Sahaj software | inspiring brilliance

Problem Statement

A key challenge for the insurance industry is to charge each customer an appropriate price for the risk each customer represents. As is widely known in the insurance industry, the coverage risk varies widely from customer to customer. A deep understanding of different risk factors helps predict the likelihood and cost of potential insurance claims.

An Insurance company wants to take your help in identifying whether a customer will claim for the medical insurance or not and if he/she claims for insurance, predict the claim amount.

In this context, you'll work with the data provided by the company to solve the above mentioned problem. As the information is finance related, to maintain the confidentiality, the feature names and the values provided are masked. It is up to you how you would want to treat these features.

Data Set

You are provided with two csv files - “traindata.csv” and “testdata.csv”. The “traindata.csv” has the target variable (whether the customer is satisfied or not) and the “testdata.csv” does not have a target.

Evaluation

As specified in the problem statement, this would be a two-stage problem. **In Stage-1, you would be predicting whether a customer claims for insurance or not. In Stage-2, you would be predicting the insurance amount for those customers who were predicted to be claiming insurance.**

For Stage-1: We aim at better F1 statistic

For Stage-2: We aim for lower MAPE values

Step1: **Visualizations**

Since this forms an important aspect in data science problems, we would want you to use visualizations to obtain any insights from the data that could be a value-add to the company.

Extensive Exploratory Data Analysis (EDA), statistical analysis with appropriate reasoning is expected.

Step2: **Benchmark F1 statistic and MAPE**

We will evaluate your predictions submission file with the actuals that we have at our end for F1 score and MAPE. The benchmark for F1 is 40% on minority class and MAPE is 10%. Please note that the values mentioned as benchmark are quite liberal and these values should not refrain you from making attempts to further improve your predictions. The higher F1 statistic and lower MAPE are always better.