**Customer.csv** dataset is a binary classification problem. I have imported required libraries like **pandas, numpy, matplotlib, seaborn** and different library of **sciket-learn** for predictive model building and for **accuracy checking**.

Before model building, I have tried to do extensive **data-preprocessing** like finding the size of dataset using **shape**, getting information about data using **info()** method, checking the null value, filling null value using **fillna()** method

I have plotted graphical representation for some important feature to show how data is distributed in the model. I have also plotted heatmap to show graphically how two features are correlated to each other. For that I have imported two powerful library matplotlib and seaborn

I have done feature scaling. For that I have imported **StandaradScaler**

Based on instruction, I have tried to answer all questions

I have divided data in 80:20 ratio, 80 percent of data is for training purpose and 20 percent for testing purpose.

I have used different classification models and retrieved the performance **score** for each model to check their performance. Also done **cross\_validation\_test** to check **precision and accuracy** of model

LogisticRegression

DecisionTreeClassifier

RandomForestClassifier

KNeighborsClassifier

XGBoost =XGBClassifier

Types of Variables

Predictor

Age,driving\_experience,education,income,postal\_code,vehicle\_type,credit\_score,mileage,speeding\_violation, past\_accidents

Target

outcome

**Data Types**

Numeric

Credit\_score, postal\_code, mileage,speeding\_violations, past experience

Object

Age,gender,race,driving\_experience,income,education, vehicle\_year,vehicle\_type

Bool

Outcome,children,vehicle\_ownership,married

**Variable Category**

**Categorical**

Age,gender,race, driving\_experience,education,income,vehicle\_ownership,married,children,postal\_code,vehicle type, outcome

**Continuous**

credit\_score, mileage,speeding\_violations,past\_accidents

EDA findings

1. little more than 31 percent customer has claimed for insurance amount in last year
2. on an average, every customers with driving experience between 20- 29y have done speeding violations more than one time
3. Only 4.7(477) percent customer drive sports car, where male and female both equally represents. 32 percent customer from sports\_car has claimed for insurance and the proportion of male customer claiming insurance is higher than female. Male customers from sports car has also history of spped violation and past accidents. We can say that female can be potential customer base for the profit of company

1. 2818.44 is the standard deviation for annual mileage
2. There are particular customers type like
3. Customers using sedan is potential base
4. More than 80 % customers are literate
5. Female customers should be potential target by seeing their past history according to this database
6. Before 2015 customer is in large number, it means company need to add new customer base
7. Postal code 10238 has claimed more than 60 percent of total claim. So, there must be cautious steps before giving policy to that location customers. And company need to target other location which doing significantly good and having less number of claim.

**\*\*What do claimants have in common?\*\***

>past accidents> all have commited accidents in past

>vehicle ownership and postal code> most of the drivers claim vehicle ownership and most of them belongs to same postal code

>speeding violations> all have history of past accidents

How does the number of claims vary between postcodes?

10238 0.6940

32765 0.2456

92101 0.0484

21217 0.0120

Name: postal\_code, dtype: float64

Are there any problems with the data you have been given that should be kept in mind when modelling?

1. Some features have missing value data, missing value should be imputed for building model and for prediction
2. Many categorical features are there,so, need to perform hot encoding to convert data into numeric
3. Feature scaling need to be done on numeric dataset so that all value looks similar. I have done StandardScaler to get an appropriate training data
4. Some features are not useful, therefor it should be dropped out from final dataframe

Has the client collected the right data for their business needs?

What recommendations would you make to the client for future data collection?

1. Client need to gather more data
2. More features need to be added like premium amount, duration of insurance policy, city wise data collection or choose large area rather chosing postal\_code

**Wayforward**

Company need to increase new customer base

They need to focus small cars rather sports\_car

They should target literate customer base

Need to add more features in data collection