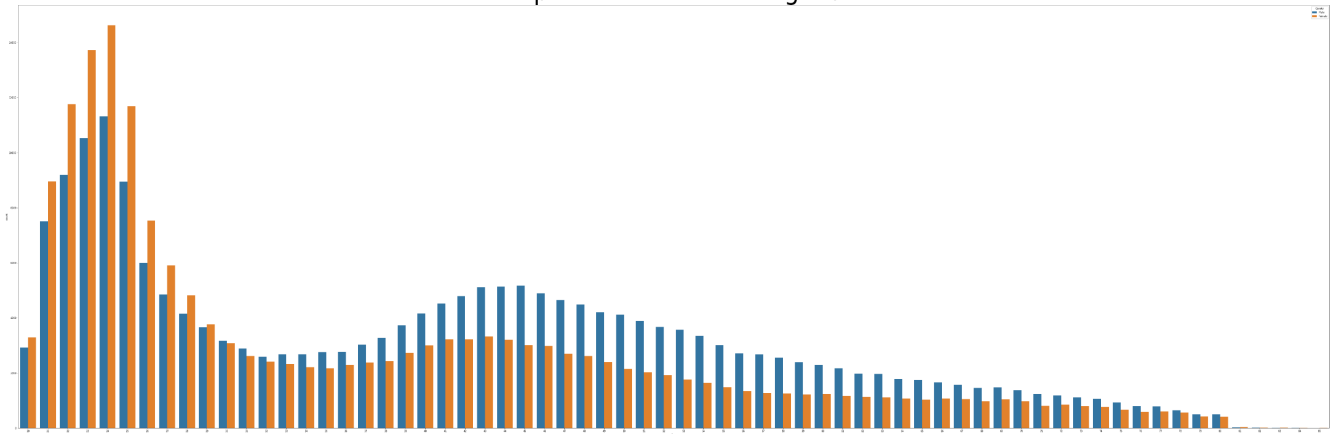
This chart shows the left-skewed distribution that means almost of people in age range from 20 to 28 owns a vehicle in dataset.
This trend goes down by older people.



Exploratory Data Analysis

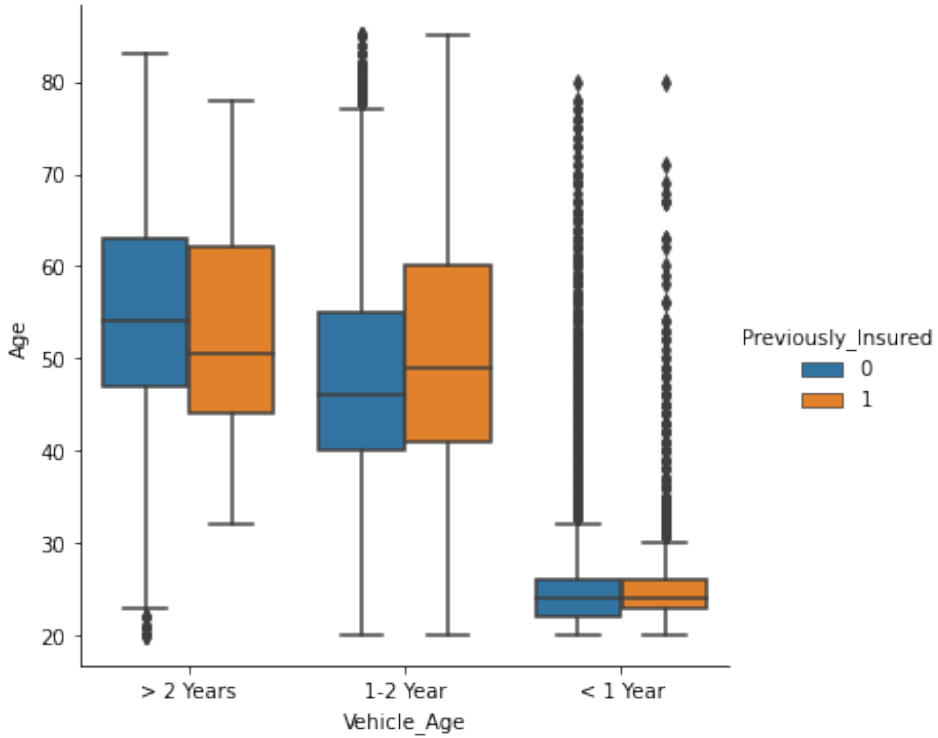
The blue chart that is represented for females has the more amount of people that owned vehicles after 40 years old.
In contrast, young males are preferred owning vehicles to young females.

Comparison the owner of gender



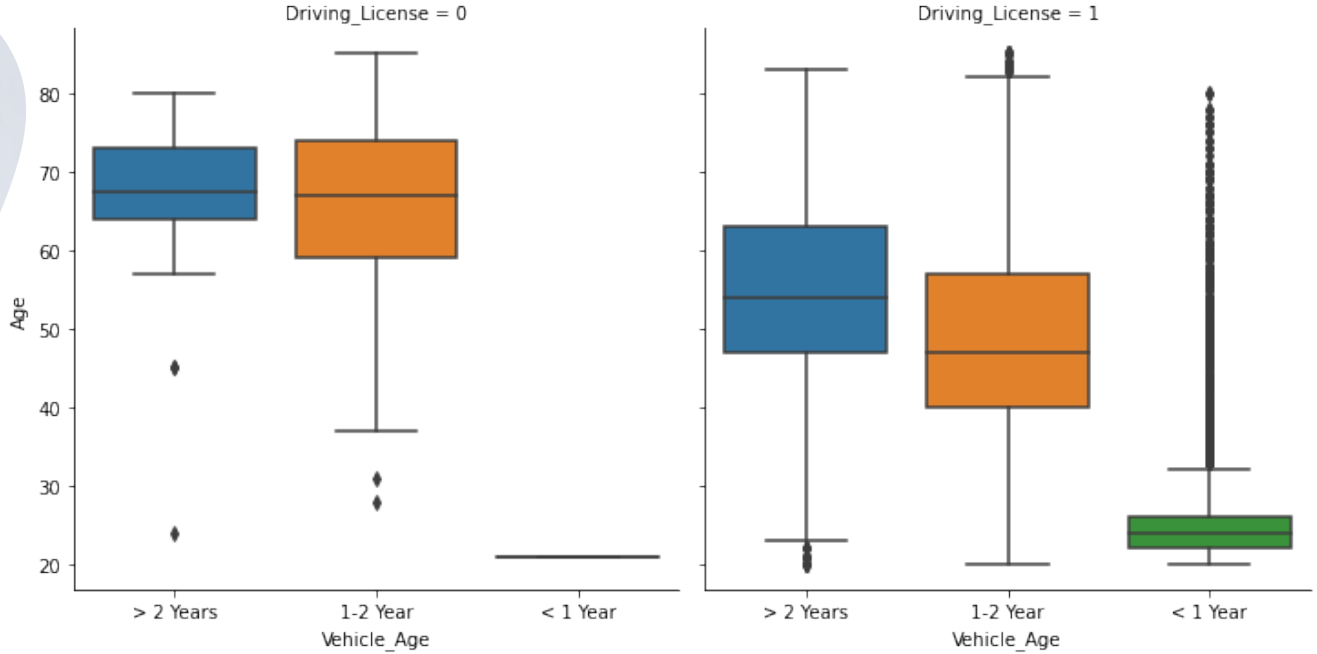
Exploratory Data Analysis

This boxplot chart shows that the newer vehicles that people purchase, the less insurance they buy.
The main group of customers who buys insurance own vehicle for more than one year.



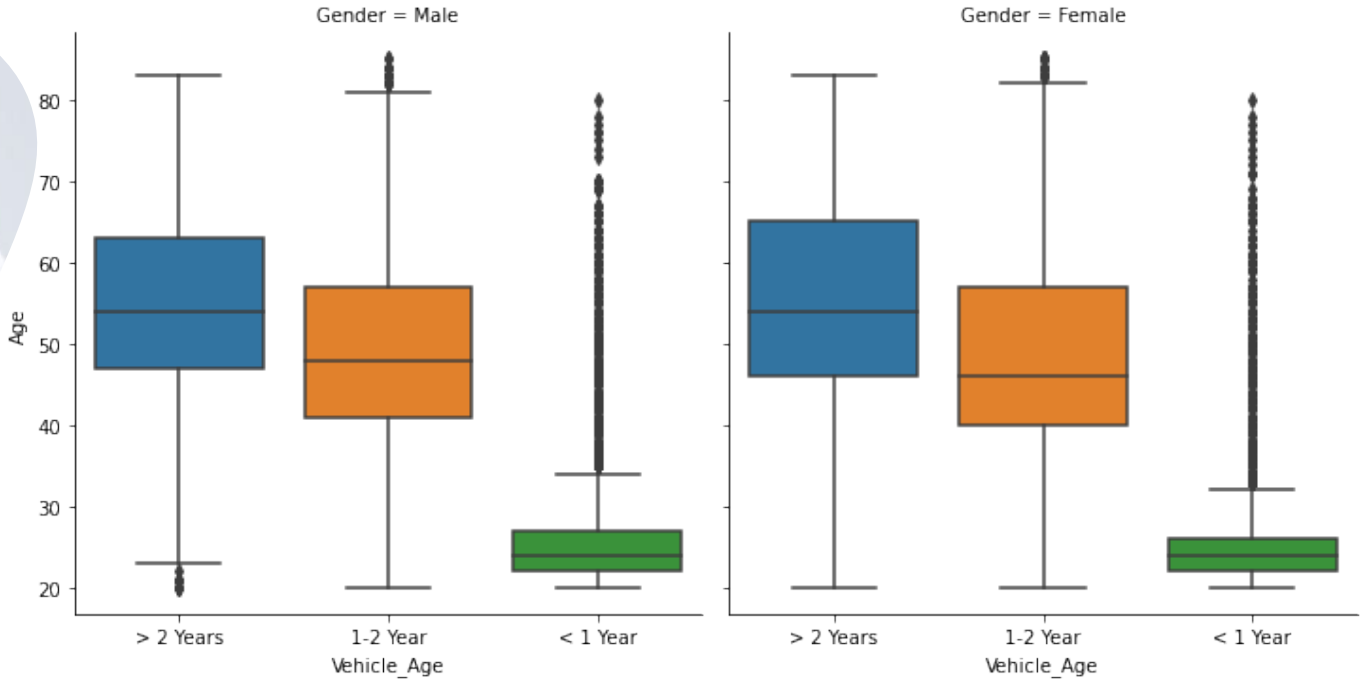
Exploratory Data Analysis

Almost younger people who own new cars have driving license later. This group is the most people who buy new vehicle.
Surprisingly, group of people from 20 to 30 owns new vehicle. However, a little of them have driving licenses.



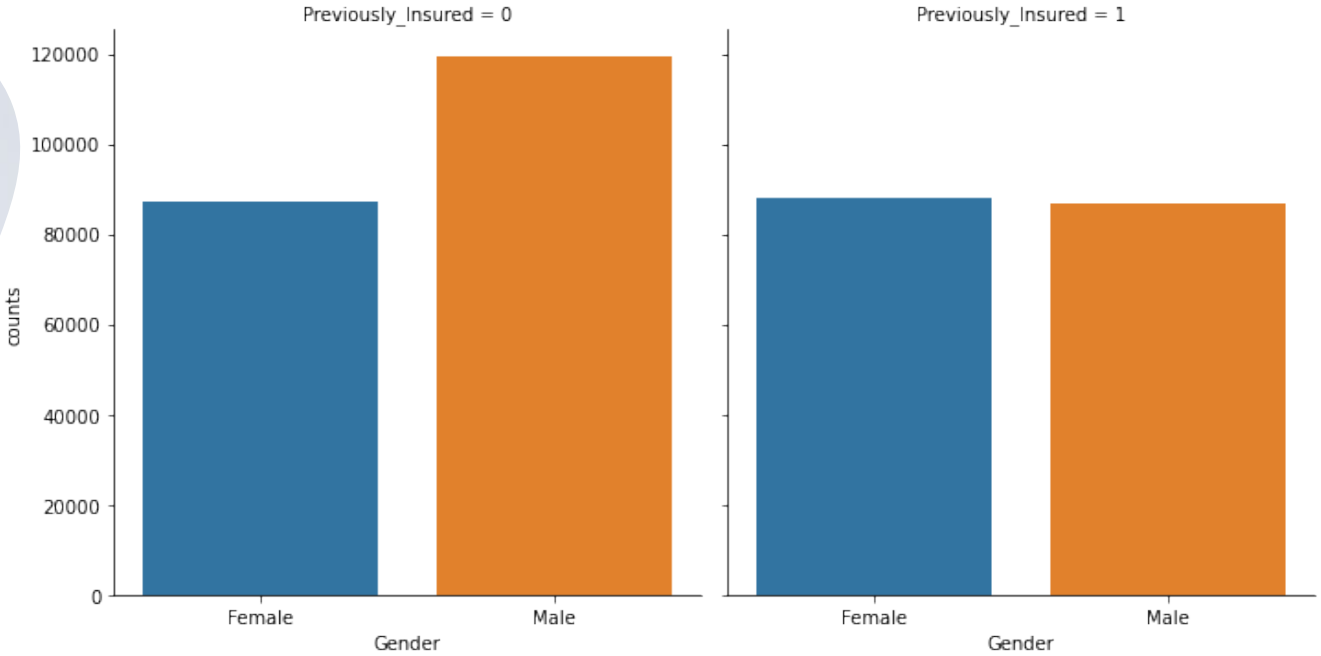
Exploratory Data Analysis

In this chart, we can see that females have the same trend with males when having the normal distribution of owning an old or new vehicle.



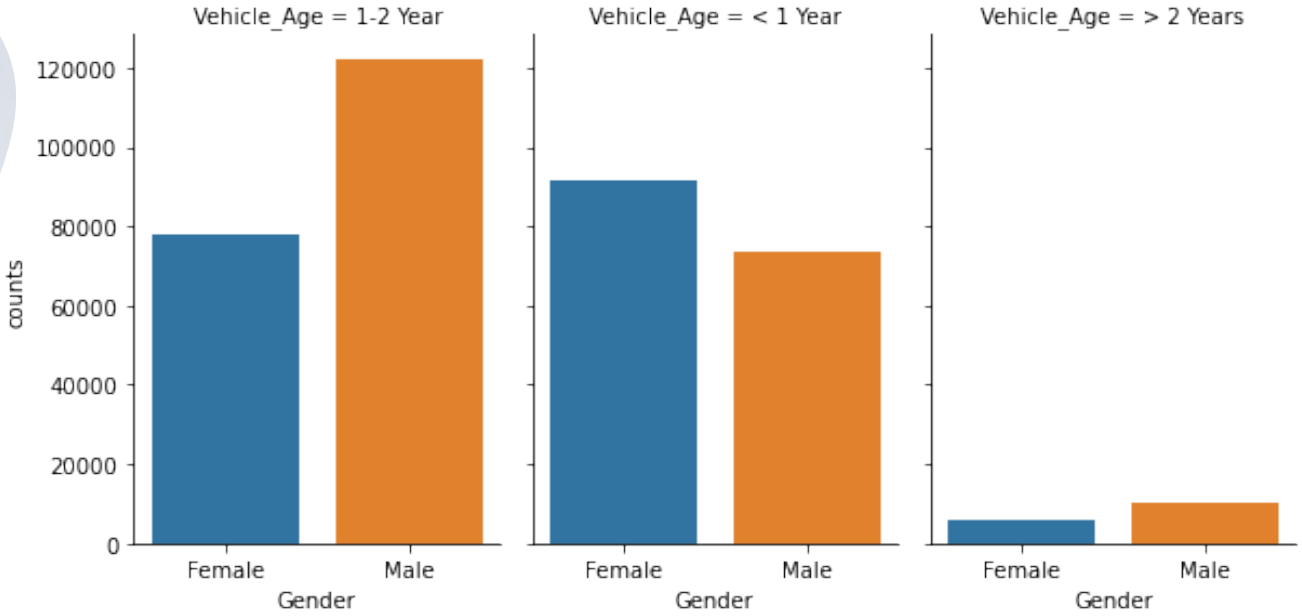
Exploratory Data Analysis

There is no difference of gender in having previously insured.



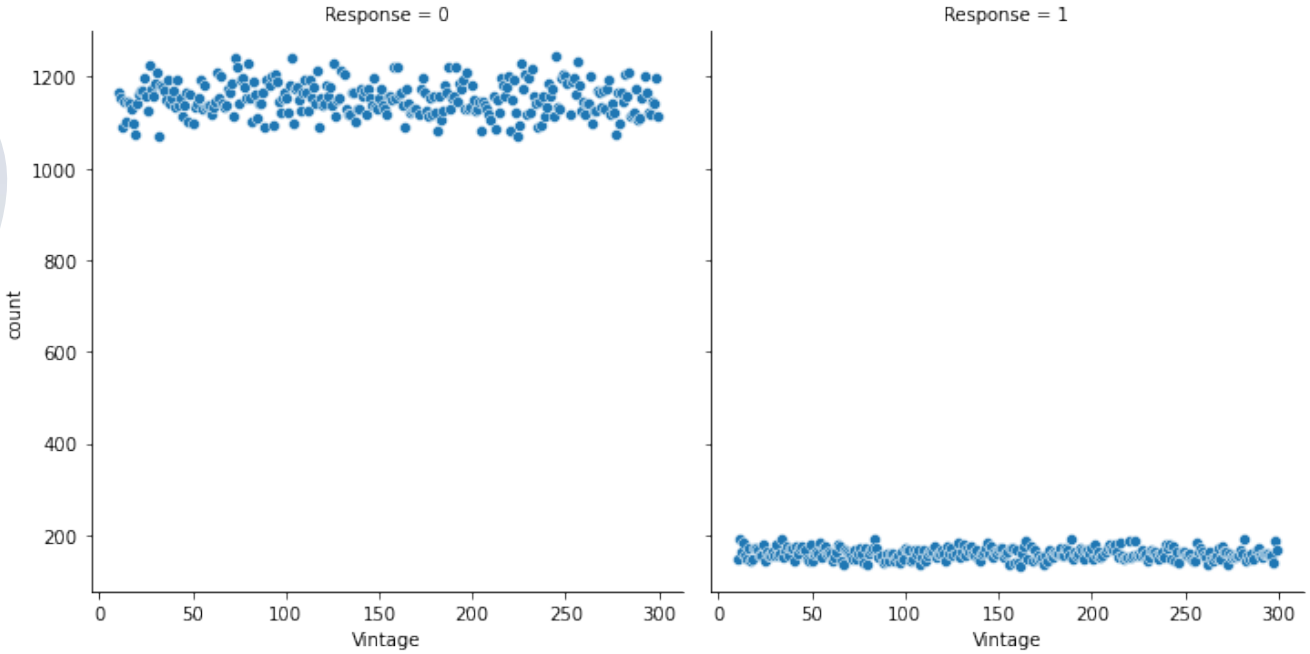
Exploratory Data Analysis

We can see that people in dataset who own vehicle lower 1 year and from 1 to 2 years are the highest group.



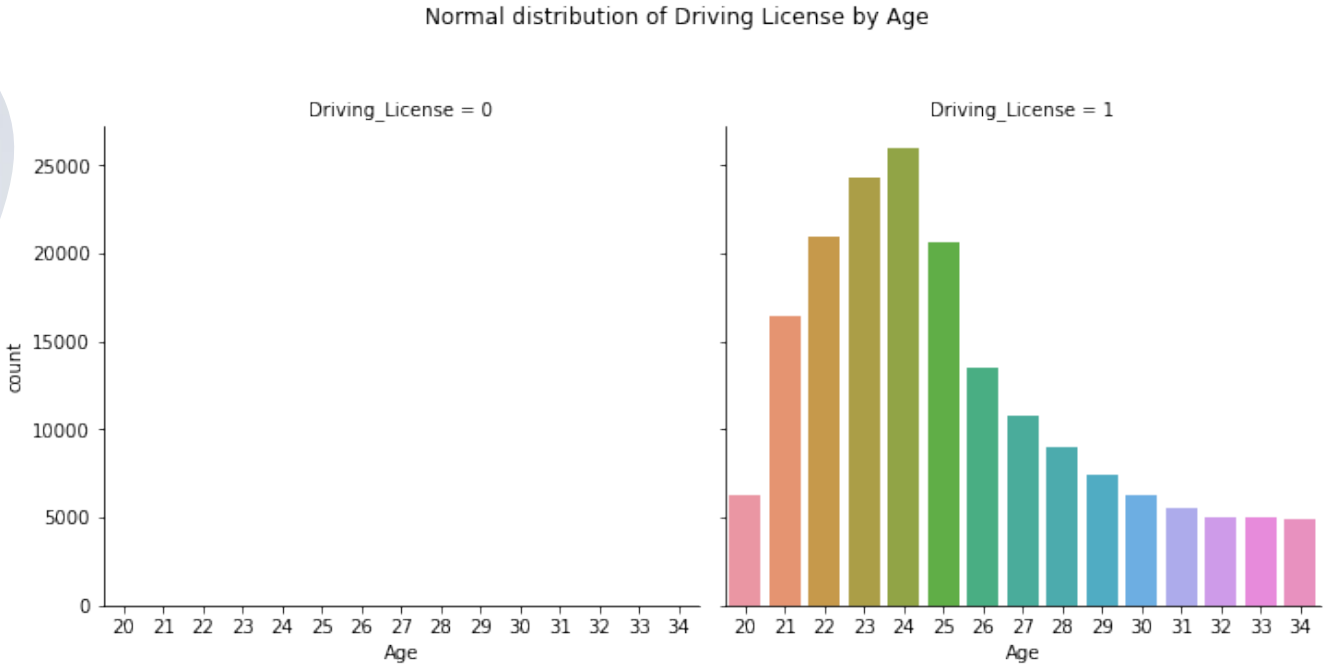
Exploratory Data Analysis

The difference of customers' response in dataset.



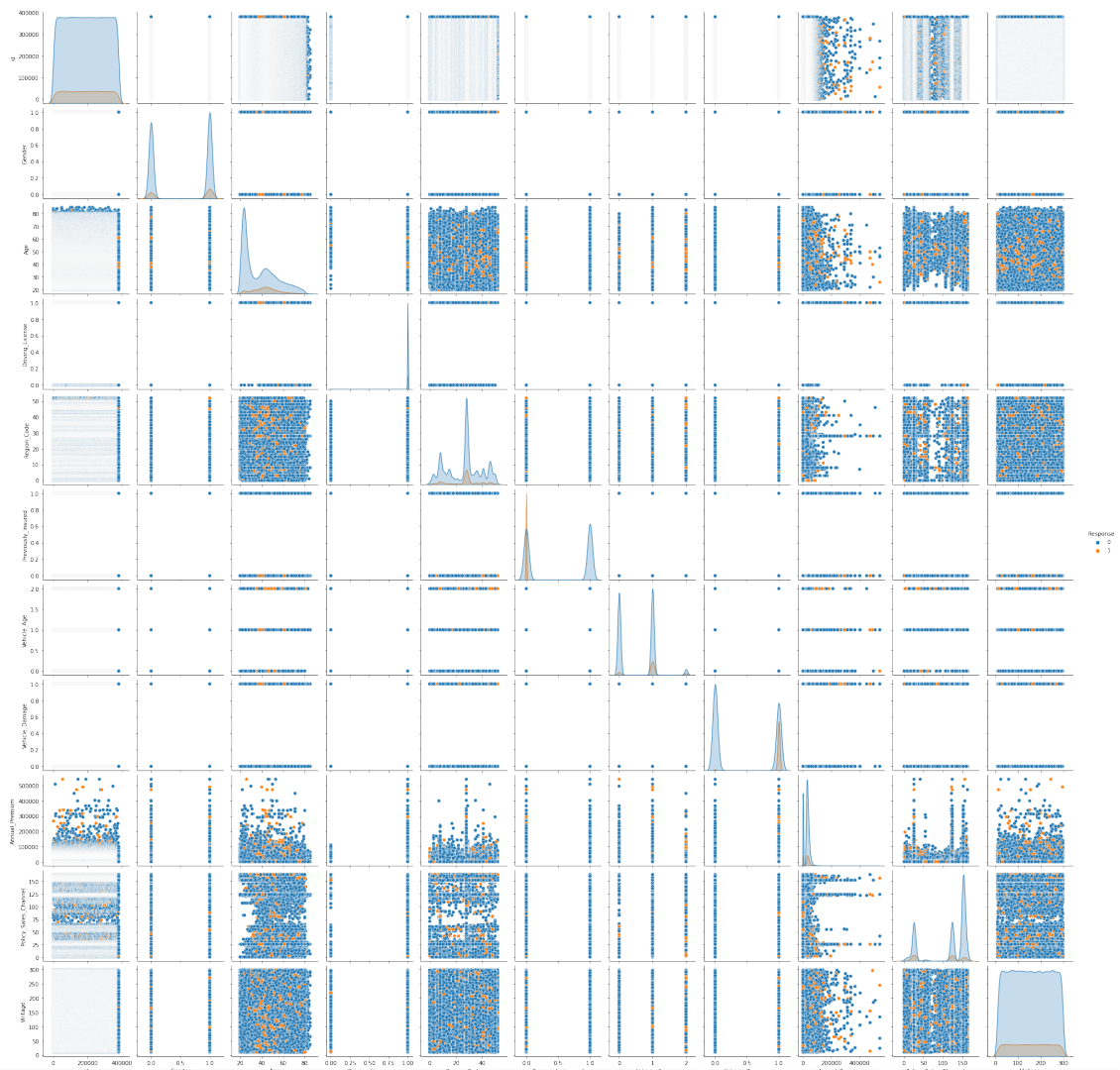
Exploratory Data Analysis

In a group of young people, lower than 35 years old, there are only 4 groups of people who do not have a driving license.
181,872 is the number of this group that has a driving license, and almost all of them are from 21 to 25 years old.



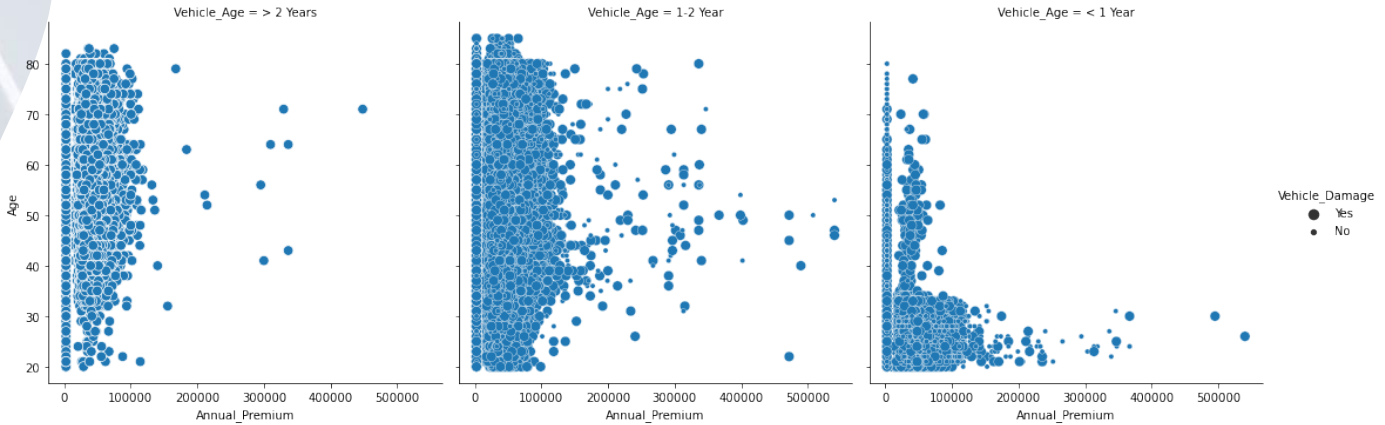
Exploratory Data Analysis

Region does not show a very strong connection to the Annual Premium.



Exploratory Data Analysis

Once again, we see that almost of young people owns new vehicles.
Surprisingly, vehicles which are more than 2 years have less damage than newer.
The same behavior of people owning vehicle, have less than 2 years, are willing to pay high annual premium.



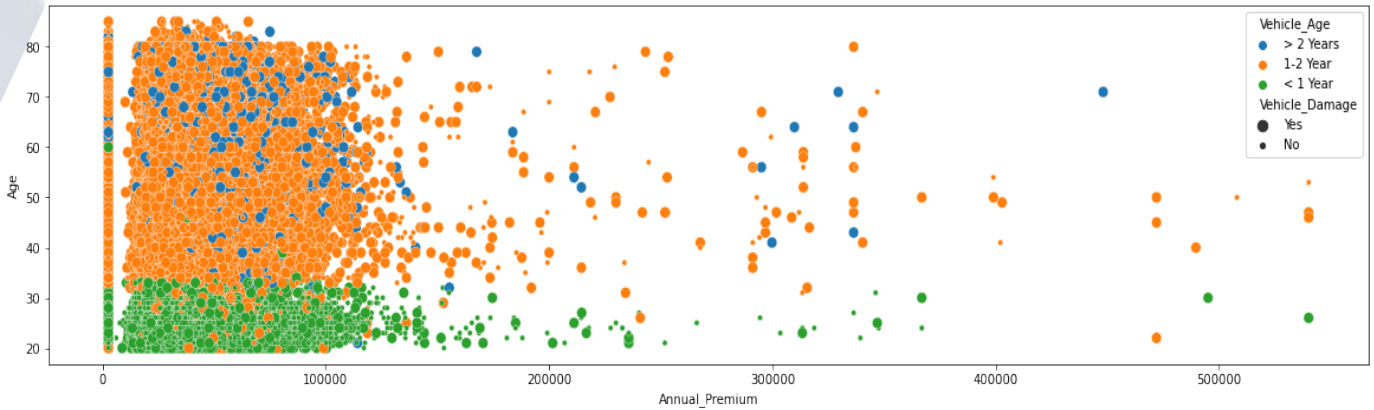
Modelling & Model Evaluation

Modelling

In order to predict the response for the ‘test.csv’ which is a binary variable, so recommendations of the algorithm are Logistic Regression, Random Forest, Decision Tree Classifier, LGBM Classifier, Gaussian Naïve Bayes and Catboost.

In this dataset, I used the SMOTE method to resample the imbalance of data, thus increasing the model performance.

I used the ROC curve for illustrating in a binary classifier system the discrimination threshold created by plotting the true positive rate vs false positive rate.



Modelling

As mentioned above, the ‘train.csv’ and ‘test.csv have 5 categorical columns, especially ‘Gender’, ‘Vehicle_Age’ and ‘Vehicle_Damage’ have string values. So these values have to convert into integer values for the training model.

In this dataset, I used 70% data for training, 30% for tests, and random_state=4 and stratify parameter.

Using GridSearchCV and RandomizedSearchCV to apply best parameters, hyper-parameters tuning, for models.

Because I used 6 models for comparing their value, so I record the result of each model and export them then.

Modelling

The Logistic Regression Model uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, it is estimating the parameters of a logistic model.

The Random Forest Model consists of a large number of individual decision trees that operate as an ensemble. Each individual tree spits out a class prediction and the class with the most votes. It can be used for both Regression and Classification model.

Decision Tree Model is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation.

Modelling

Gaussian Naïve Bayes is a simple but powerful algorithm for predictive modeling under supervised learning algorithms. Naive Bayes has higher accuracy and speed when we have large data points.

The LBBM Classifier has become a de facto algorithm for machine learning competitions when working with tabular data for regression and classification predictive modeling tasks. As such, it owns a share of the blame for the increased popularity and wider adoption of gradient boosting methods in general.

Catboost Model is is an algorithm for gradient boosting on decision trees. It is developed by Yandex researchers and engineers, and is used for search, recommendation systems, personal assistant, self-driving cars, weather prediction and many other tasks at Yandex and in other companies, including CERN, Cloudflare, Careem taxi.

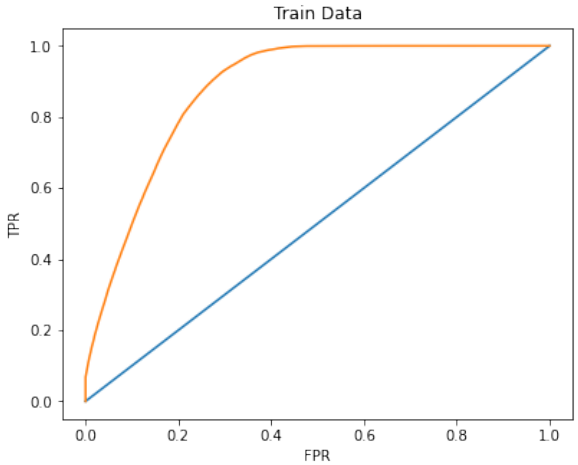
Model Evaluation

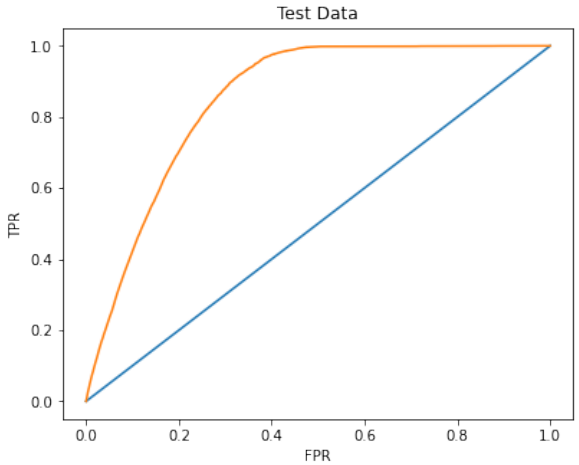
Model Name	Accuracy score test	ROC auc score test	F1 score test	Precision score test	Recall score test
Logistic Regression	0.87743696	0.600270055	0	0	0
Logistic Regression with tuning	0.639141805	0.831257217	0.398536358	0.250425957	0.975451367
RandomForestClassifier	0.69394663	0.851336501	0.42677413	0.276966256	0.929565404
DecisionTreeClassifier	0.700646358	0.848414646	0.429243238	0.280068766	0.918432884
LGBMClassifier	0.725556051	0.848570503	0.434732481	0.290768007	0.861057589
GaussianNB	0.639133059	0.823191752	0.398670806	0.250494505	0.976022265
Catboost	0.746547366	0.851891571	0.445312201	0.304271626	0.830086348

The highest result of these model is Catboost Classifier when it has 0.852, 0.445 and 0.747 at ROC AUC, F1-Score and Accuracy respectively .

Model Evaluation

ROC curve of Catboost model





Luckily, this model seems return good result when True and False Positive Rate are high.

5

Conclusion

Conclusion

People having Vehicles with age > 2 years have to pay more amount of annual premium and that has lead to higher number of people from that category not taking insurance. We need to modify the amount little bit so that people from that category do not skip taking insurance.

People having Vehicle Damage tend to buy insurance as compared to the ones who do not have any damage.

Annual Premium does not depend on how many days people are associated with company. So we can modify the premium policy so that insurance company can attract more customers.

Conclusion

After applying The CatBoost into to the “test.csv” to predict, the output will have 84,731 and 42,306 users who have negative and positive response respectively.

Thanks to the ROC curve of the CatBoost model, we have high accuracy and can use this output. Because this dataset is imbalance, left-skew, so ROC curve is the most important value in applying this prediction model.

Thank you

Any questions?