Capstone Project

Health Insurance Cross Sell Prediction

BY – AYUSH SHARMA

CONTENT

- ❑ Problem Statement
- ❑ Data Description
- ❑ EDA
- ❑ Models Used
- ❑ Models Evaluation metrics
- ❑ Model Selection
- ❑ Conclusion

Problem Statement

Our client 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.

Data Description

- The Dataset provided consisted of 381109 rows and 12 columns.

- There were zero duplicates and null values in the Dataset.

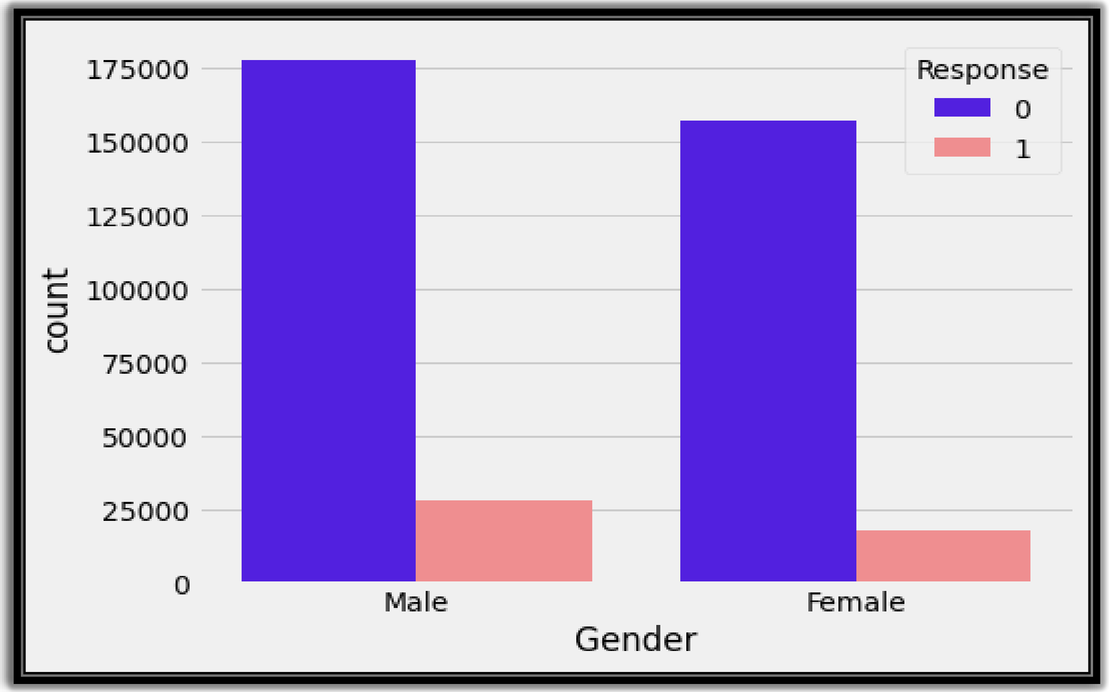
Dependent Variable - Response

Independent Variables –

- Id
- Annual Premium
- Gender
- Age
- Driving License
- Region
- Previously Insured
- Vehicle Age
- Vehicle Damage
- Policy Sales Channel
- Vintage

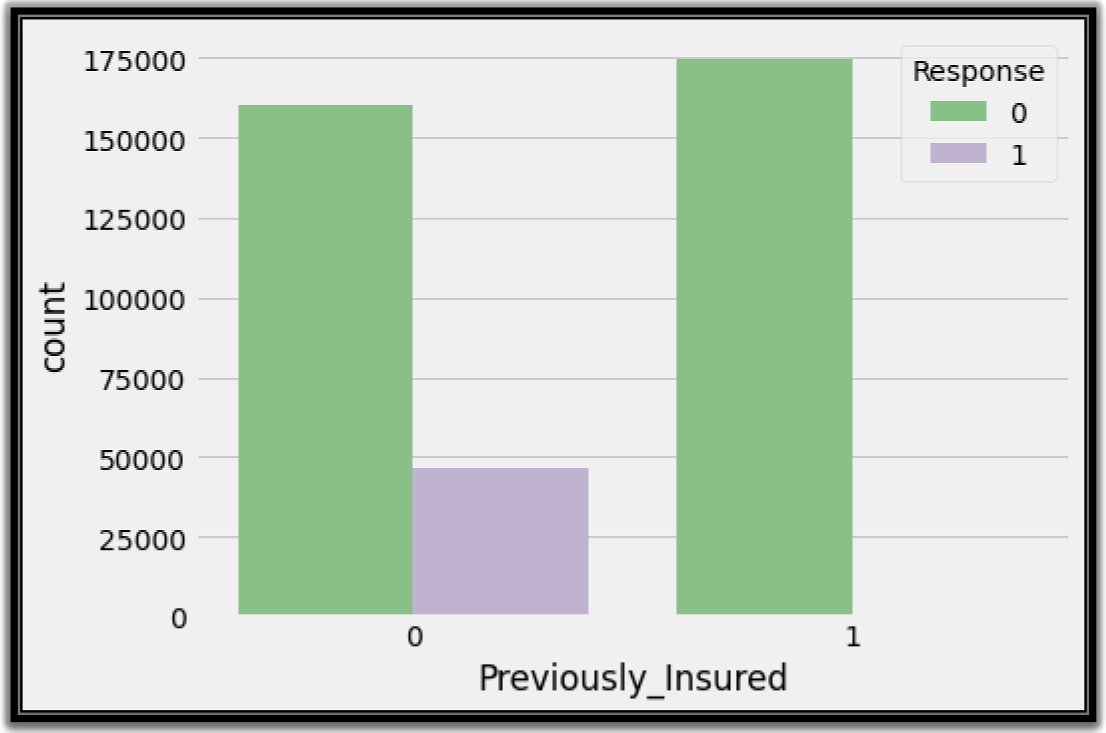
EXPLORATORY DATA ANALYSIS (EDA)

Gender Wise Response



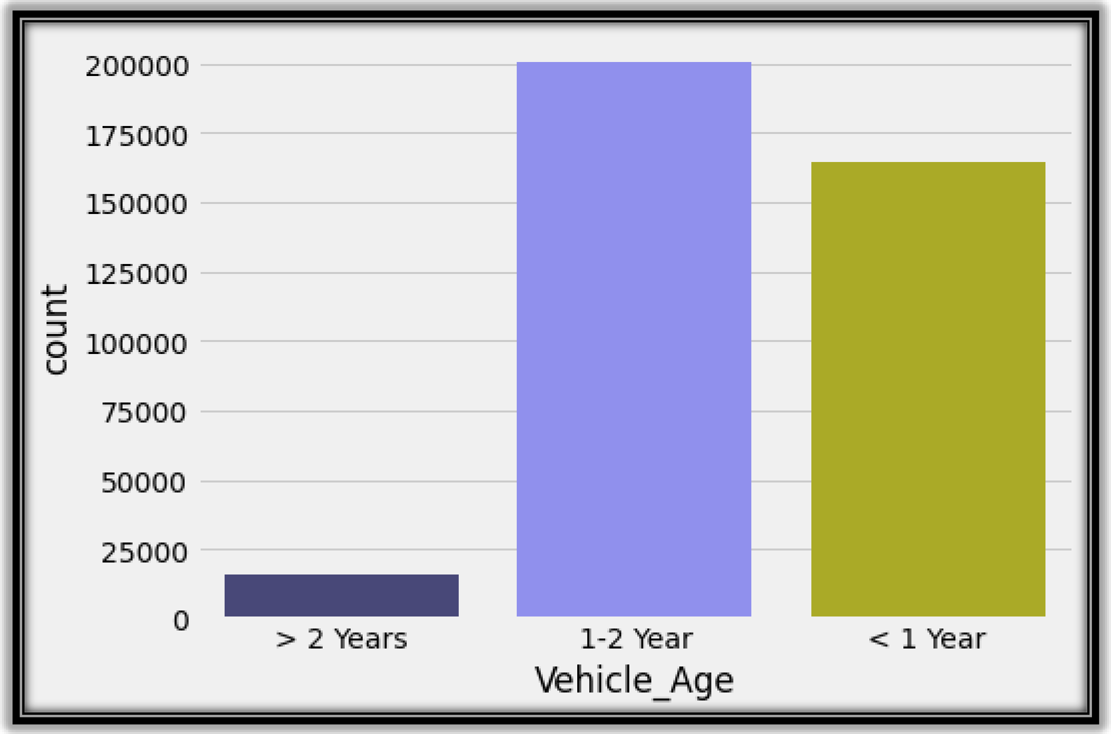
In the above plot we can see that there are slightly more chances of positive response if the customer is male

Response With Respect To Previously Insured



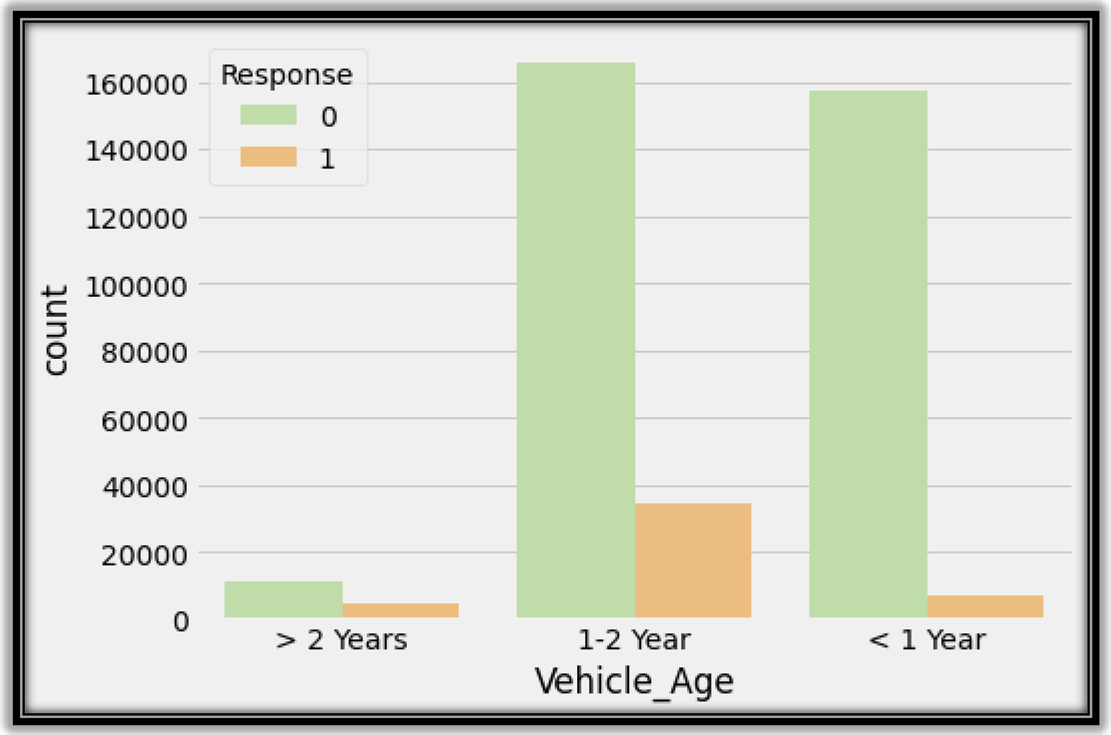
People who already have a vehicle insurance are not interested

Vehicle Age



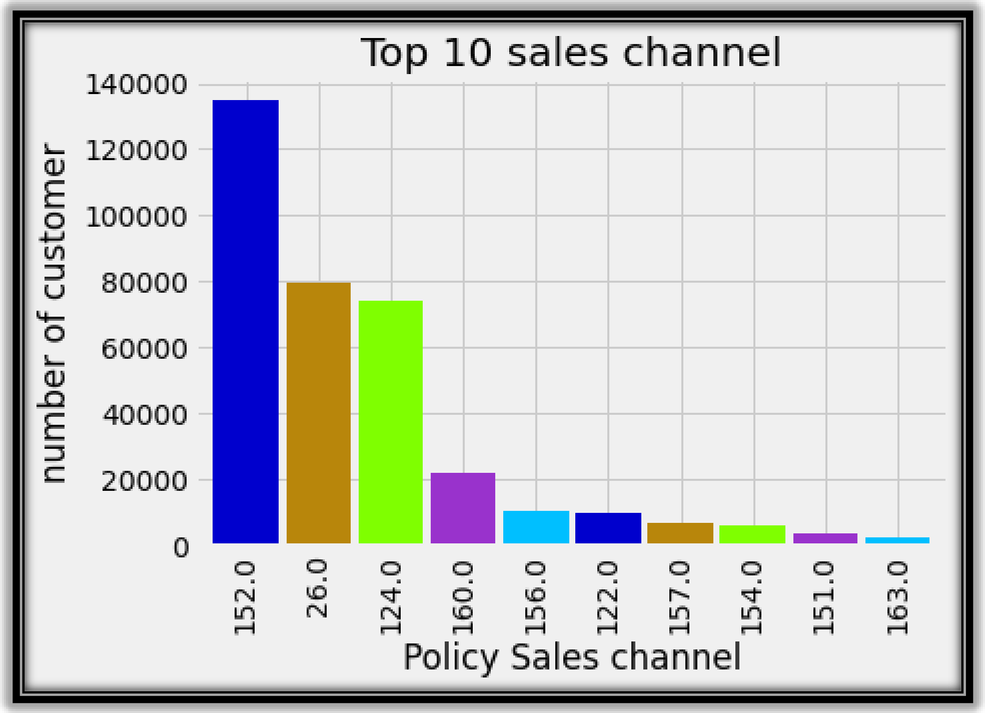
Most people either have Vehicle for less than a year or for 1-2 years

Response With respect To Vehicle Age



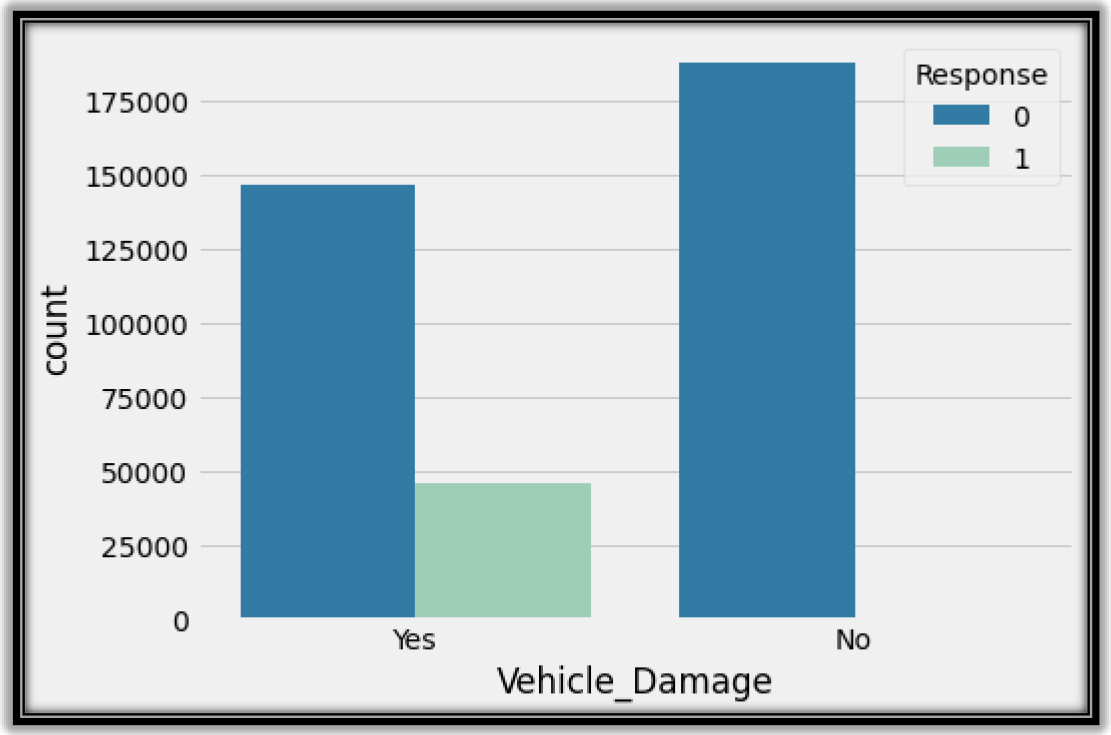
- **People having vehicle for 1-2 years are more likely to have a positive response towards insurance**

Top 10 Sale Channel



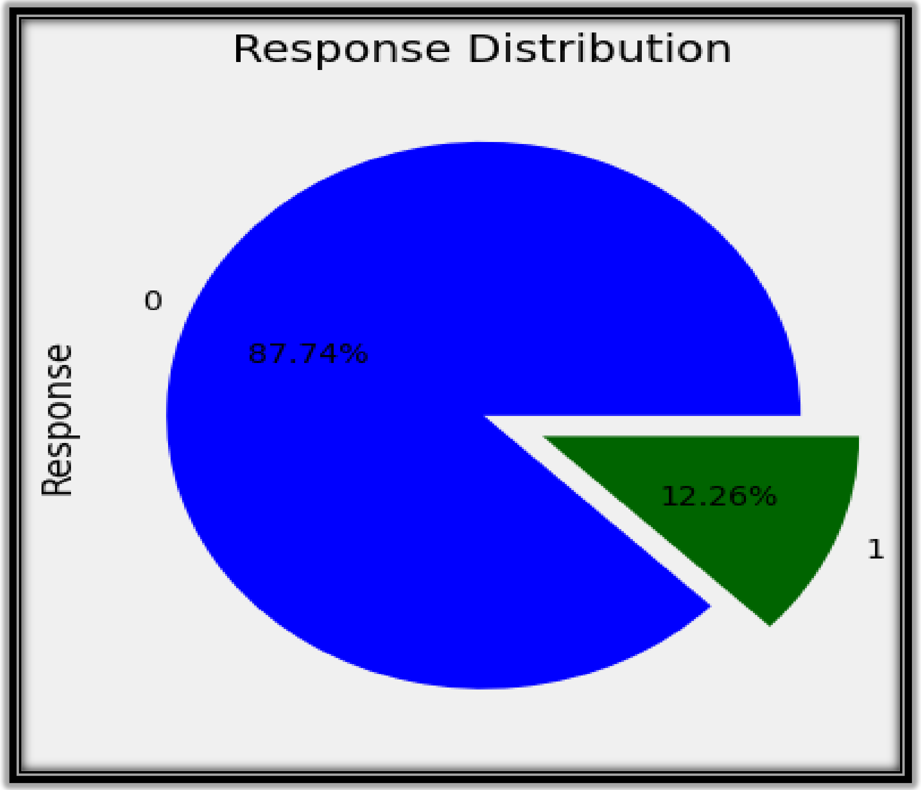
- In the above plot we clearly see that policy sales channel 152.0 was able to reach most number of customer followed by number 26.0

Response With Respect To Vehicle Damage



- People who got their vehicle damaged in the past are more likely to show positive response towards vehicle insurance

Response Count

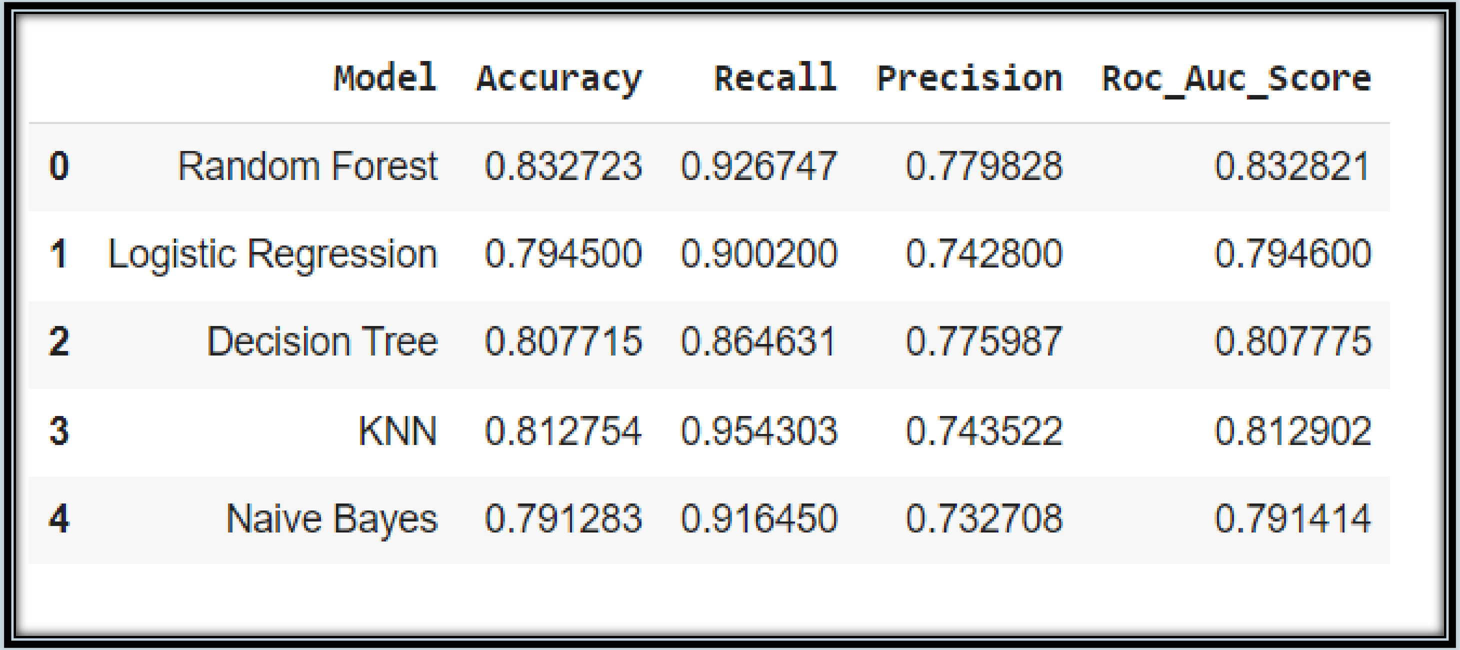


- From the above plot we can see that the dependent variable is highly imbalanced

Model Used

- Random Forest
- Logistic Regression
- Decision Tree
- K-Nearest Neighbor
- Naive Bayes

Evaluation Metrics



	Model	Accuracy	Recall	Precision	Roc_Auc_Score
0	Random Forest	0.832723	0.926747	0.779828	0.832821
1	Logistic Regression	0.794500	0.900200	0.742800	0.794600
2	Decision Tree	0.807715	0.864631	0.775987	0.807775
3	KNN	0.812754	0.954303	0.743522	0.812902
4	Naive Bayes	0.791283	0.916450	0.732708	0.791414

Model Selection

- By looking at the Evaluation metrics for each the model we can clearly see that all of the models performed similarly. There was not much difference in accuracy, precision, recall and roc auc score.

- The best performing model was Random forest with highest accuracy, precision and roc auc score

- In terms of recall K-nearest Neighbors was the best model.

Conclusion

- Out of all the models used, with Random Forest model we were able to get the highest accuracy, precision, roc auc score of 0.832, 0.779, 0.832 respectively.

THANK YOU