Capstone Project

Health Insurance Cross Sell Prediction

By

Ravi Kumar

Content

AI

- Problem Statement
- Data summary
- Exploratory Data Analysis(EDA)
- Feature Engineering
- Building Model
- Machine Learning Algorithm
- Result of all the model
- Conclusion

Problem Statement

AI

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 the 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. 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.

Data Description

AI

- id: Unique ID for customer
- Age: Age of the customer
- Driving_License 0 : Customer has DL or not
- Region_Code: Unique code for the region of the customer
- Previously_Insured 1 : Customer already has Vehicle Insurance or not
- Vehicle_Age: Age of the Vehicle
- Vehicle_Damage 1 : Past damages present or not
- Annual_Premium: The amount customer needs to pay as premium
- Policy Sales Channel: Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person
- Vintage: Number of Days, Customer has been associated with company
- Response :Customer is interested or not

Insights from the Problem and Data

AI

What is an insurance firm?

- **If a loss occurred a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium.**

What is the probability of buying an insurance?

- **In insurance industry, it refers to a situation in which people only buy insurance when they expect high risks. Buying insurance is not appropriate for all levels and types of risks. In many cases, people are better off taking actions to avoid risk, retain (accept) risk or reduce risk. Buying insurance makes the most sense when the potential loss is great and there is a significant probability of loss.**

Insights from the Problem and Data

AI

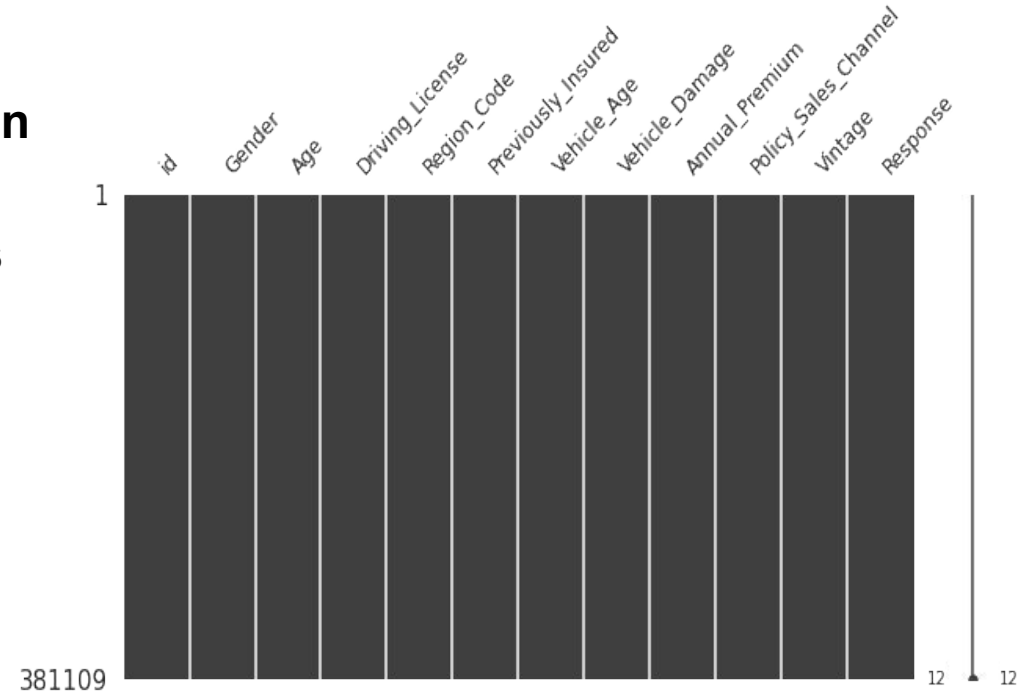
How many people are knowledgeable about insurance policy and how many of them claim insurance?

- **Let's say about four in 10 men describe themselves as being very knowledgeable about life insurance. As in the problem statement, about 2 or 3 get hospitalized out of 100, which means 2 to 3 percent claim the insurance. This way everyone shares the risk of everyone else.**

So we need to 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. Now, we need to predict whether the customer would be interested in Vehicle insurance or not.

Checking Dataset

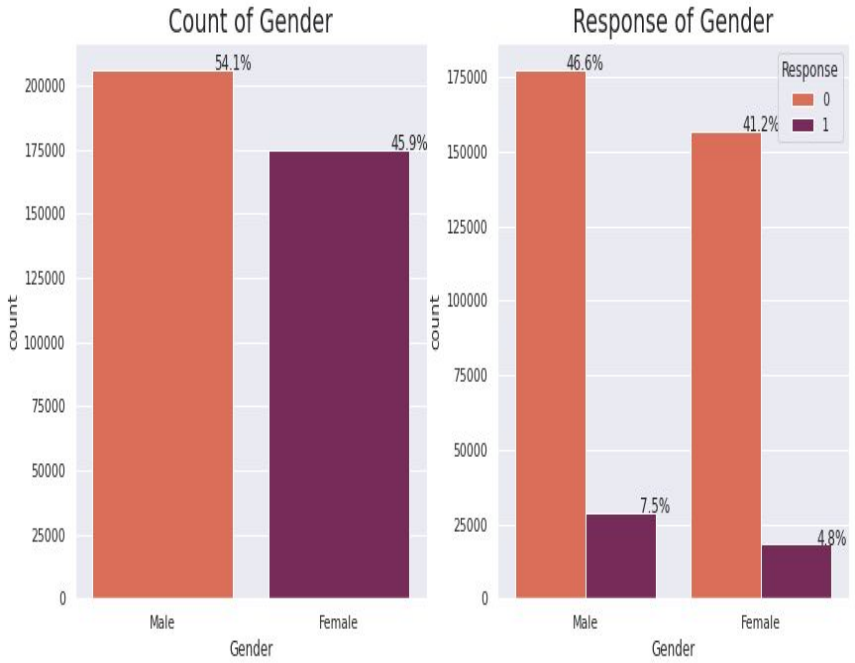
- There are no missing values in the dataset
- There are no duplicate values found in the dataset
- The shape of the dataset is 381109 rows and 12 columns



EDA: Gender variable

As we can see from the graph,

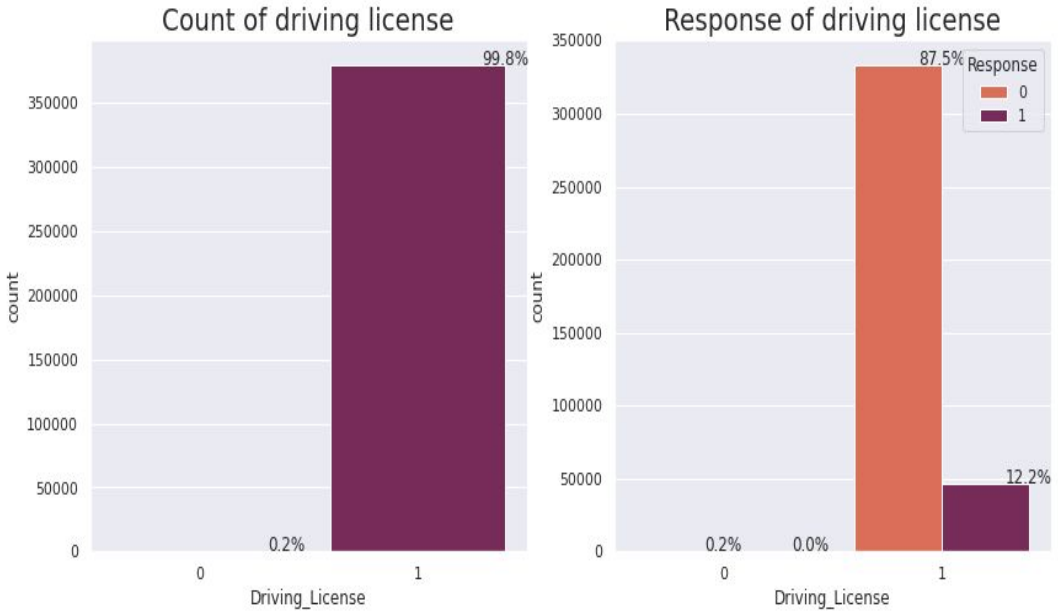
- **The gender variable in the dataset is spread nearly evenly. The male category is marginally larger than the female category, and the likelihood of purchasing insurance is also slightly higher.**
- **Only 12.3% people are interested in buying vehicle insurance and 87.7% are not interested to buy vehicle insurance**



Driving License

As we can see from the graph,

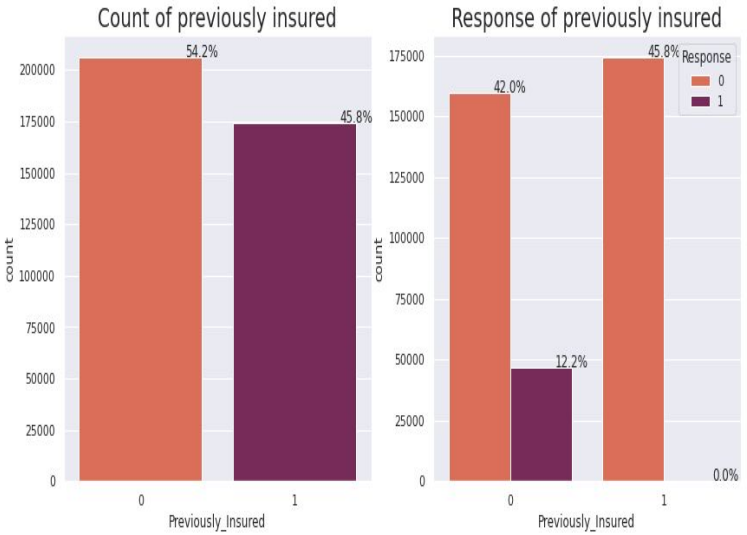
- **99.8% of customers have DL, whereas 0.2% do not have DL.**
- **Only a small percentage of people who have a DL (12.2%) are interested**



Previously Insured

As we can see from the graph,

- **45.8% people are insured previously, in that 12.2% people interested to buy the vehicle insurance again, Which means people are aware of insurance policy and ready to pay a premium amount, for better off taking actions to avoid certain risks or reduce risk.**
- **So buying insurance makes the most sense when the potential loss is great and there is a significant probability of loss.**

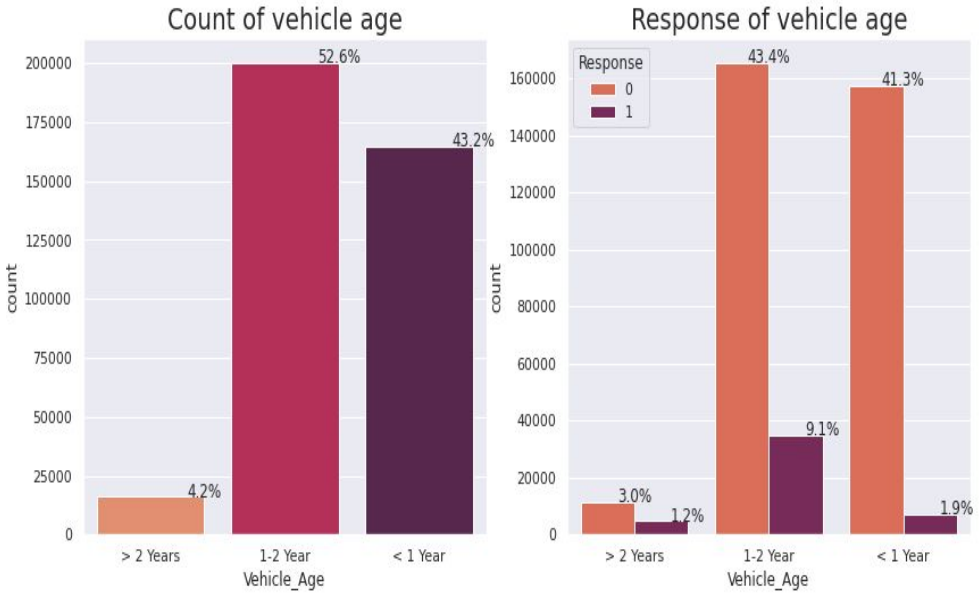


Vehicle Age

As we can see from the graph,

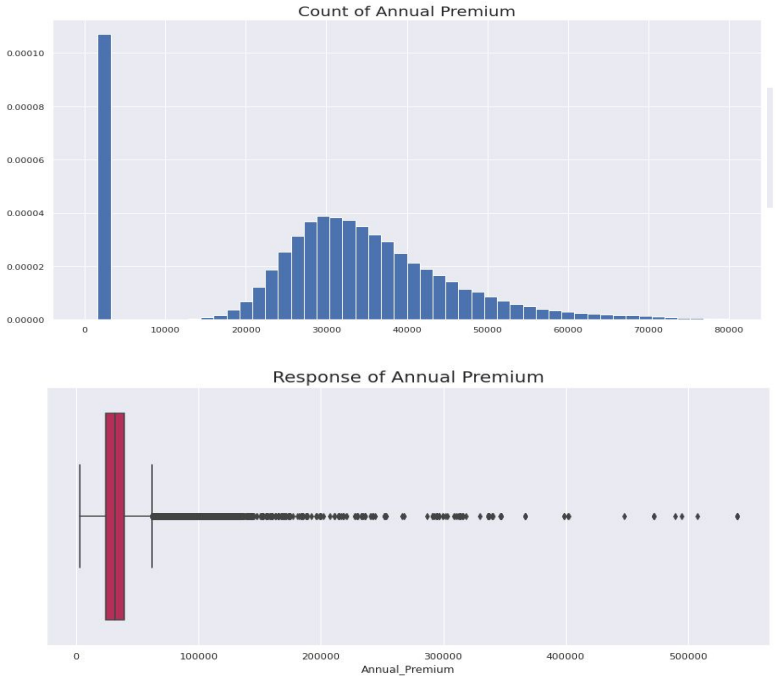
- **Around 4.2% of vehicles are more than two years old, 52.6% are between one and two years old, and 43.2% are under one year old.**

As vehicle age increases most of the people are aware of insurance and interested to buy the insurance for reducing the risk



Annual Premium

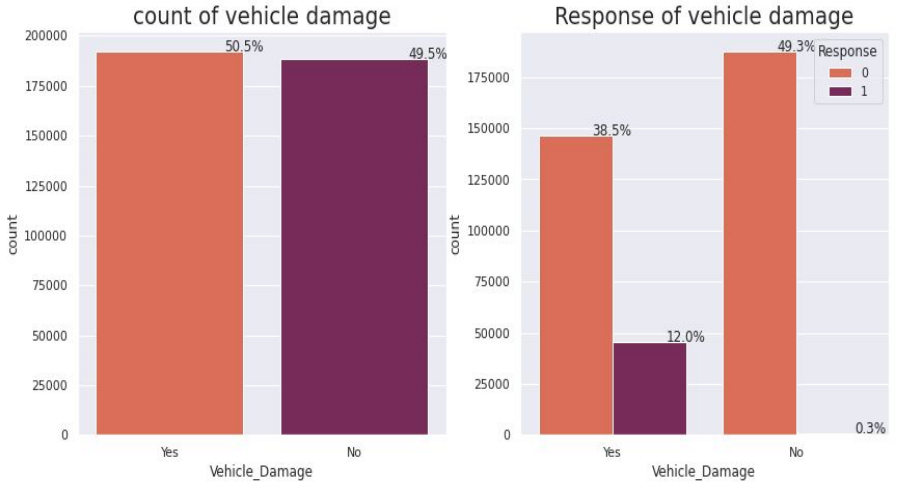
- From the distribution plot we can infer that the annual premium variable is right skewed
- From the boxplot we can observe lot of outliers in the variable



Vehicle Damage

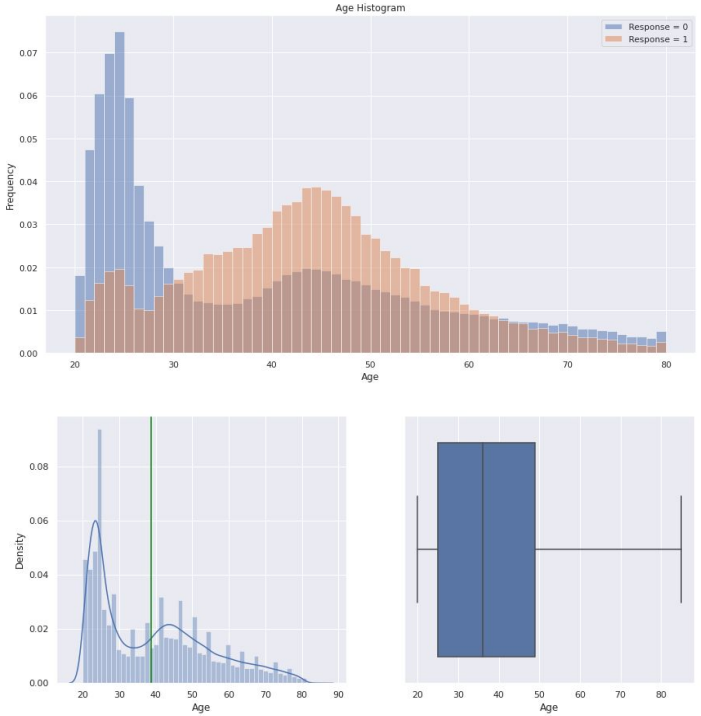
As we can see from the graph,

- **50.5% of the vehicles have past damage**
- **12.0% of people who have had a damaged vehicle in the past want to acquire vehicle insurance**



Age

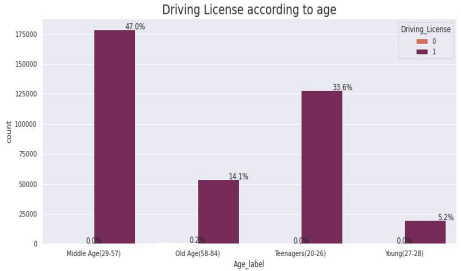
- The dataset has more individuals with an age of 24.
- 40 to 60-year-olds had a higher likelihood of purchasing vehicle insurance.
- From the above boxplot we can see that there no outlier in the Aege column.

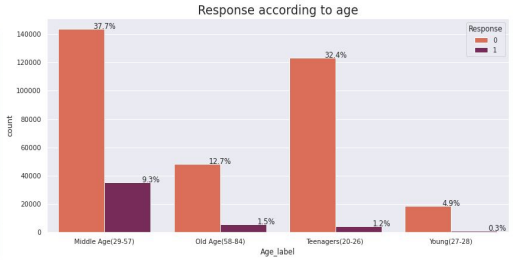


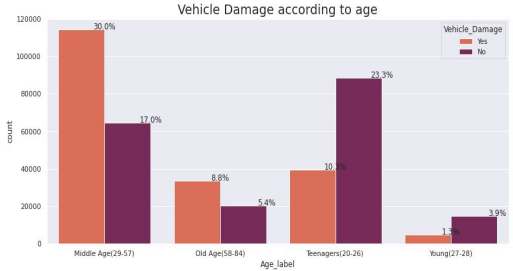
Age According to Response

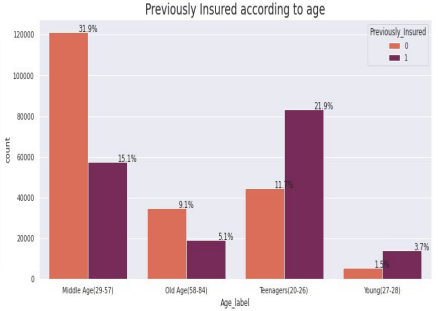
According to age wise:

- **9.3% of people in their middle age people are interested in purchasing insurance.**
- **Almost 47% of middle-aged individuals have a driver's licence.**







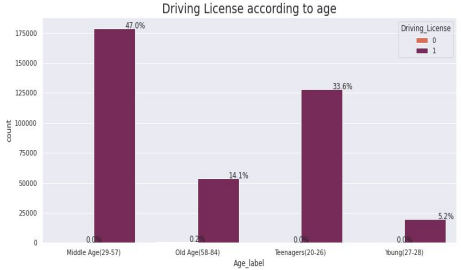


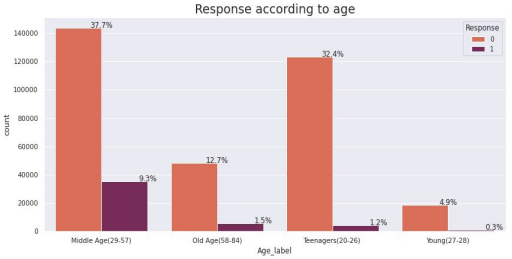
Age According to Response

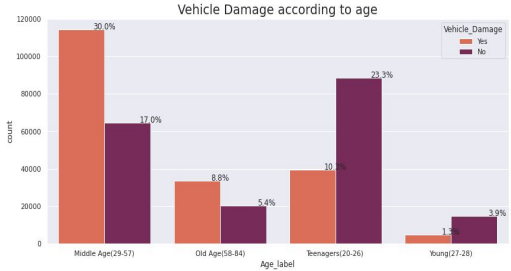
According to age wise:

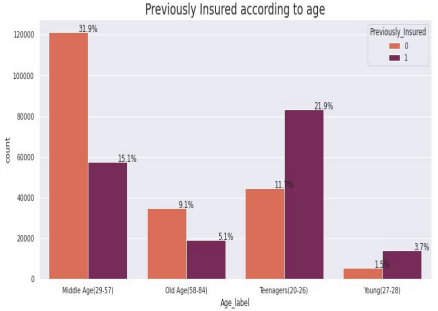
- **Around 21.9% of persons in their teens have insurance previously.**

So most teenagers have insurance and are aware of their policy. So the target audience might be middle-aged people and teenagers to generate more leads for insurance companies.





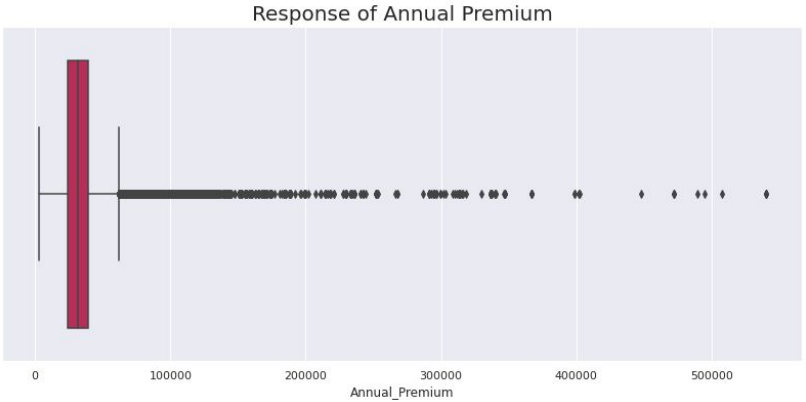




Outliers

Handling Outliers:

- **For identifying outliers, scatter plots and box plots are the most used visualisation techniques.Here we use boxplot.**
- **We have used the quantile approach to address outliers and eliminate Outlier.**



Feature encoding:

- We have encoded gender,vehicle damage and vehicle age which where of object type

Model Building

AI

- As we can see from the figure, target variable response is not balanced.
- ML techniques such as Decision Tree and Logistic Regression have a bias towards the majority class, and they tend to ignore the minority class. So solving this issue we use resampling technique.
- After Random Over Sampling Of Minor Class Total Samples are : 668798
- Original dataset shape Counter({0: 334399, 1: 46710})
- Resampled dataset shape Counter({1: 334399, 0: 334399})
- We split the dataset into train and test and performed scaling technique using standardscaler

Machine Learning Algorithms

AI

Let's try various machine learning models on our data set to see how they each perform.

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **KNN**
- **Gradient Boost**
- **XGBoost**
- **LightGBM**

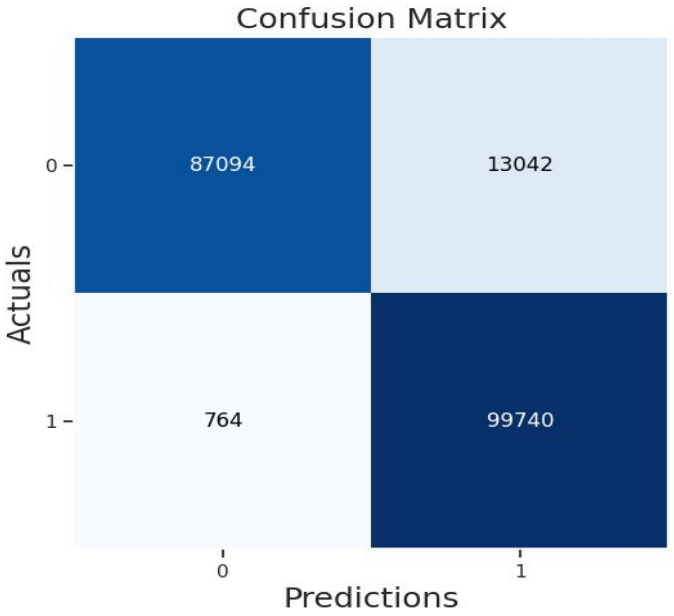
Machine Learning Algorithms

AI

- **Precision: Precision is the ratio of true positive predictions to the total number of positive predictions.**
- **Recall:Recall is the ratio of true positive predictions to the total number of actual positive cases.**
- **F1-Score: F1-Score is the harmonic mean of precision and recall.**
- **Support: The support is the number of samples of the true response that lie in that class.**
- **Confusion Matrix: The confusion matrix shows the number of true positive, false positive, true negative, and false negative predictions made by the model.**
- **ROC Curve:An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification threshold.**
- **Classification Report: A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification model.**

DecisionTree Classifier

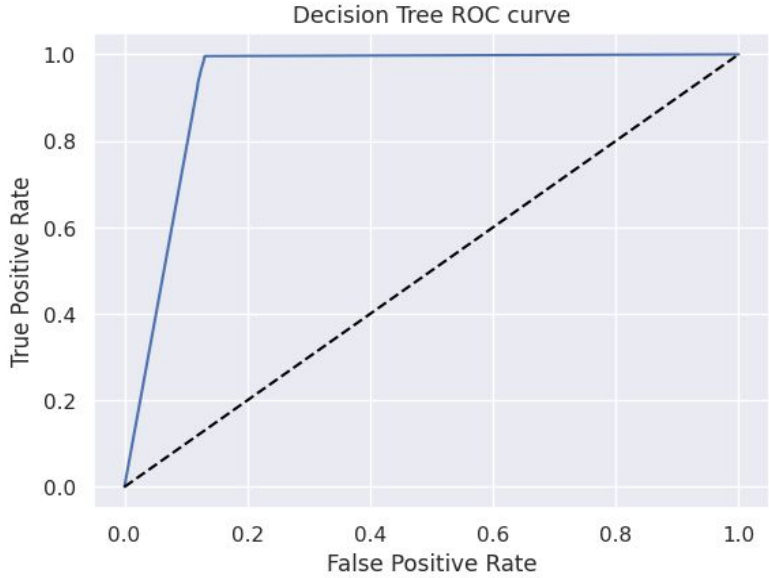
- **A decision tree is a model composed of a collection of "questions" organized hierarchically in the shape of a tree. The questions are usually called a condition, a split, or a test. We will use the term "condition" in this class. Each non-leaf node contains a condition, and each leaf node contains a prediction.**



DecisionTree Classifier

- **ROC Curve**
- **print(classification_report(dt_pred, y_test))**

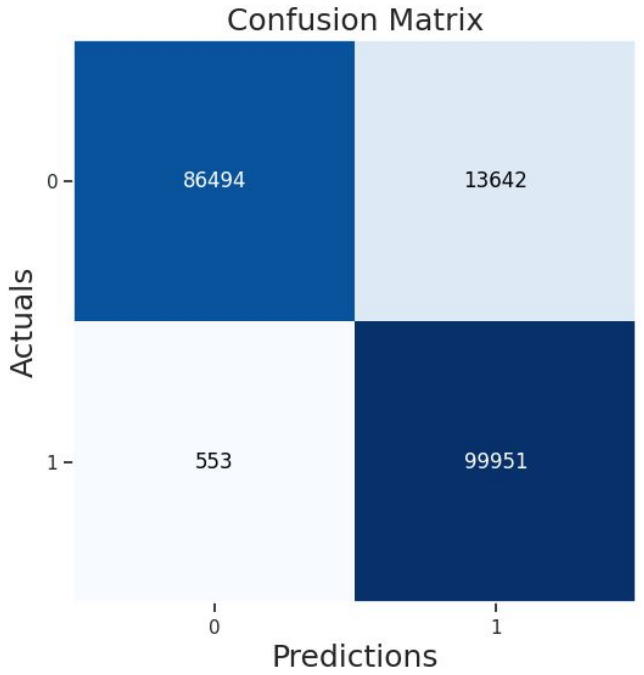
	precision	recall	f1-score	support
0	0.87	0.99	0.93	87858
1	0.99	0.88	0.94	112782
accuracy			0.93	200640
macro avg	0.93	0.94	0.93	200640
weighted avg	0.94	0.93	0.93	200640



Random Forest Classifier

AI

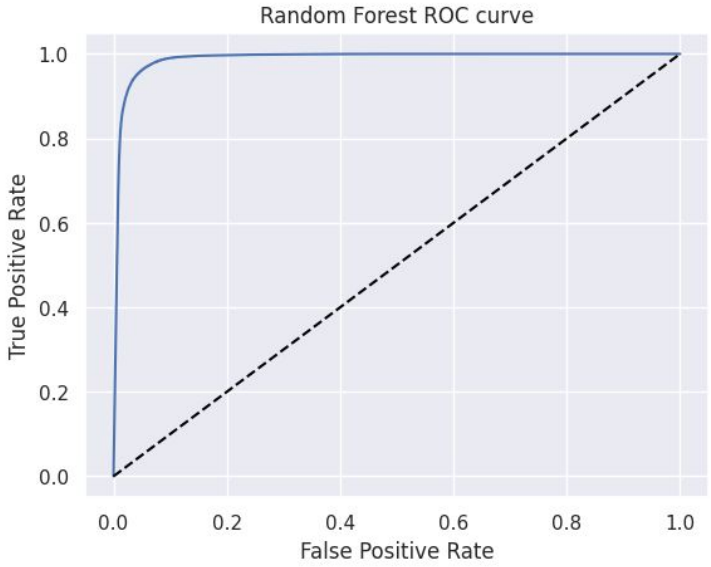
- **Random Forest is a robust machine learning algorithm that can be used for a variety of tasks including regression and classification. It is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, called estimators, which each produce their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction.**



Random Forest Classifier

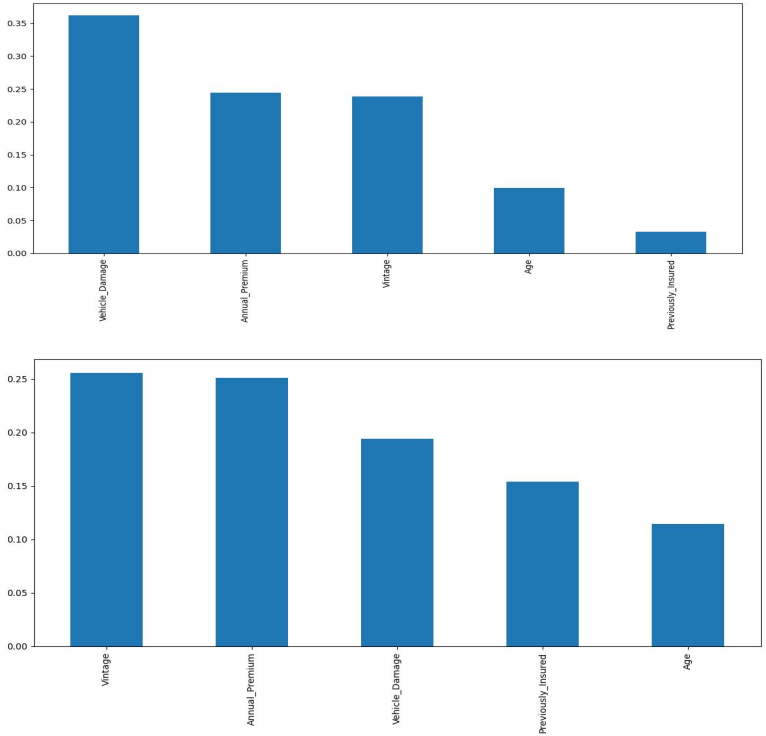
- **ROC Curve**
- **print(classification_report(rf_pred, y_test))**

	precision	recall	f1-score	support
0	0.86	0.99	0.92	87047
1	0.99	0.88	0.93	113593
accuracy			0.93	200640
macro avg	0.93	0.94	0.93	200640
weighted avg	0.94	0.93	0.93	200640



Important Feature

- **Vehicle damage and Annual premium are significant features seen in decision trees, while vintage and annual premium are significant features seen in random forests.**



Result of all models

- **Comparing the result of all the model Decision Tree, Random Forest gives the best Result.**
- **We can deploy this models into production.**

model_name	Recall_Score	Precision_Score	f1_Score	Accuracy_Score	ROC_AUC Score
Logistic Regression	0.976130	0.707201	0.820183	0.785601	0.834228
Decision Tree	0.992398	0.992398	0.935270	0.931190	0.931078
Random Forest	0.994498	0.879905	0.933698	0.929251	0.929132
KNN	0.965156	0.790859	0.869357	0.854695	0.854492
Gradient Boosting	0.927376	0.734615	0.819817	0.795803	0.795562
Extreme Gradient Boosting	0.932908	0.743262	0.827356	0.804974	0.804739
LGBM	0.927744	0.736864	0.821360	0.797852	0.797613

model_name	Recall_Score	Precision_Score	f1_Score	Accuracy_Score	ROC_AUC Score
Logistic Regression	0.976130	0.707201	0.820183	0.785601	0.834228
KNN	0.994508	0.798951	0.886068	0.871890	0.871665
LGBM	0.926282	0.736156	0.820347	0.796775	0.796537

Summary and Conclusion

AI

After loading our dataset, the first thing we did was look for duplicates and null values. There were no duplicates or null values, thus there was no need to treat them. With the aid of exploratory data analysis:

- **The gender variable in the dataset is spread nearly evenly. The male category is marginally larger than the female category, and the likelihood of purchasing insurance is also slightly higher. The response rate of those who are not interested in purchasing vehicle insurance is higher than that of those who are interested in buying vehicle insurance.Only 12.3% people are interested in buying vehicle insurance and 87.7% are not interested in buying vehicle insurance. So people who own a vehicle may already have vehicle insurance, or people might not be aware of insurance policy and pricing factors, which means the firm needs to come up with good marketing techniques and a pricing strategy to create awareness and offer an affordable price to the customers in order to reach out to more customers to generate more leads.**

Summary and Conclusion

AI

- The response rate of those who are not interested in purchasing vehicle insurance is higher than that of those who are not interested in buying vehicle insurance. 99.8% of customers have DL, whereas 0.2% do not have DL. Only a small percentage of people who have a DL (12.2%) are interested in purchasing vehicle insurance. So almost all the people who own vehicles have DL because it's mandatory when you have a bike, and only a small percentage of people are interested in buying vehicle insurance. The possible reason might be that people who own the bike may already have vehicle insurance or insurance might be expired.
- 45.8% of people are insured previously, in that 12.2% of people interested to buy the vehicle insurance again, Which means people are aware of insurance policy and ready to pay a premium amount, for better off taking actions to avoid certain risks or reduce risk. So buying insurance makes the most sense when the potential loss is great and there is a significant probability of loss.

Summary and Conclusion

AI

- Around 4.2% of vehicles are more than two years old, 52.6% are between one and two years old, and 43.2% are under one year old. 1.2% are interested in purchasing vehicle insurance for vehicles older than 2 years, 9.1% are interested in purchasing insurance for vehicles between 1 and 2 years old, and 1.9% are interested in purchasing insurance for vehicles older than 1 year. As vehicle age increases most of the people are aware of insurance and interested to buy the insurance for reducing the risk.
- 50.5% of the vehicles have past damage.12.0% of people who have had a damaged vehicle in the past want to acquire vehicle insurance. So 50 percent of vehicles are damaged and 50 percent are not damaged, which means people with damaged vehicles (12%) are interested in buying insurance and are aware of vehicle insurance policies and its benefits, while the rest of the people might already have purchased insurance and do not need to purchase again.

Summary and Conclusion

- Around 4.2% of vehicles are more than two years old, 52.6% are between one and two years old, and 43.2% are under one year old. 1.2% are interested in purchasing vehicle insurance for vehicles older than 2 years, 9.1% are interested in purchasing insurance for vehicles between 1 and 2 years old, and 1.9% are interested in purchasing insurance for vehicles older than 1 year. As vehicle age increases most of the people are aware of insurance and interested to buy the insurance for reducing the risk.
- The dataset has more individuals with an age of 24. 40 to 60-year-olds had a higher likelihood of purchasing vehicle insurance. 9.3% of people in their middle age are interested in purchasing insurance. Almost 47% of middle-aged individuals have a driver's licence. About 21.9% of people in their teens have health insurance. Around 21.9% of persons in their teens have had insurance previously. So most teenagers have insurance and are aware of their policy. So the target audience might be middle-aged people and teenagers to generate more leads for insurance companies.

Summary and Conclusion

AI

By using the interquartile range, we eliminated outliers and dealt with null data. We split the dataset into train and test splits after feature encoding three columns. Further, we applied 6 machine learning algorithms to see which customers might be interested in purchasing vehicle insurance and we also used hyperparameter tuning for three models to discover which model gives the best results. Vehicle damage and annual premium are the two most significant features seen in decision trees, while vintage and annual premium are seen in random forests. With 93% and 92% ROC AUC scores, Decision Tree and Random Forest outperform all other models.

Thank You