Capstone Project-3

HEALTH INSURANCE CROSS SELL PREDICTION

Kartika Sharma

Is Insurance a good Idea?

- Problem Statement
- EDA
- Feature Selection
- Preparing dataset for modelling
- SMOTE
- Model Fitting
- Evaluation
- Hyperparameter tuning

Rest Insured

- 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.

- Now, in order to predict, 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.

Data Summary

- id : Unique ID for the customer
- Gender : Gender of the customer
- Age : Age of the customer
- Driving_License 0 : Customer does not have DL, 1 : Customer already has DL
- Region_Code : Unique code for the region of the customer
- Previously_Insured : 1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
- Vehicle_Age : Age of the Vehicle

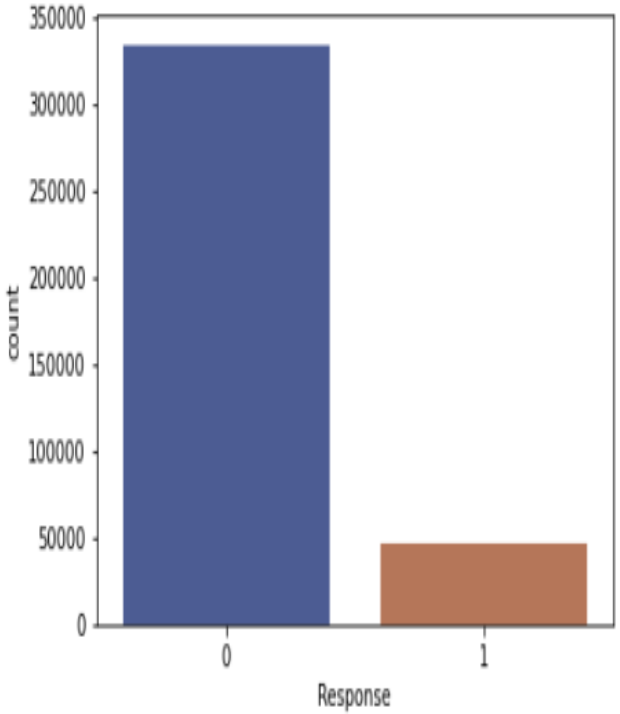
Data Summary

- Vehicle_Damage :1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get 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 : 1 : Customer is interested, 0 : Customer is not interested

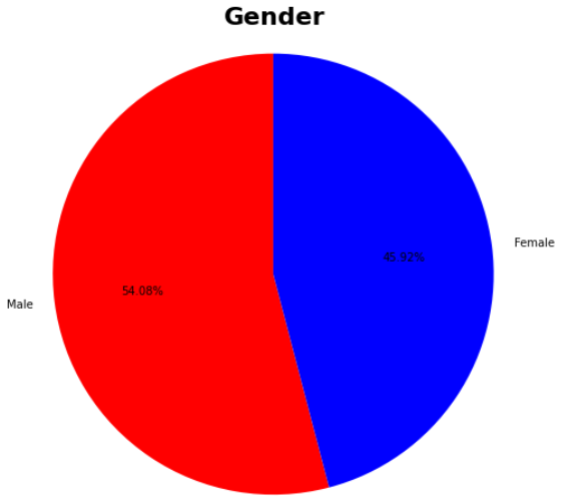
EDA (Target)

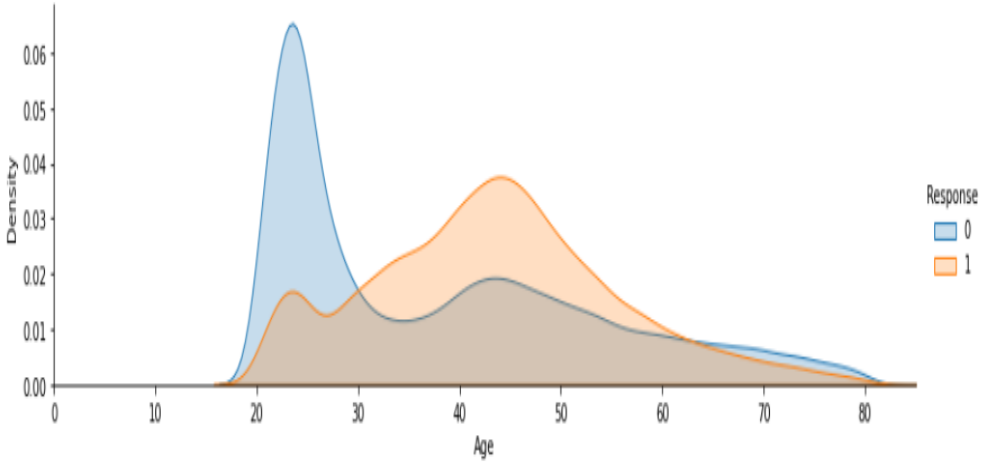
We analysed that response 1 values are too low to give desired results
i.e is has a high possibility
of giving false positives

Countplot shows that response 1 is less than 5k and other one around 3.5 lakhs

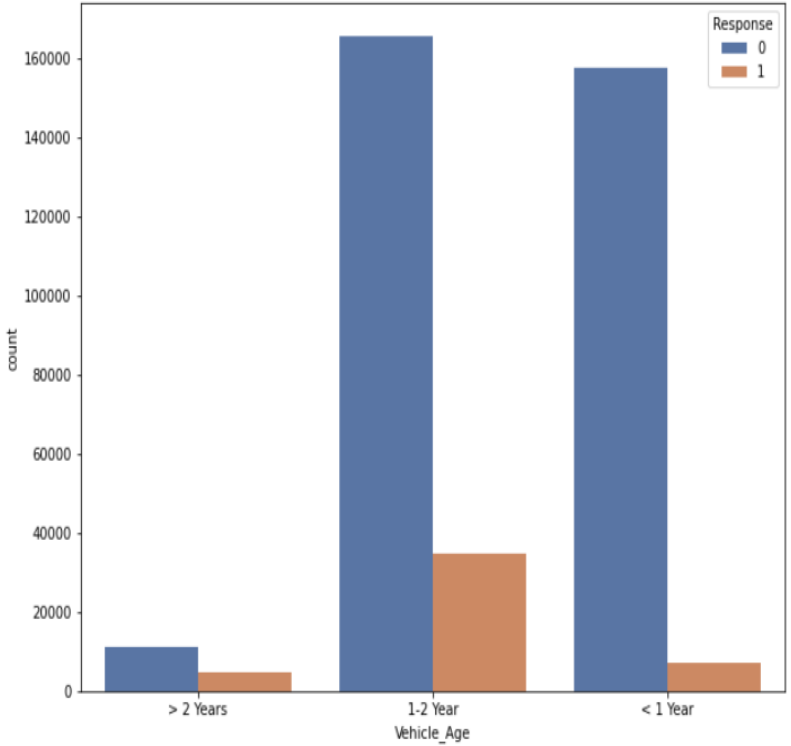


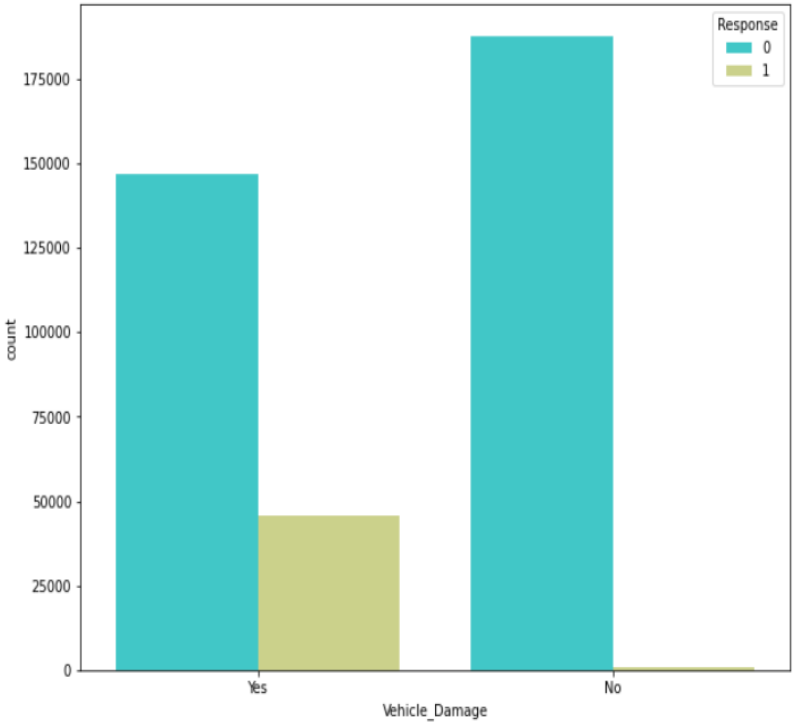
EDA


Gender
Female
Male
45.92%
54.08%

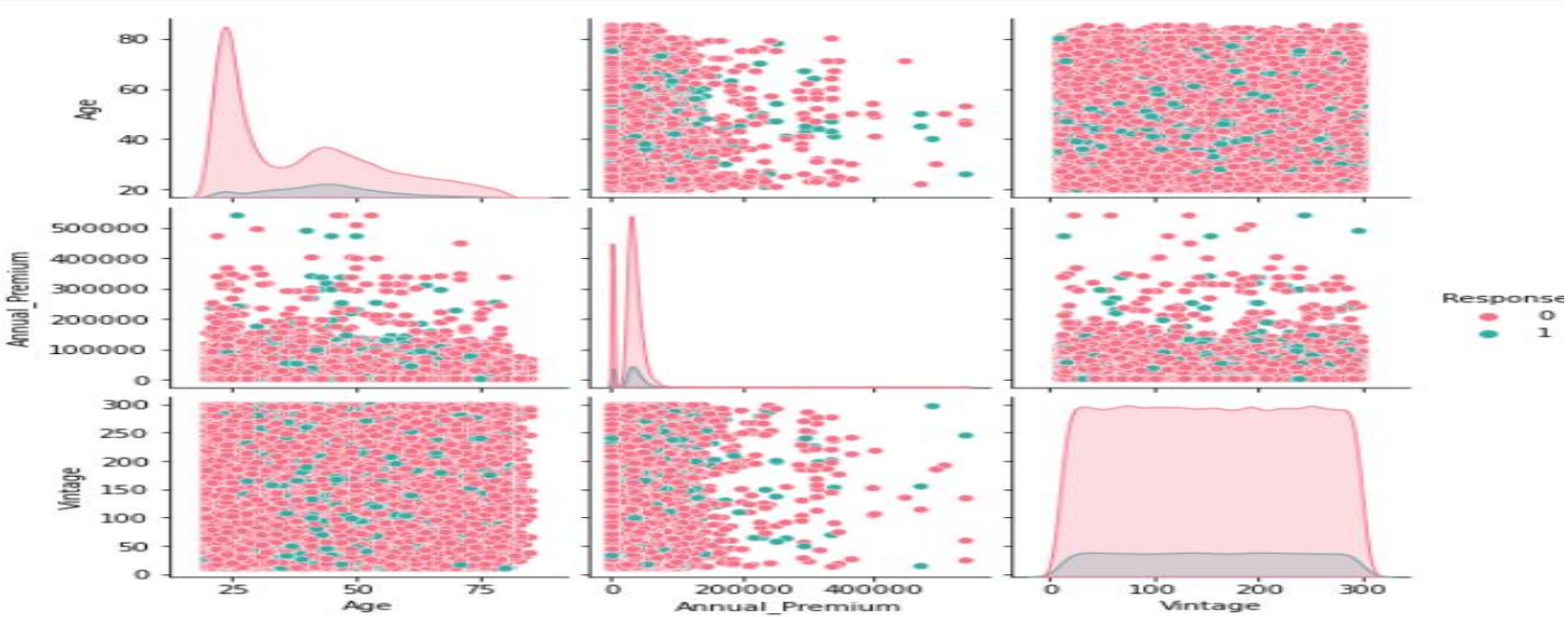

Density
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50
60
70
80
Age
Response
0
1

EDA

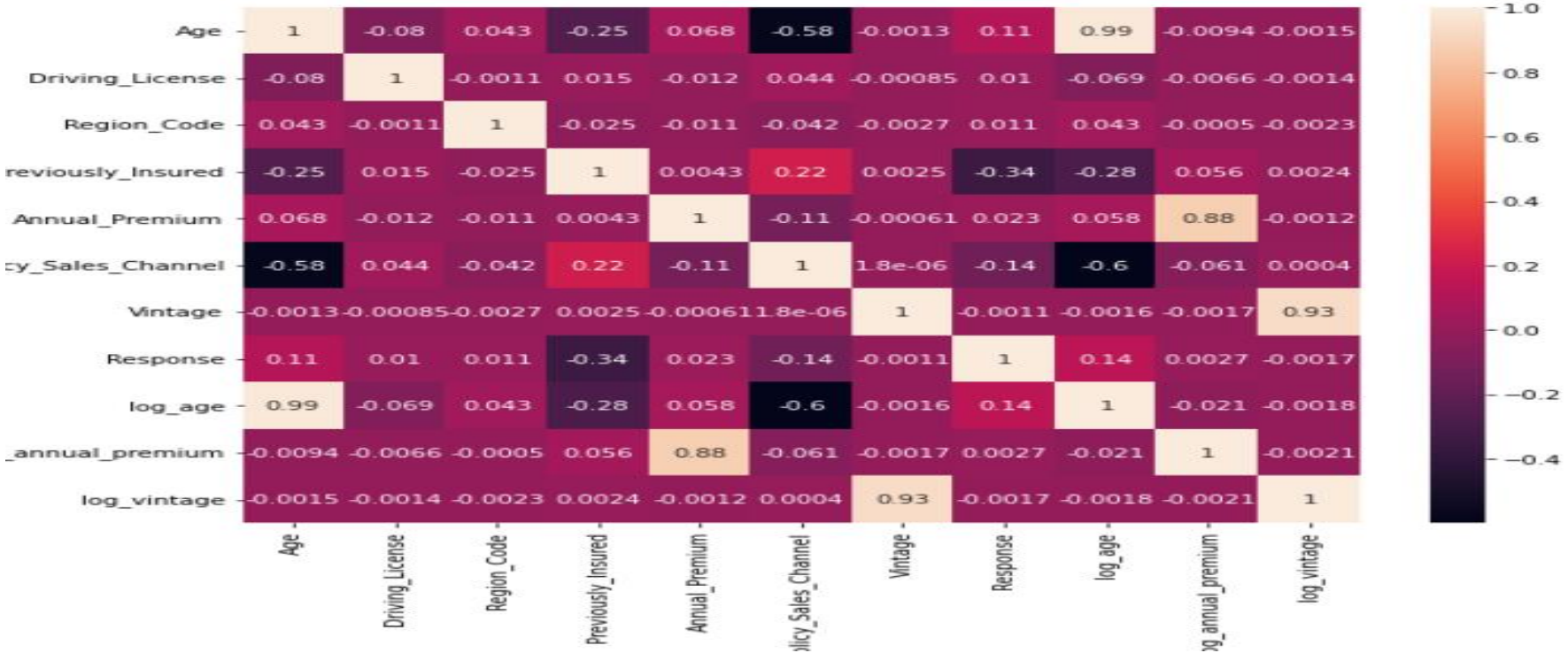

Response
0
1
160000
140000
120000
100000
80000
60000
40000
20000
0
count
> 2 Years
1-2 Year
< 1 Year
Vehicle_Age


Response
0
1
175000
150000
125000
100000
75000
50000
25000
0
count
Yes
No
Vehicle_Damage

CORRELATION(EDA)

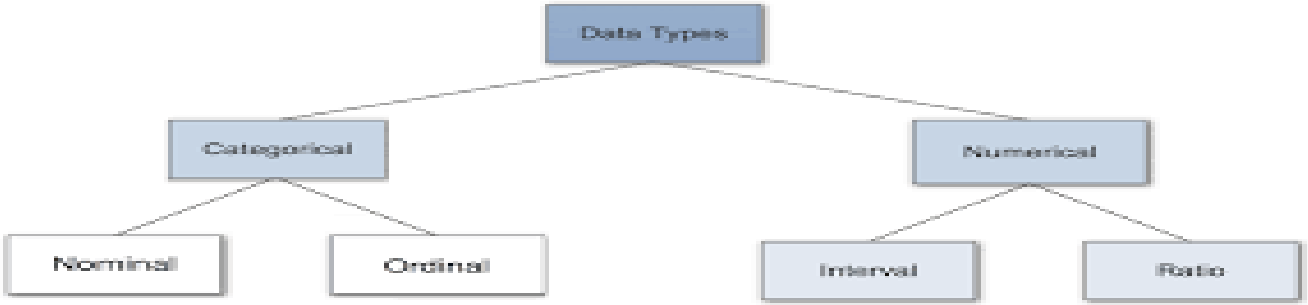


CORRELATION



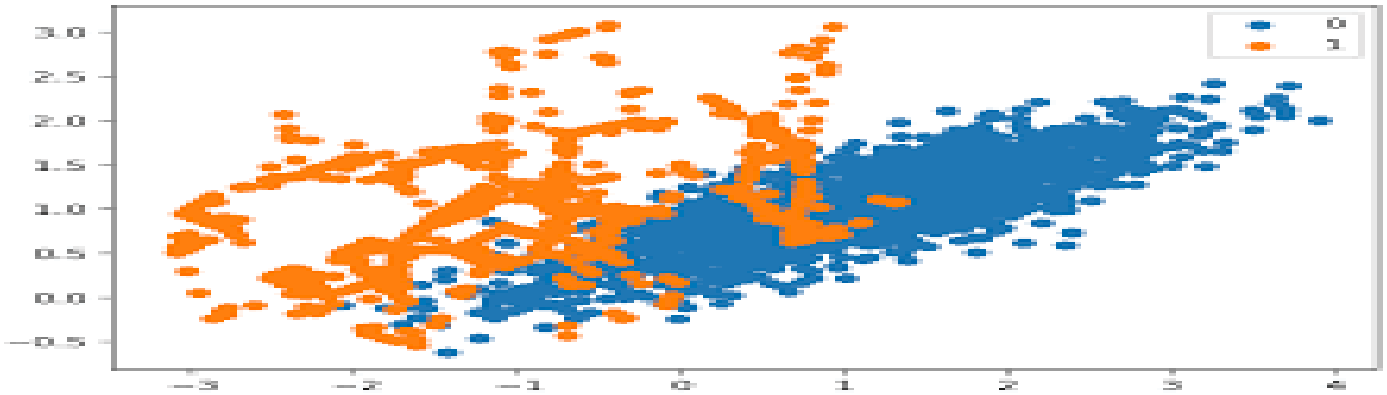
CATEGORICAL VARIABLES

- Numerical Data
- Ordinal Variables
- Nominal Variables



SMOTE

- SMOTE stands for **Synthetic Minority Oversampling Technique**. This is a statistical technique for increasing the number of cases in your dataset in a balanced way. The module works by generating new instances from existing minority cases that you supply as input



EVALUATION

	Accuracy	F1 Score	Precision	Recall	ROC AUC
Decision Tree	0.764017	0.799042	0.867982	0.764017	0.789809
Random Forest	0.762600	0.798253	0.869635	0.762600	0.822426
Gradient Boosting	0.690863	0.744457	0.899840	0.690863	0.853096
XGBoost	0.687924	0.742038	0.900235	0.687924	0.853234
CatBoost	0.762012	0.799101	0.879443	0.762012	0.846597
LGBM	0.724595	0.771543	0.894817	0.724595	0.853723

Lots of Confusion

- BEFORE SMOTE **VS** AFTER SMOTE

```
 [ 8294  3273]]
for Random Forest
[[78318  5393]
 [ 8979  2588]]
for Gradient Boosting
[[83708     3]
 [11564     3]]
for XGBoost
[[83711     0]
 [11567     0]]
```

```
for Decision Tree
[[65066 18645]
 [ 3671  7896]]
for Random Forest
[[64665 19046]
 [ 3472  8095]]
for Gradient Boosting
[[55791 27920]
 [  869 10698]]
for XGBoost
[[55432 28279]
 [  812 10755]]
```

THINGS THAT DIDN’T WORK

- Removing outliers
- Without SMOTE performance was dull
- Log normalization didn’t work
- No improvement with hyperparameters

Conclusion

- **We observed that Hyperparameter tuning with randomized search CV did not improve the model performance**
- **We are sticking to our original random forest, LGBM and catboost model as it gave food accuracy on both Responses(0 and 1) and performed well on other matrices too**
- **Random Forest ROC AUC =82% and ACCURACY = 76%**
- **ACCURACY = 88% without SMOTE**
- **Outlier removal gave worst results**

Q & A