VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

PROGRAMMING INTERGRATION PROJECT (CO3101)

Report

VEHICLE INSURANCE CROSS SELL PREDICTION

Advisor: Nguyễn Đức Dũng
Students: Trần Quang Thiện - 2053455

HO CHI MINH CITY, OCTOBER 2022

Contents

1 Introduction

There is a client which is an Insurance company that has provided Health Insurance to its customers now they need your help in building a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company.

An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee.

For example, you may pay a premium of 70,000 VND each year for a health insurance cover of Rs. 2,800,000 VND - so that if you fall ill and need to be hospitalised in that year, the insurance provider company will bear the cost of hospitalisation etc. for upto 2,800,000 VND. Now if you are wondering how can company bear such high hospitalisation cost when it charges a premium of only 70,000 VND, that is where the concept of probabilities comes in picture. For example, like you, there may be 100 customers who would be paying a premium of 70,000 VND every year, but only a few of them (say 2-3) would get hospitalised that year and not everyone. This way everyone shares the risk of everyone else.

Just like medical insurance, there is vehicle insurance where every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called ‘sum assured’) to the customer.

2 Goals of project

2.1 Classification problem

Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and revenue.

2.2 Cross sell prediction problem

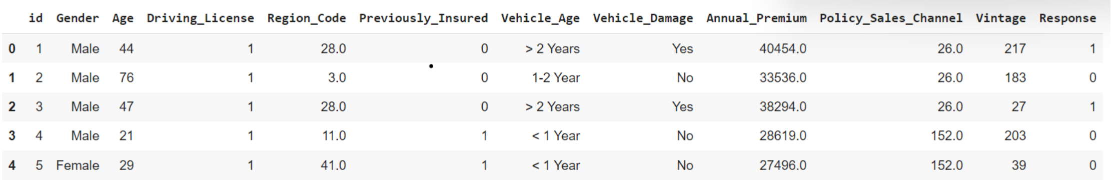
Now, in order to predict and analysis, whether the customer would be interested in Vehicle insurance, you have information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc. Especially, building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimize its business model and revenue.

3 Dataset

Using dataset of a vehicle insurance with columns:

- **id** Unique ID for the customer

- **Gender** Gender of the customer
- **Age** Age of the customer
- **Driving_License**
  0 : Customer does not have Driving License
  1 : Customer already has Driving License
- **Region_Code** Unique code for the region of the customer
- **Previously_Insured**
  0 : Customer doesn't have Vehicle Insurance
  1 : Customer already has Vehicle Insurance
- **Vehicle_Age** Age of the Vehicle
- **Vehicle_Damage**
  0 : Customer didn't get his/her vehicle damaged in the past
  1 : Customer got his/her vehicle damaged in the past
- **Annual_Premium** The amount customer needs to pay as premium in the year
- **PolicySalesChannel** Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
- **Vintage** Number of Days, Customer has been associated with the company
- **Response**
  0 : Customer is not interested
  1 : Customer is interested



	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
0	1	Male	44	1	28.0	0	> 2 Years	Yes	40454.0	26.0	217	1
1	2	Male	76	1	3.0	0	1-2 Year	No	33536.0	26.0	183	0
2	3	Male	47	1	28.0	0	> 2 Years	Yes	38294.0	26.0	27	1
3	4	Male	21	1	11.0	1	< 1 Year	No	28619.0	152.0	203	0
4	5	Female	29	1	41.0	1	< 1 Year	No	27496.0	152.0	39	0

Hình 1: Dataset

4 Exist methods to solve the problem

4.1 Imbalanced data and classes handling

4.2 Exploratory Data Analysis

4.3 Comprehensive ML Models

4.4 Analyze Machine Learning Models using SHAP

5 Tools and domain required

5.1 Data handling

- Handle many data from difference kind of data XML, RDF, JSON (CSV) by some applications.

- Processing data by using MS excel, MS power BI, Python, R, . . .

5.2 Build model to predict

- Azure Machine Learning: Azura ML from Microsoft

- Python for Probability, Statistics, and Machine Learning

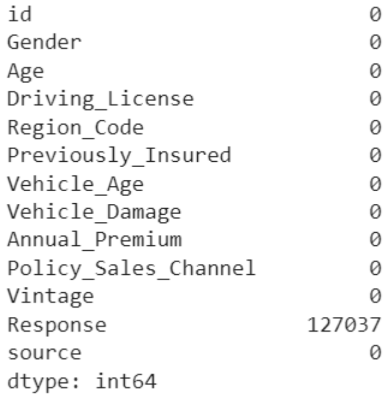
- Explore NumPy for Numerical Data, Pandas for Data Analysis, IPython, Scikit-Learn and Tensorflow

6 Implementation

6.1 Exploratory Data Analysis (EDA)

6.1.1 Check for missing values

- Detect and count missing values(NaN) for each column by default.



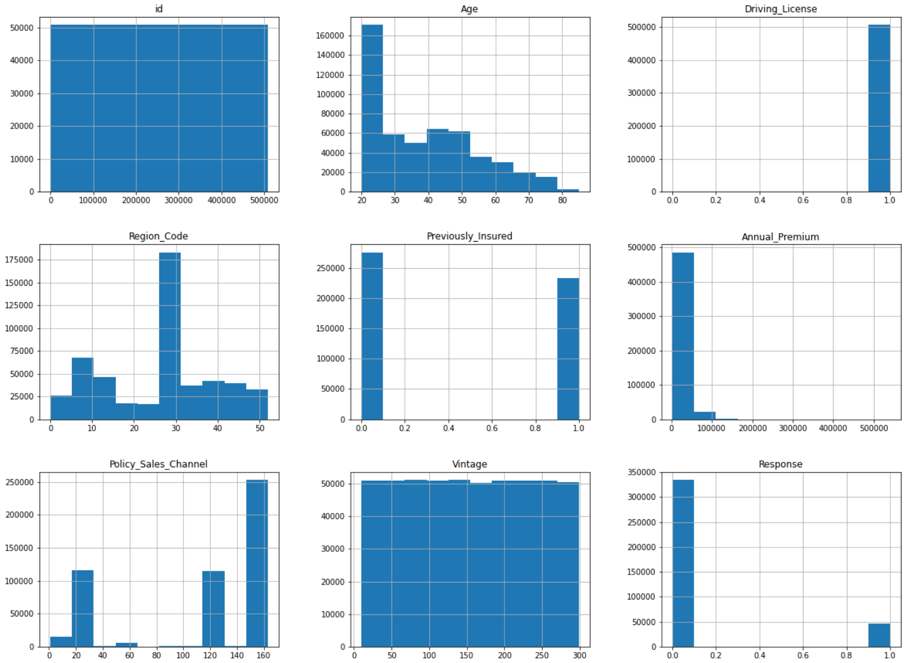
```
id                         0
Gender                     0
Age                        0
Driving_License            0
Region_Code                0
Previously_Insured         0
Vehicle_Age                0
Vehicle_Damage             0
Annual_Premium             0
Policy_Sales_Channel       0
Vintage                    0
Response              127037
source                     0
dtype: int64
```

Hình 2: The sum of null elements for each column

- Divide dataset into 2 part: num_features(having columns with number object) and num_cat(having columns without number object).
  - num_features = ['id','Age','Driving_License','Region_Code','Previously_Insured', 'Annual_Premium','Policy_Sales_Channel','Vintage','Response']
  - num_cat = ['Gender', 'Vehicle_Age', 'Vehicle_Damage', 'source']

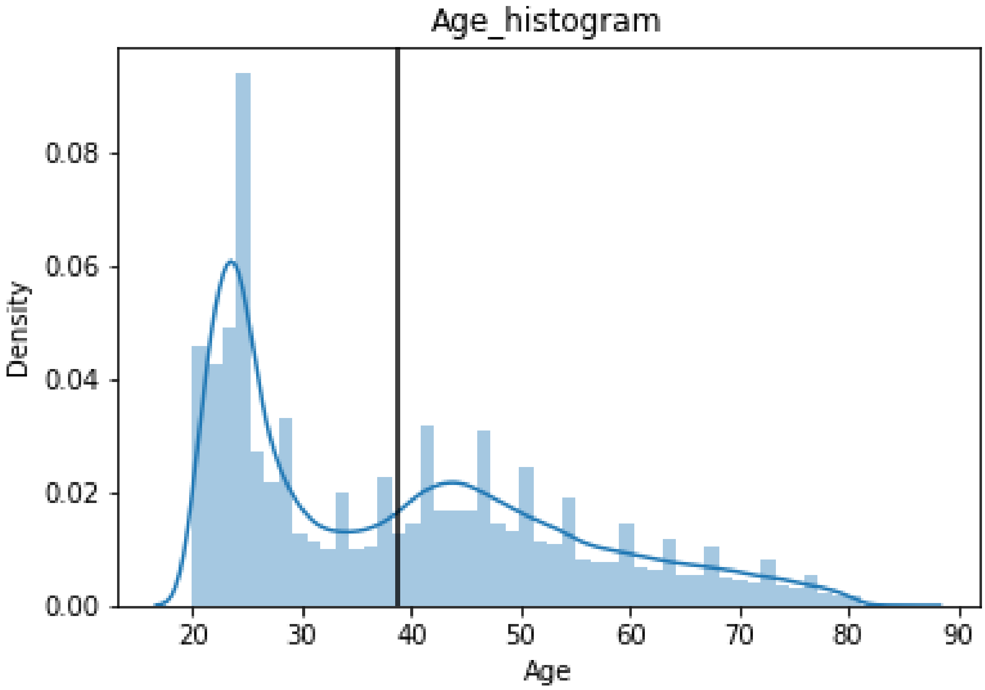
6.1.2 Numerical Features Analysis

- From the data of the num_features, draw the histograms corresponding to those columns.



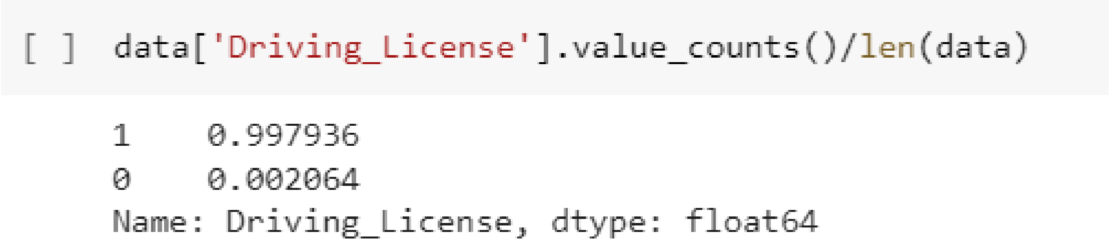
Hình 3: Histograms of the num_features

- From the above plot we can see that :
  - Age column is Right-Skewed (also known as "positively skewed" distribution, most data falls to the right, or positive side, of the graph's peak).



Hình 4: Age histogram

- Diriving_license column has unbalanced ratio between the categories.

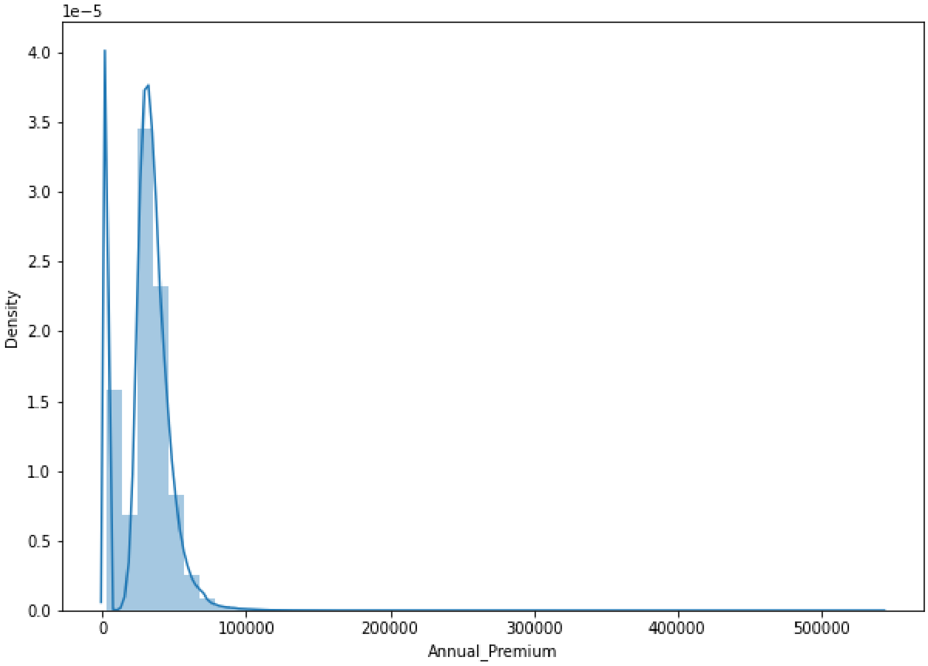


```
[ ]  data['Driving_License'].value_counts()/len(data)

     1    0.997936
     0    0.002064
     Name: Driving_License, dtype: float64
```

Hình 5: Ratio of Diriving_license column

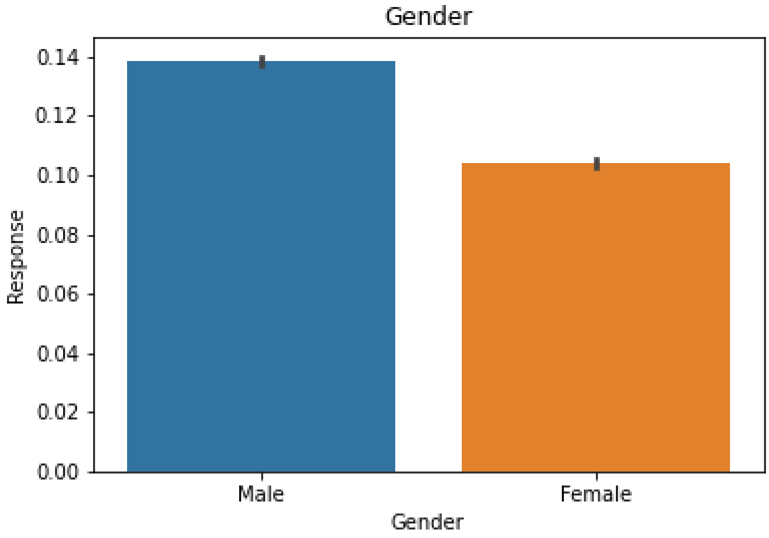
- We can see the unequal distribution in Driving_License. So we shall drop this column.
- We can see that customers generally opt for Premium which is not too high.



Hình 6: Annual Premium diagram

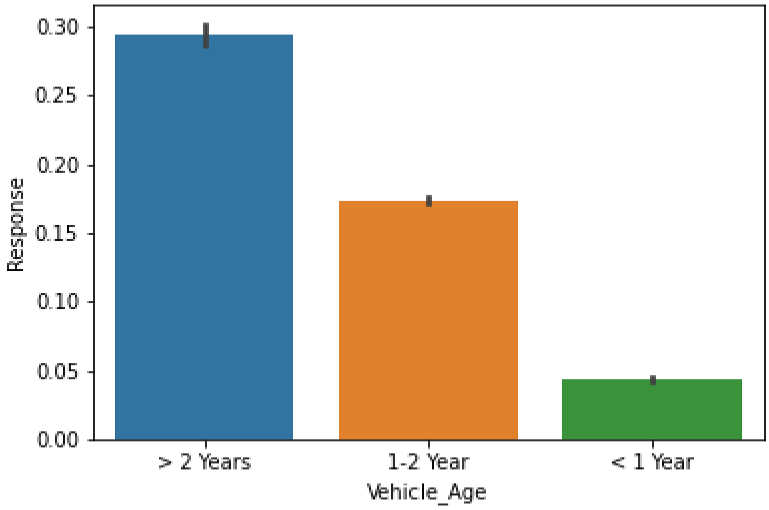
6.2 Categorical Features Analysis

– We can see that 'Males' are more likely to buy insurance.



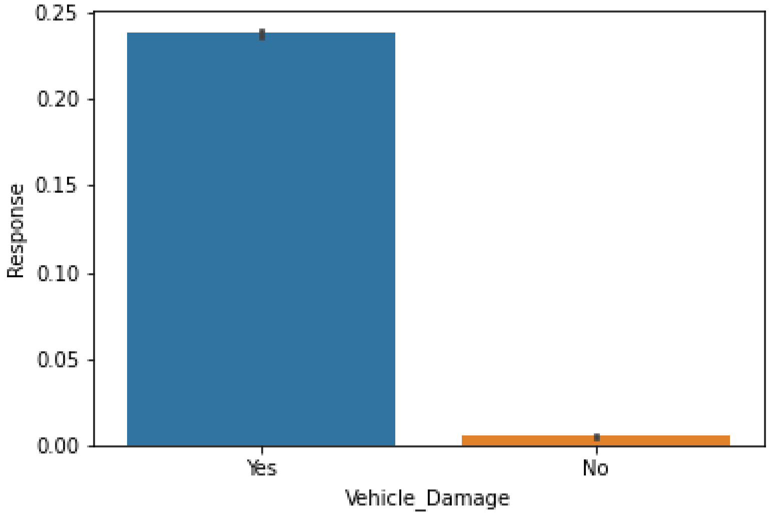
Hình 7: Gender diagram

- From the figure.8, it is clear that the more the age of vehicle the better making the vehicle insurance cheaper

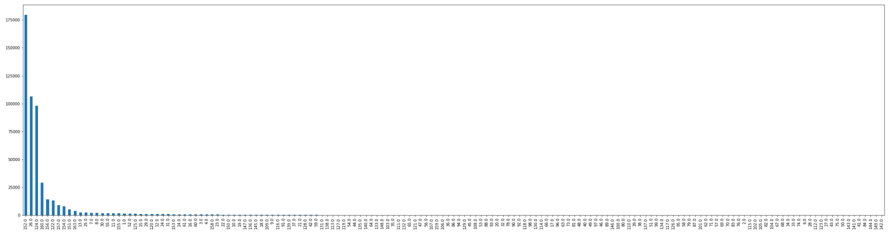


Hình 8: Vehicle age diagram

- From the figure.9, We can see that having ’Vehile_Damage’ are more likely to buy insurance.



Hình 9: Vehicle age diagram



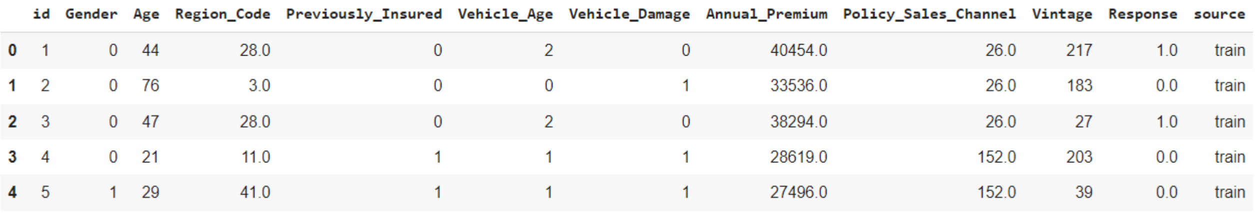
Hình 10: Vehicle age diagram

- Policy_Sales_Channel no. 152 have highest number of customers.
- Policy_Sales_Channel no. [152,26,124,160,156,122,157,154,151,163] have most of the customers.

6.2.1 Data Cleaning

- In Gender Feature:
  - 'Male' : 0
  - 'Female' : 1
- In vehicle_age_map Feature:
  - '1-2 Year' : 0
  - '< 1 Year' : 1
  - '> 2 Years' : 2
- In Vehicle_Damage_map Feature:

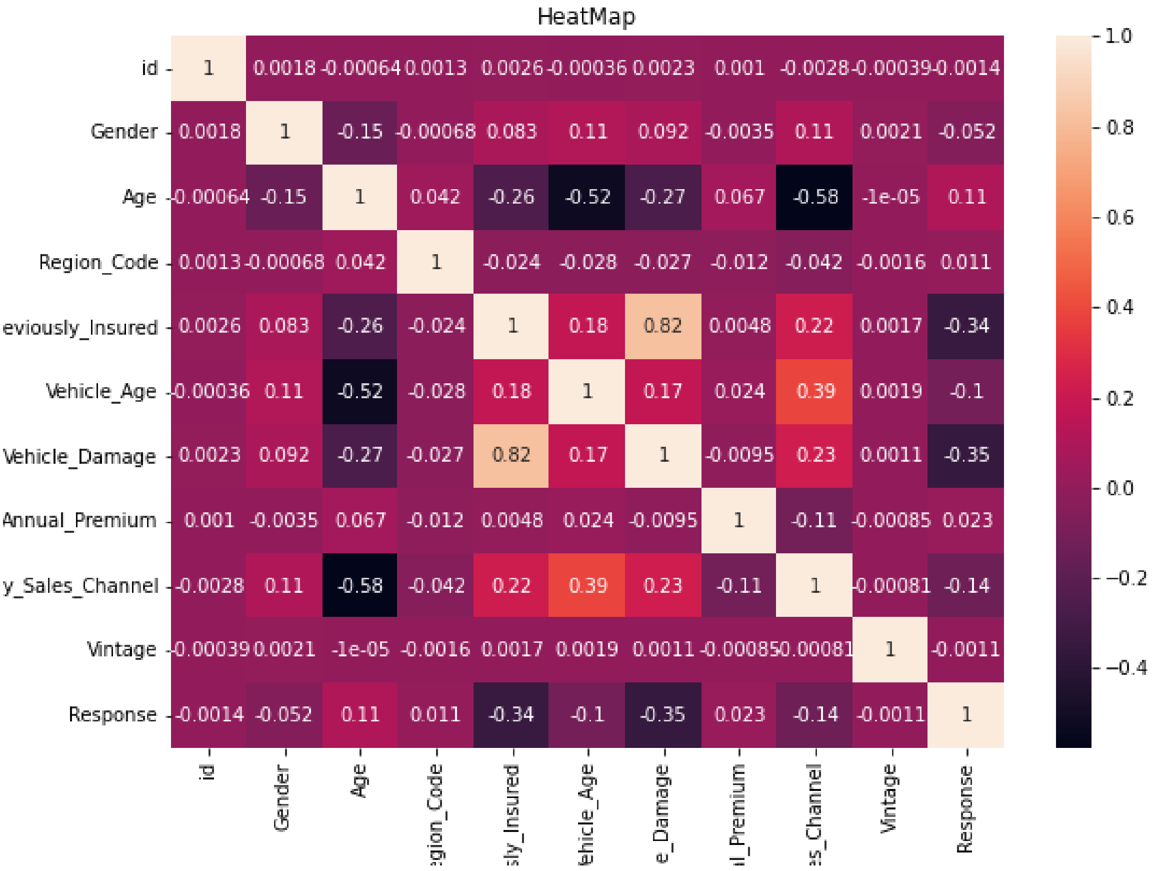
- ∗ 'Yes' : 0
- ∗ 'No' : 1



	id	Gender	Age	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response	source
0	1	0	44	28.0	0	2	0	40454.0	26.0	217	1.0	train
1	2	0	76	3.0	0	0	1	33536.0	26.0	183	0.0	train
2	3	0	47	28.0	0	2	0	38294.0	26.0	27	1.0	train
3	4	0	21	11.0	1	1	1	28619.0	152.0	203	0.0	train
4	5	1	29	41.0	1	1	1	27496.0	152.0	39	0.0	train

Hình 11: Cleaned dataset

- From above dataset, we can draw Corresponding map between all features.



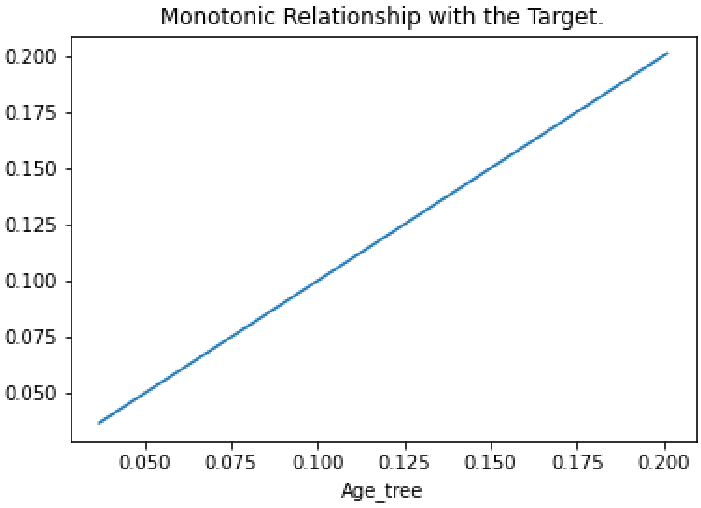
	id	Gender	Age	:gion_Code	sly_Insured	ehicle_Age	e_Damage	ıl_Premium	:s_Channel	Vintage	Response
id	1	0.0018	-0.00064	0.0013	0.0026	-0.00036	0.0023	0.001	-0.0028	-0.00039	-0.0014
Gender	0.0018	1	-0.15	-0.00068	0.083	0.11	0.092	-0.0035	0.11	0.0021	-0.052
Age	-0.00064	-0.15	1	0.042	-0.26	-0.52	-0.27	0.067	-0.58	-1e-05	0.11
Region_Code	0.0013	-0.00068	0.042	1	-0.024	-0.028	-0.027	-0.012	-0.042	-0.0016	0.011
eviously_Insured	0.0026	0.083	-0.26	-0.024	1	0.18	0.82	0.0048	0.22	0.0017	-0.34
Vehicle_Age	-0.00036	0.11	-0.52	-0.028	0.18	1	0.17	0.024	0.39	0.0019	-0.1
Vehicle_Damage	0.0023	0.092	-0.27	-0.027	0.82	0.17	1	-0.0095	0.23	0.0011	-0.35
Annual_Premium	0.001	-0.0035	0.067	-0.012	0.0048	0.024	-0.0095	1	-0.11	-0.00085	0.023
y_Sales_Channel	-0.0028	0.11	-0.58	-0.042	0.22	0.39	0.23	-0.11	1	-0.00081	-0.14
Vintage	-0.00039	0.0021	-1e-05	-0.0016	0.0017	0.0019	0.0011	-0.00085	-0.00081	1	-0.0011
Response	-0.0014	-0.052	0.11	0.011	-0.34	-0.1	-0.35	0.023	-0.14	-0.0011	1

Hình 12: Corresponding map

- We can see that Previously_insured & Vehicle_Damage are both Positively correlated. We will delete column which has strong relation with the Target column.
- This time we will delete column: Vehicle_Damage as it is more negatively correlated to the Target class.

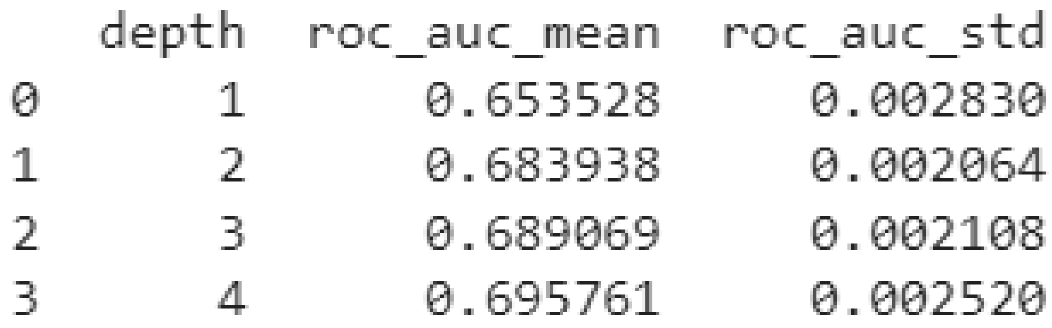
6.3 Decision Tree for Age column

- Before building Decision Tree, we must check the unique values, the no of customer per bin. And from there we can draw Monotonic Relationship between age tree and the Target.



Hình 13: Monotonic Relationship with the Target

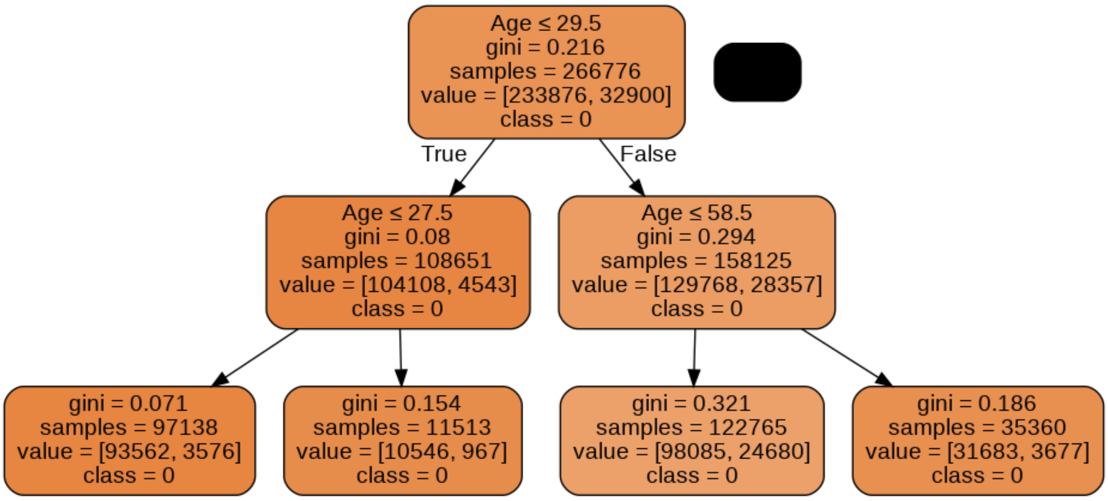
- We can see that the new column: Age_tree is a good predictor of the Target. We can use the Predicted_probability to create the Bins.
- The next step we must do is choosing the maximum dept of Decision Tree. After using Hyper-parameter Tuning to check the best depth, we can see from the Figure.14 that We can see that with depth=2 is a better choice to avoid overfitting. We replaced the continuous column with the bins.



```
   depth  roc_auc_mean  roc_auc_std
0      1      0.653528     0.002830
1      2      0.683938     0.002064
2      3      0.689069     0.002108
3      4      0.695761     0.002520
```

Hình 14: Result of Hyper-parameter Tuning to check the best depth Age's Decision Tree

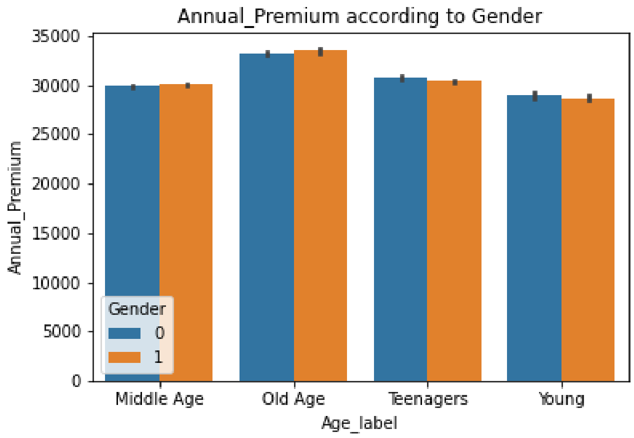
- And I build Age's Decision Tree with maximum depth = 2.



[redacted]

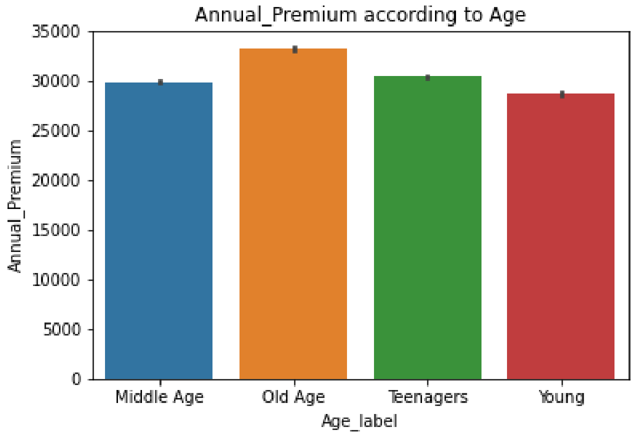
Hình 15: Age Binary Decision Tree

- From above Age's Decision Tree, we can divide Age Features into 4 group.
  - People with age smaller than 27: Teenagers
  - People with age from 27 to 29: Young
  - People with age from 29 to 58: Middle Age
  - People with age from 58 to 85: Old Age
- After that, we shall check whether the relationship between Age group with Gender and Annual Premium.



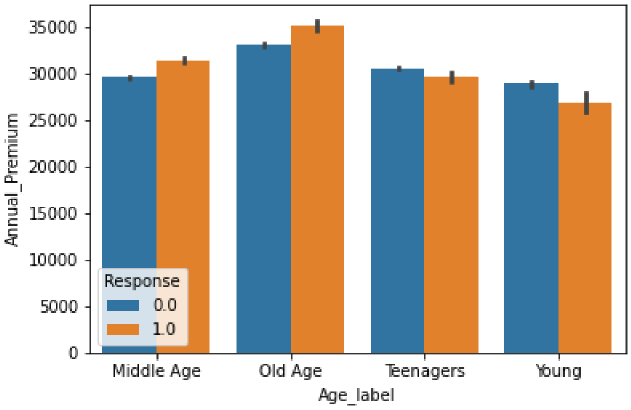
Hình 16: Gender according to Age

- There is not much to see from the above plot.



Hình 17: Annual Premium according to Age

- From the above Plot we can see that Annual Premium is directly dependent on the Age of the Customer. The higher the age higher the Annual Premium.
- And relationship between age group, Annual Premium and response(Figure.18)

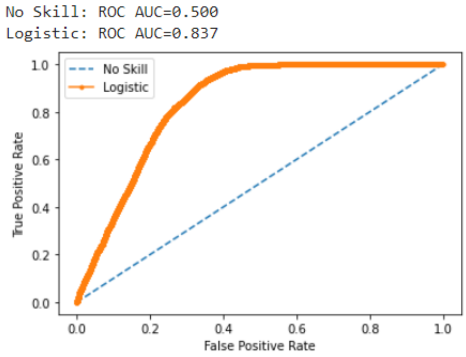


Hình 18: Annual Premium, response according to Age

- We can see that:
  - Age group 20-27 usually do not take Insurance as they are just starting with their lives and may not have money to pay the Premium.
  - Age group from 29-85 opt for Insurance as they have driving License and resources to pay the Premium

6.4 Build Model Cross Sell Prediction

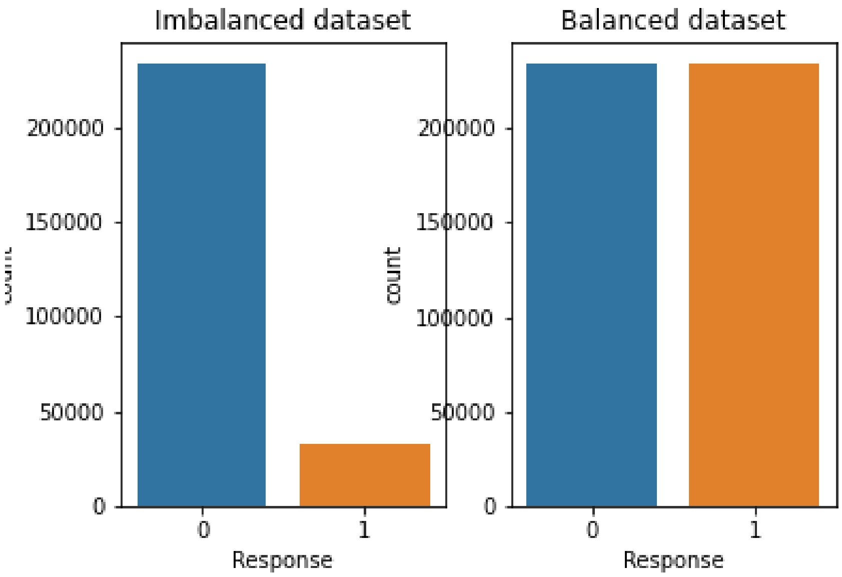
- In this stage we will create our model using Machine Learning libraries.
- I have also used Cross Validation to select across different models.
- First, I solve the imbalanced data problem, in particular i used Logistic Regression to get ROC-AUC score of Model.



```
No Skill: ROC AUC=0.500
Logistic: ROC AUC=0.837
```

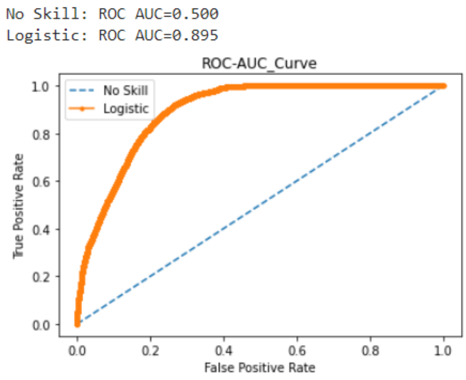
Hình 19: ROC-AUC_Curve with unbalanced data

- We can see that with using simplest Classification Algorithm we are able to get ROC-AUC score as 0.83. And its is a case of Imbalanced dataset. So we will apply some techniques to balance the dataset. In this case, I use over_sampling to solve this problem. And the result is generalized Figure.20.



Hình 20

- After solving the imbalanced data problem, we have ROC-AUC_Curve with balanced data. (Figure.21)



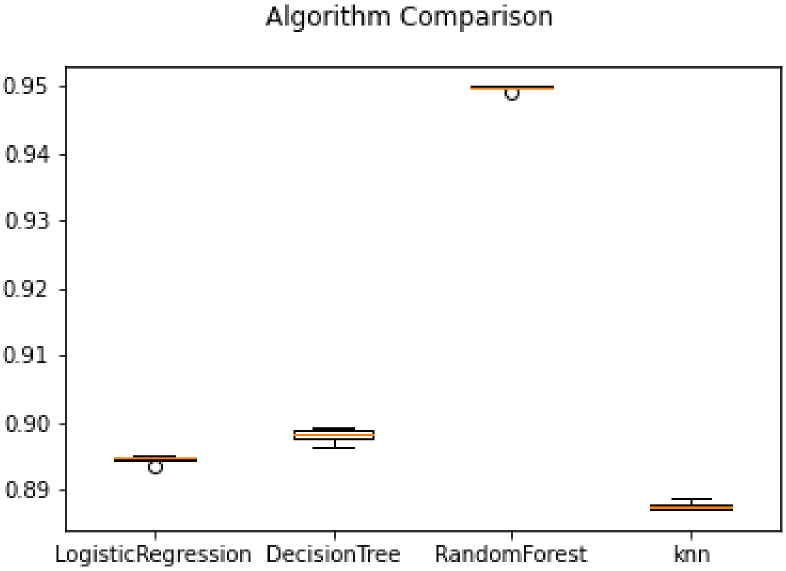
```
No Skill: ROC AUC=0.500
Logistic: ROC AUC=0.895
```

Hình 21

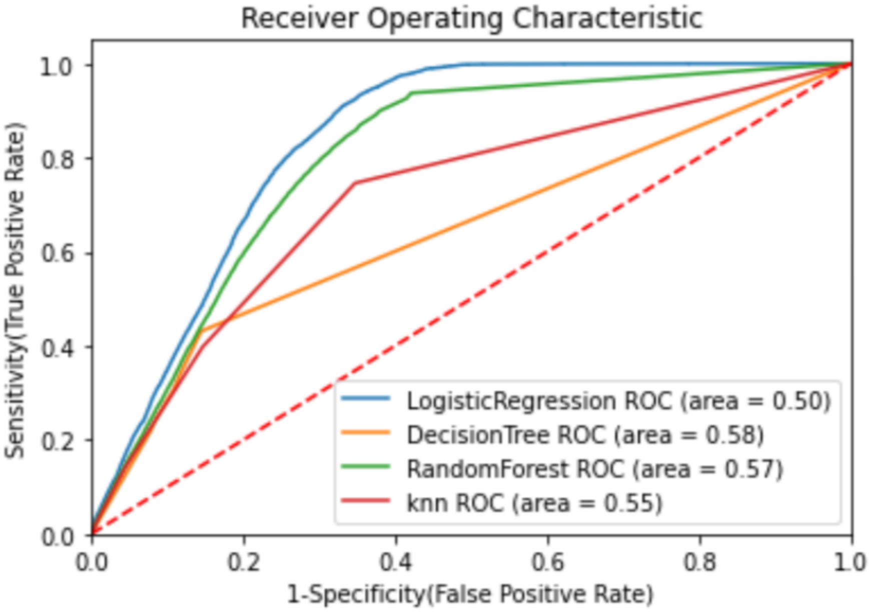
- We can clearly see that after Balancing the ratio of target variable. The ROC-AUC score has increased to from 0.83 to 0.89.

- We can clearly see that after Balancing the ratio of target variable. The ROC-AUC score has increased to from 0.83 to 0.89.

- Finally, choosing Algorithm to build model.

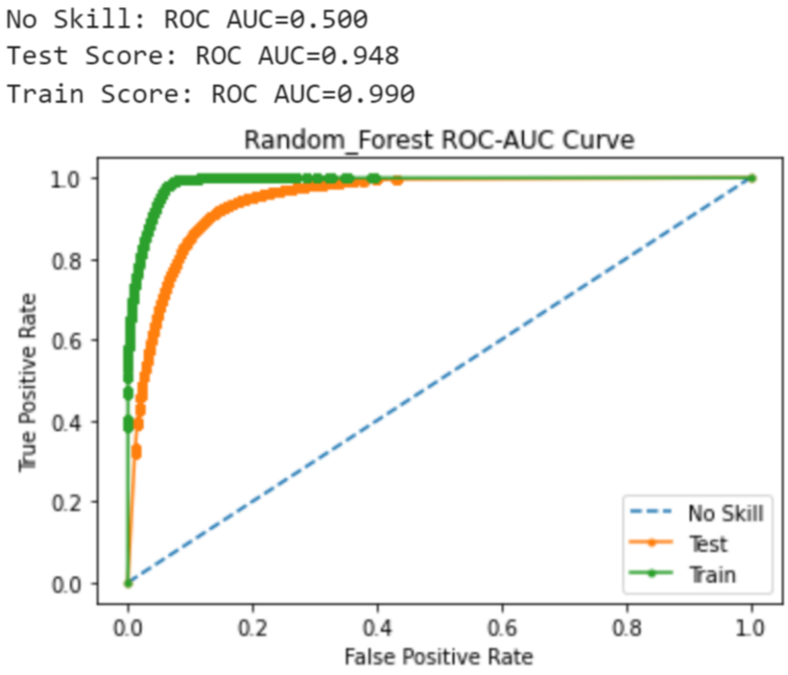


Hình 22



Hình 23

- We can see that RandomForest has performed the best so we will go with it.
- And the result ofRandom_Forest ROC-AUC Curve. (Figure.24)



```
No Skill: ROC AUC=0.500
Test Score: ROC AUC=0.948
Train Score: ROC AUC=0.990
```

Hình 24: Random Forest ROC-AUC Curve

7 References