Business Objective:

This is a classification task, where we need to predict if the customer contacted during campaign is going to buy the Car Insurance or not.

Approach:

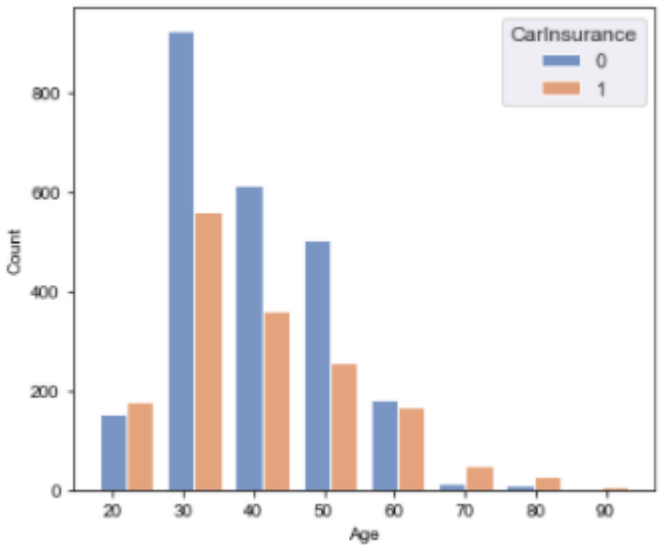
I have followed CRISP-DM approach, which is an Industry standard and easy to understand the flow, to tackle this task.

Business Understanding is clear from the objective, now we will focus on the other steps.

Data Understanding:

Exploratory Data Analysis (EDA) plays an important role to begin the analysis and understand the data. It helps in providing us with the stats and hidden insights of the data, which features are possibly going to have the impact on our target variable. In general, it gives the idea about the group of people which should be contacted to achieve the call success (assuming Customer buying Car Insurance is success).

For example: In the below graph, it can be inferred that people below 25yrs of age and more than 60yrs of age are high likely to buy Car Insurance. Similarly, there are other examples which are discussed in notebook in detail.



Data Preparation:

Once, we have understood about the data, we can start with the data preparation for modelling. There were missing values, where 'Outcome' column shown ~75% of missing data which I have removed. The 'Job' & 'Education' columns which had less number of missing data was replaced by mode value and 'Communication' columns have shown ~25% of missing data which was replaced by new label (as replacing it by mode value for this column can create the bias in data). There was the datetime data (CallStart & CallEnd), from which I have calculated the CallDuration to make those two features in useful form.

As the class labels for the data were in ratio of 40-60, which is slight imbalance in data. In real world, it can be assumed as closed to balanced data. So I haven't used any sampling techniques.

For Machine Learning modelling, we require the data to be in numerical format. I have used one hot encoding to replace the text label fields. Initially, I have used all the features as it is and after analysing the results from first round of model training, I saw that some of the continuous features (which should be significant from EDA) are not showing significant impact on the target variable. So I have made those type of records in categorical format by binning method.

Modelling:

As the majority of features are categorical variable, so ideally Tree & Ensemble based algorithms should perform better in these type of data. But for comparison purposes, I have used three types of algorithms.

- Probability Based Algorithms
- Distance Based Algorithms
- Tree & Ensemble Based Algorithms

And as expected, Tree & Ensemble based methods performed well than the other algorithms.

After, further feature engineering the conversion of continuous variable into categorical format, have further improved the accuracy of the model.

Evaluation:

As the data was almost balanced, I have preferred to use accuracy score to measure the performance of the models. Further reviewing the classification report-

	precision	recall	f1-score	support
0	0.90	0.87	0.88	484
1	0.81	0.85	0.83	316
accuracy			0.86	800
macro avg	0.85	0.86	0.86	800
weighted avg	0.86	0.86	0.86	800

We can say that 81% of time the model is precisely predicting the '1' whereas 85% of the time our prediction for '1' is correct. Depending on the cost associated with each event, we can focus on the Precision or Recall of the model while evaluating and comparing with others.

I have also used 10-Fold Cross Validation, to identify the model performance for all the subset of data and how it will behave for future data. And we can see that accuracy for the model is 85% with standard deviation of 0.015.