Capstone Project - 3

Supervised ML Classification : Health Insurance Cross Sell Prediction

By
Pankaj Kumar,
Soniya kumawat,
Ankit Sharma
Data Science Trainee,
AlmaBetter

Data Summary

Health Insurance

- ID
- Gender
- Age
- Driving License
- Region code
- Previously Insured
- Vehicle Age
- Vehicle damage
- Annual premium
- Policy Sales Channel
- Vintage
- Response

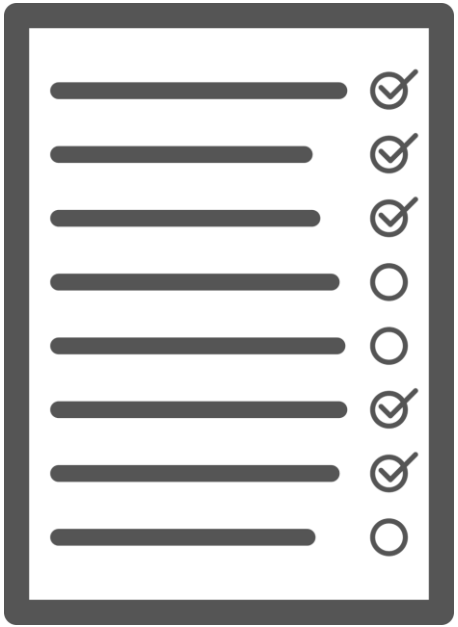
So, what factors influence buying a vehicle insurance?

Buying a vehicle insurance Depends upon below mentioned factors:-

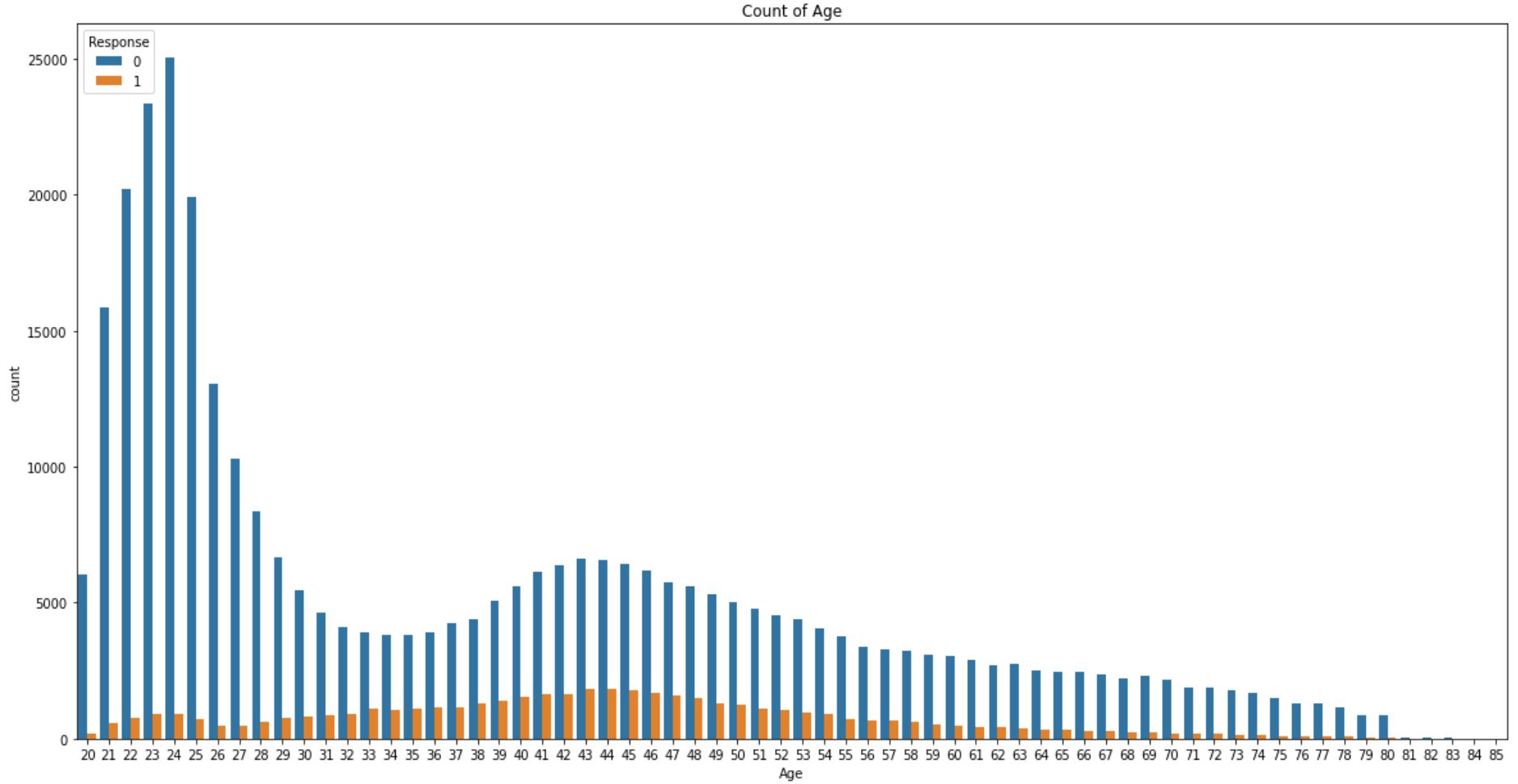
- Gender of the customer i.e., male or female
- Age of the customer
- Area in which customer reside
- Vehicle age
- Condition of Vehicle
- Types of policies

Agenda

- Health Insurance data analysis
- Categorical Analysis
- Group of people interested in buying vehicle insurance
- Average age of vehicle for which customers are interested in buying vehicle insurance
- Condition of vehicle(level of damage)
- Average Annual premium paid by the customer
- Data realisation
- Model performance analysis



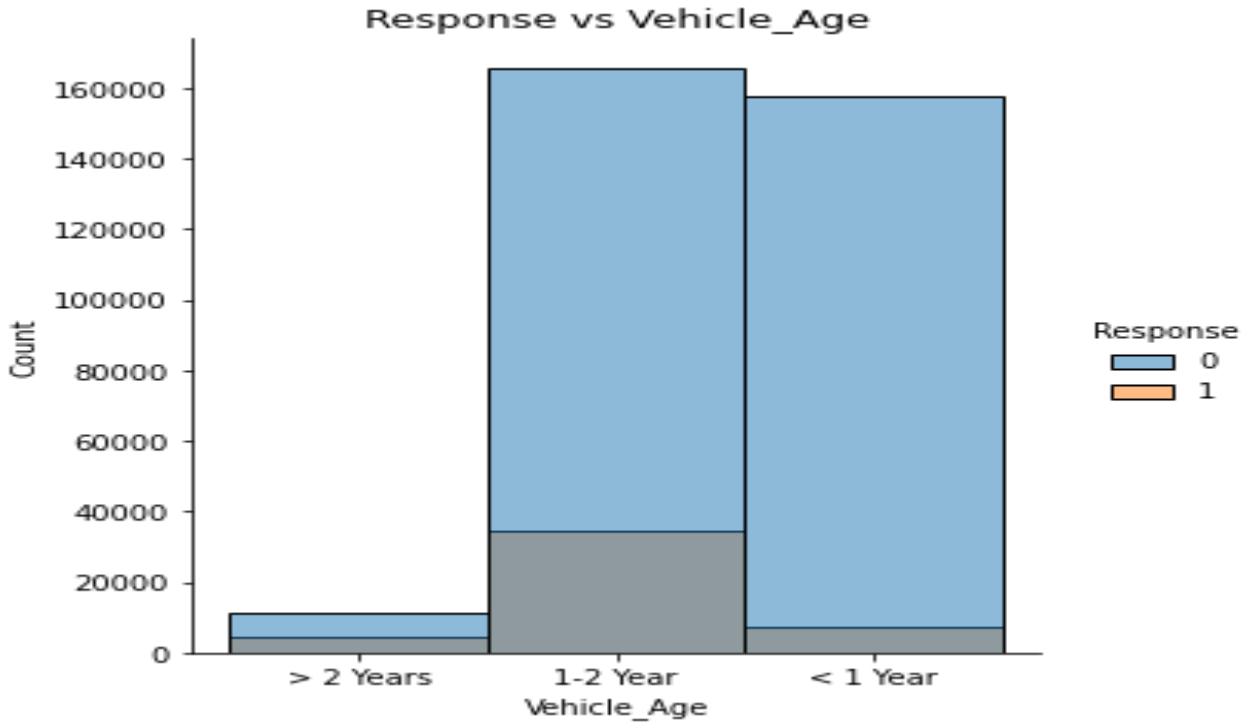
Age Group Vs Response



Categorical Analysis

Response Vs Vehicle age

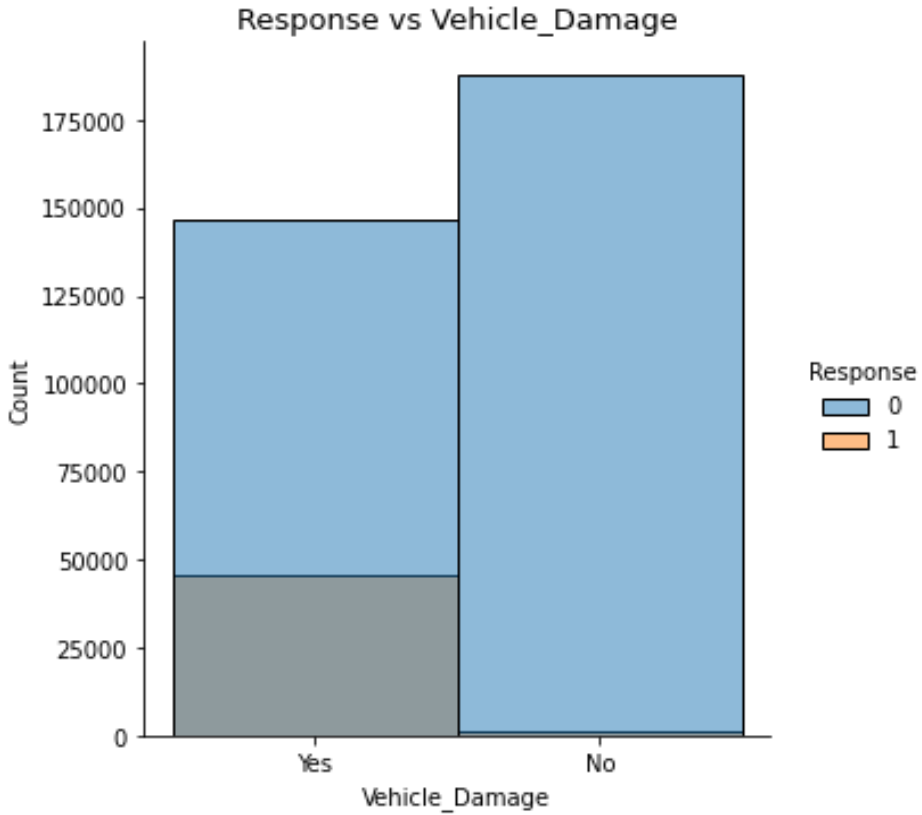
- This graph shows us that in with comparison to the vehicle age to intertest of people wanting to buy Insurance.
- vehicle with age 1-2 years are interested in taking insurance



Categorical Analysis (Contd.)

Response Vs Vehicle Damage

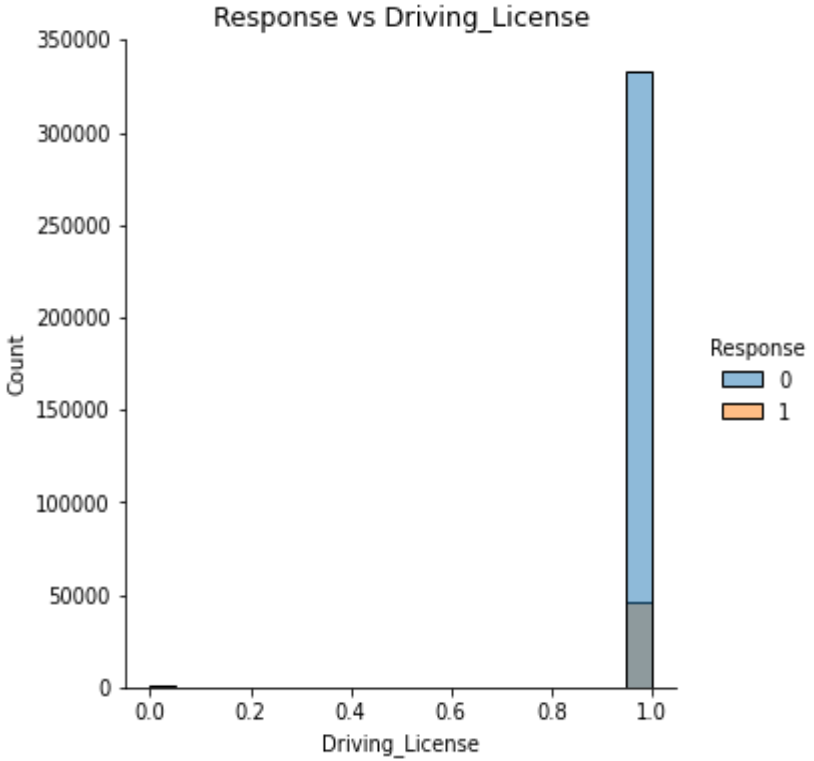
- This graph shows us people's interest in buying insurance in comparison to their level of vehicle damage.
- Persons with vehicle damage only are interested in insurance

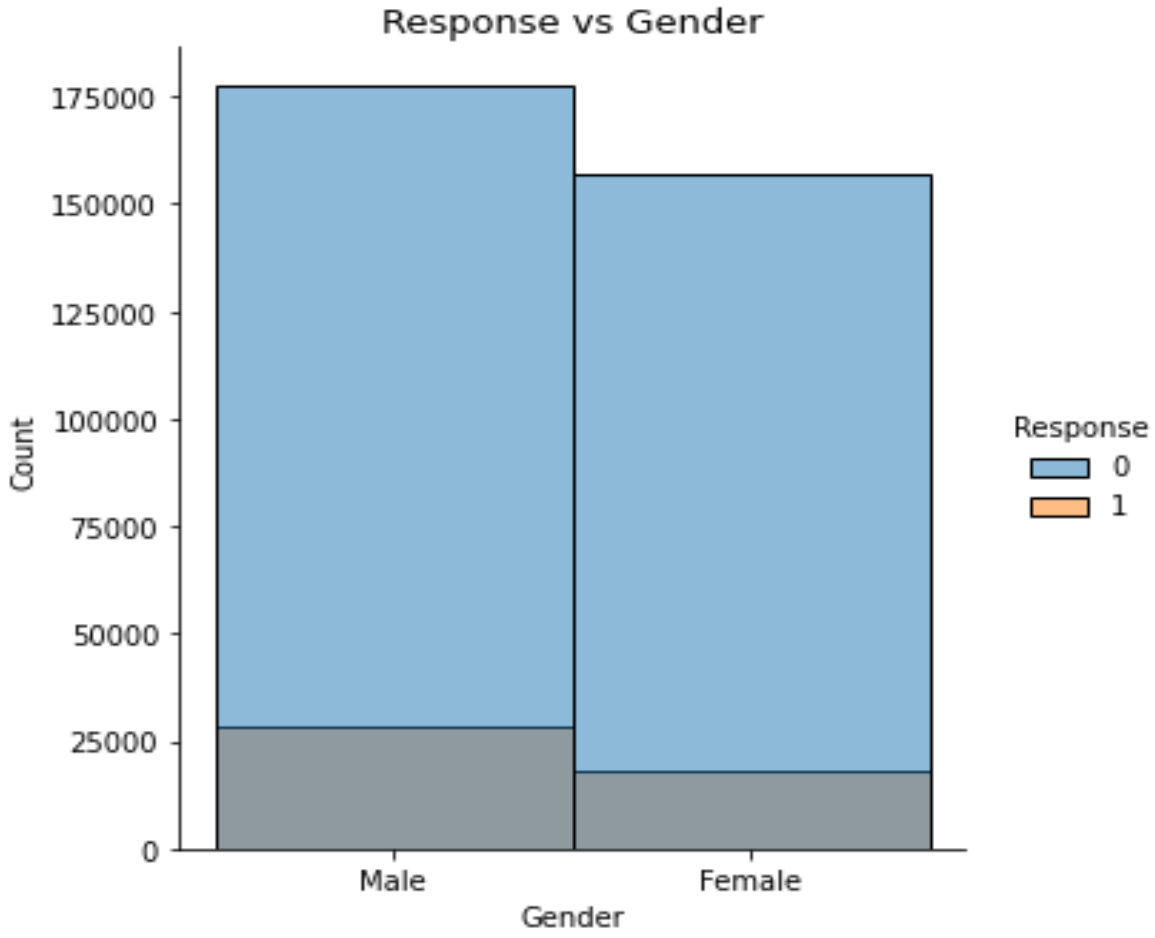


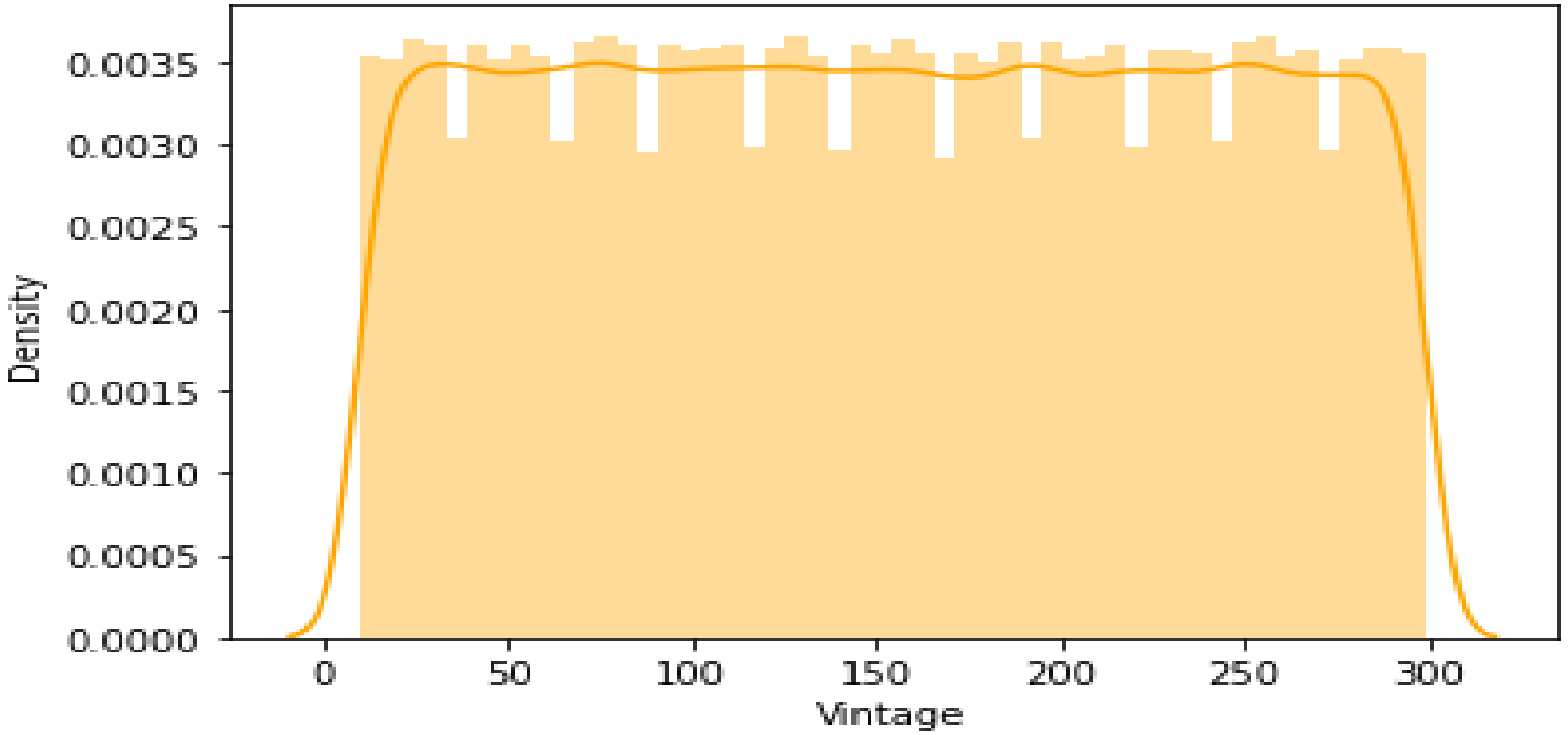
Categorical Analysis Continued

Response Vs Driving License

- This graph shows us that how many respondents willing to buy Insurance were having a Driving License.
- Customers who are interested in Vehicle Insurance almost all have driving license

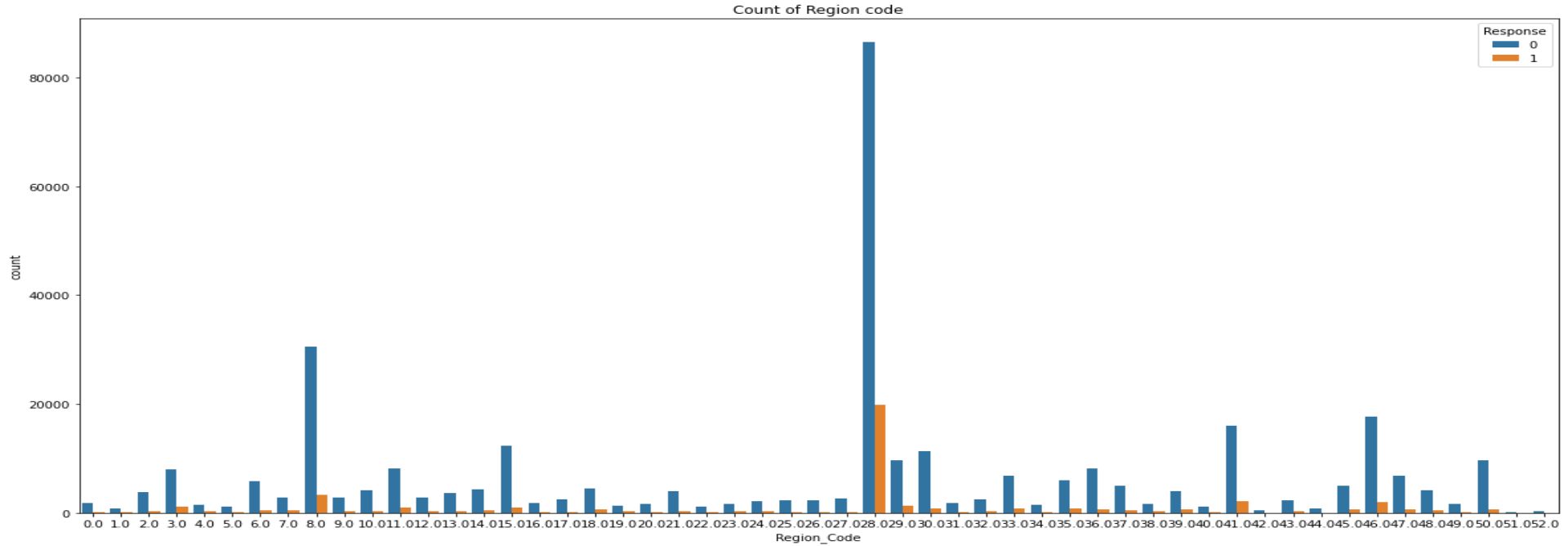



Response vs Gender
175000
150000
125000
100000
75000
50000
25000
0
Count
Male
Female
Gender
Response
0
1



The Feature Vintage has very less information and is Uniformly Distributed , With no skew .Also, the Values are uniformly mixed , in both the classes of the target variable response .

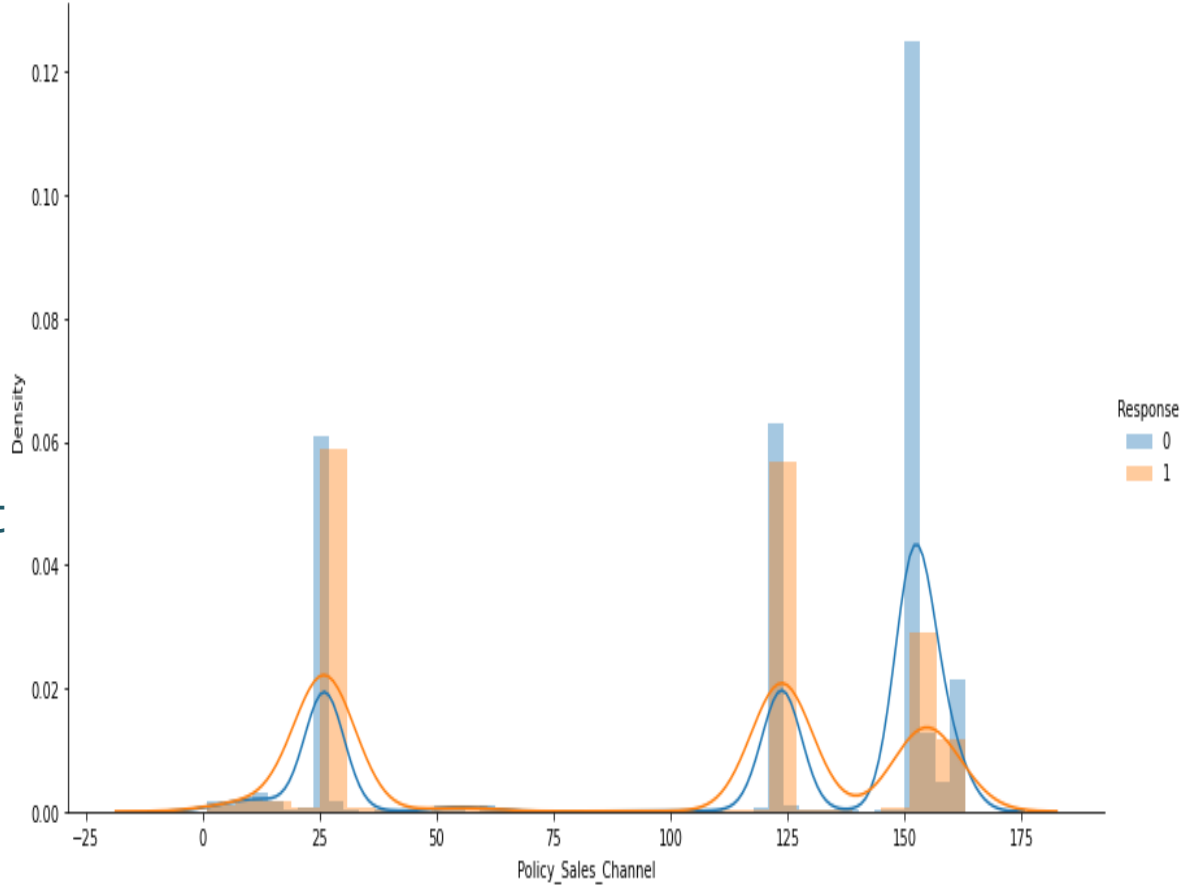
Area Vs Response



- The individuals with region code 28 are the most as compared to the other ones are interested in insurance

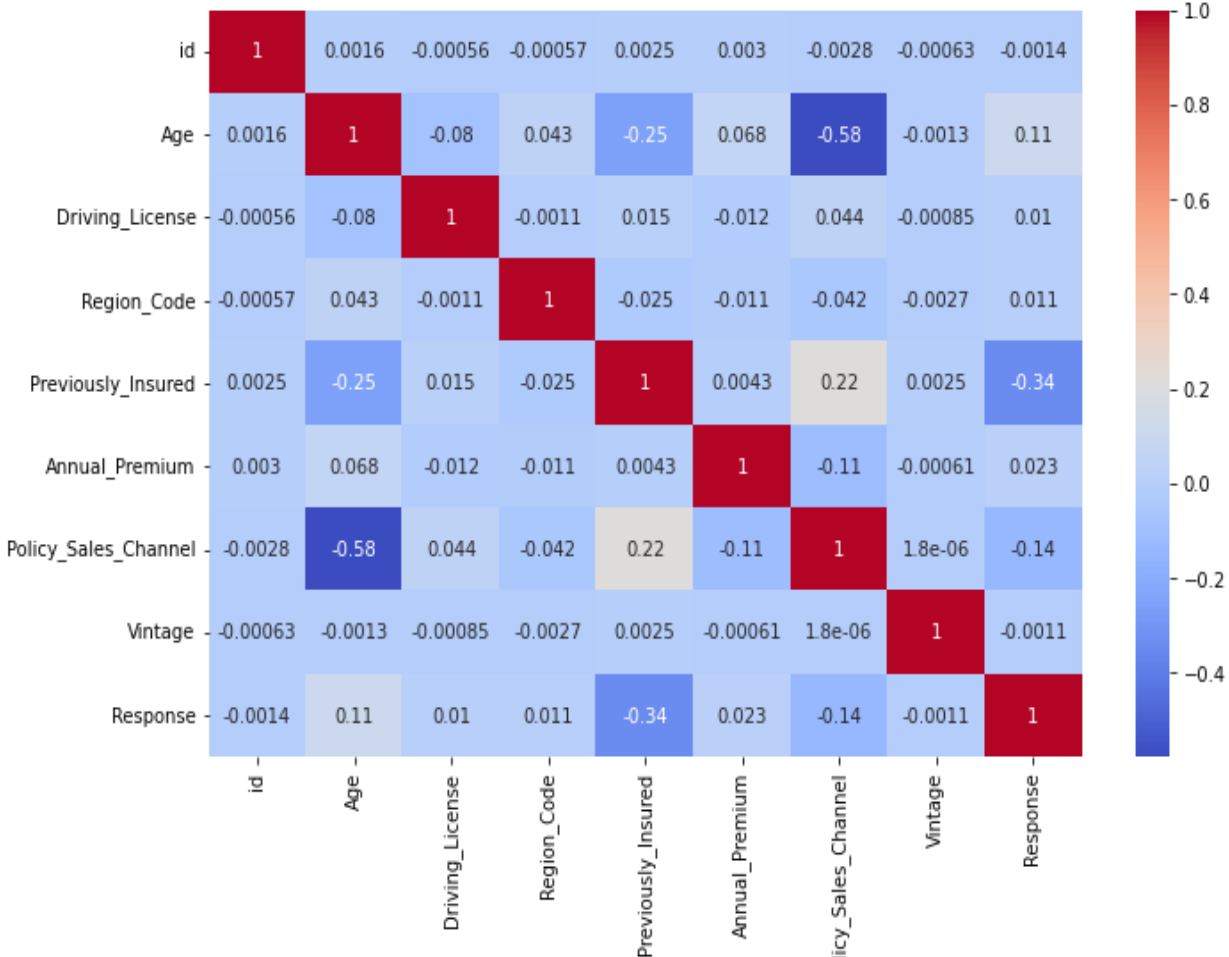
Response Vs Policy sales channel

- This graph shows us various responses recorded for various policy sales channel.
- It is clear from this graph that the most used sales channels are 152, 26 and 124. The best channel that results in customer interest is 152



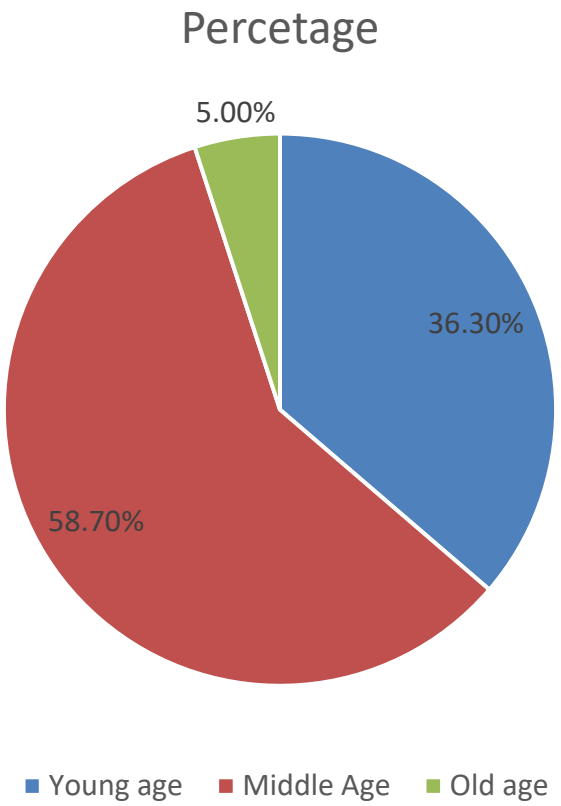
Heat map for correlation between different variables

- This graph is a heatmap in which we have taken for various variables to find the relation between them.
- We can infer from this graph that Age and annual premium is highly corelated i.e. 68% which means they are highly related with each other in comparison to other variable relations.



Response Vs Age group

- In this pie we are showing that how many percent of people of various age groups are interested in taking Insurance.
- It is clear from the Pie chart that Middle age people are the most interested in taking Insurance and Old age are the least Interested.



Conclusion

- We've drawn many inferences from the health insurance cross-sell prediction data , here's a summary of the few of them: this is confirmed with both the bivariate analysis of each feature , as well as the Feature Importance's returned by the notebook.

- Customers of age between 30 to 60 are more likely to buy insurance.

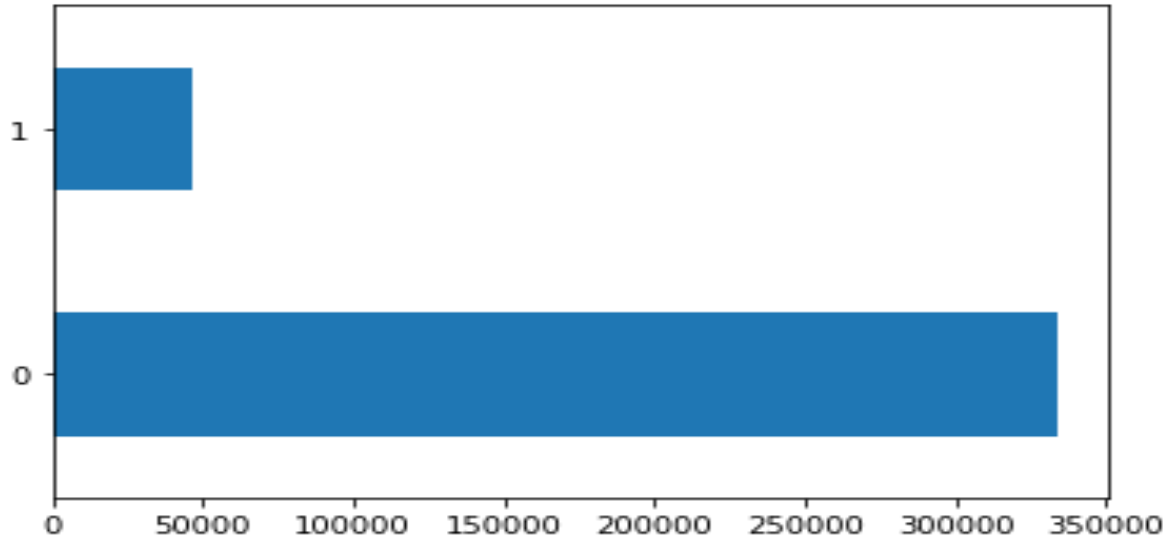
- Customers with Driving License have higher chance of buying Insurance.

- Customers with Vehicle Damage are likely to buy insurance.

- The variables : Age, Previously insured, Annual premium are more affecting the target variable.

Label encoding and Data preprocessing

- Changed all the categorical values to integer values
- Plotted a bar graph to check whether data is balanced or not
- Used smote analysis to balance the data



Machine Learning Modeling

1)Logistic Regression

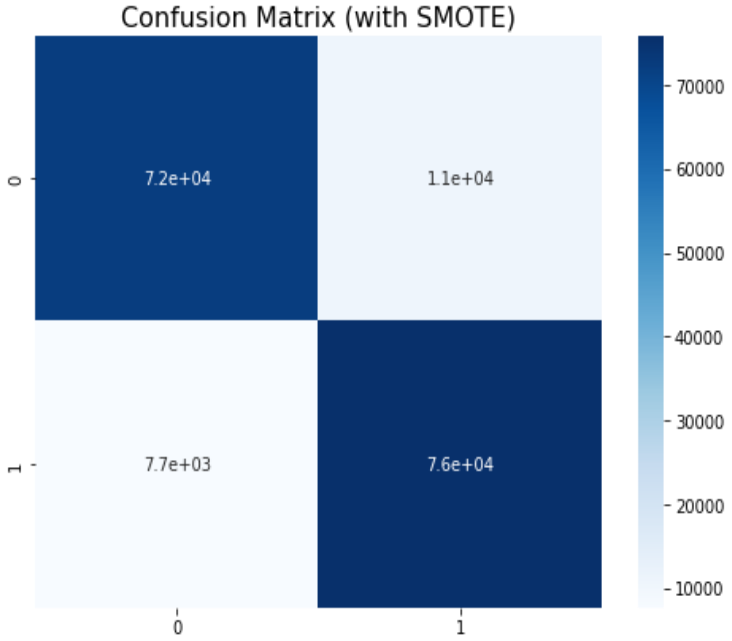
2)Random forest classifier

3)Decision Tree

4)K nearest neighbor

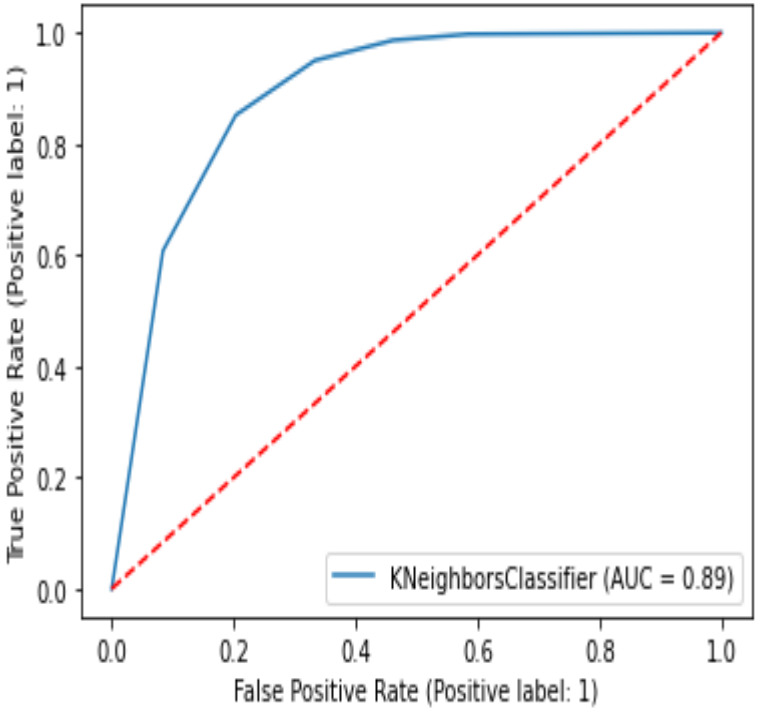
Random forest classifier

- Confusion matrix is a tool used in machine learning to evaluate performance of a classification model.
- It compares actual value with the predicted value.
- .We could not find much difference w.r.t before/after Hyperparameter Tuning. slight difference in Auc and F1 score
- Performance of Random forest Classifier
  - Accuracy : 0.889683014354067
  - Precision : 0.9082779723414174
  - Recall : 0.8734221846132149
  - F1 Score : 0.8807874417033582



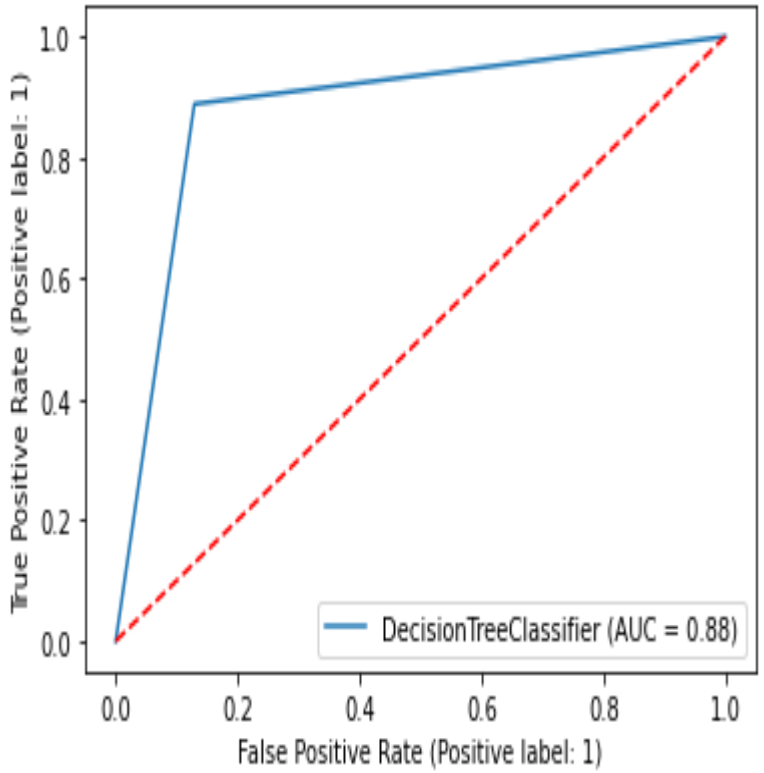
Performance of KNN classifiers

- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- From the given graph we infer that KNN stores all the available data and classify a new data point based on the similarity.
- Performance of KNN Classifier
  Accuracy : 0.8085526315789474
  Precision : 0.9499420293319627
  Recall : 0.7407010447636001 F1 Score : 0.832373271889401



Performance by Decision Tree Classifier

- In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- Performance of Decision Tree Classifier
  Accuracy : 0.879683014354067
  Precision : 0.8882779723414174
  Recall : 0.8734221846132149 F1 Score : 0.880787441703358 2

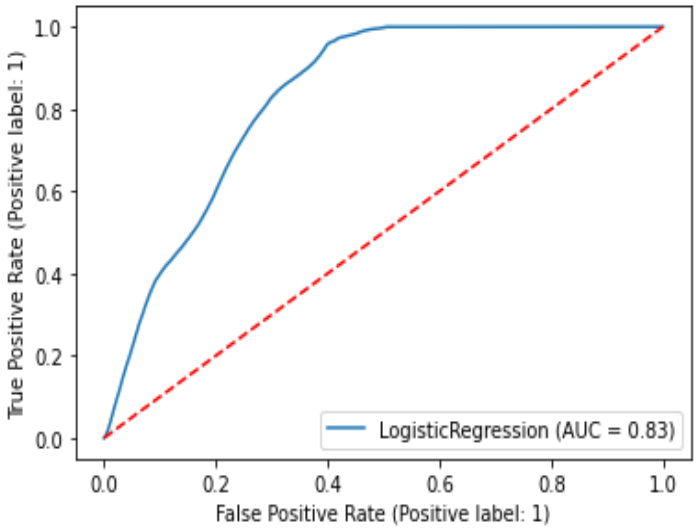


Logistic regression

Performance
Accuracy : 0.80
Precision : 0.94
Recall : 0.74
F1 Score : 0.83

After hyperparameter tuning

Accuracy : 0.80
Precision : 0.94
Recall : 0.75
F1 Score : 0.83



Conclusion(Contd.)

- Comparing ROC curve we can see that Random Forest model preform better. Because curves closer to the top-left corner, it indicate a better performance.

- In this modelling we used various methods which are as follows:-
- Hyperparameter tuning

- Feature Engineering such as concatenation, aggregation, binning.

Thank You