Project Report

Task: To find potential customers for different services provided by the company using past data and provide possible changes for sales conversions.

Project:

The data provided contained the list of the Customer IDs along with the number of services bought (divided into 9 categories from P1 to P9), ratios of services being used, amount of services assigned, customer size, segment (S_1 to S_5), time since purchase and few unknow variables(Var1, Var2, etc.). It contained approx. 6000 observations spread across 32 columns. The company could charge money only for the number of services being used and hence, to increase revenue, we had to find customers which could use more amount of services based on different factors present in the dataset. Firstly, we performed the cleaning of the data to remove inconsistencies like ratios being more than 1 or less than 0 and removing missing values data. This was followed by performing time analysis i.e. the time since the purchase of the data so that it may be used as a feature for future purposes. In order to predict the potential customers, the following approach was used: The data was divided into categories(0 and 1) based on the percentage of the services used where '1' meant the customer was using services above a certain value (ex: more than 0.45 quantile or above 0.57 ratio for P_1 services) and then the data were classified using various classification algorithms into 0 and 1 which were used as "Actual Values".

Before performing the classification, the observations which were present in the bottom 10% with respect to the number of services used were removed since they were assumed to be useless targets for conversion and hence, would only increase the noise in the data. For example: for P_1 Ratio, after dividing the data into '0' and '1' categories, we used 4 classification algorithms: Logistic Regression, Decision Trees, Random Forest Classifier, and XGboost Classifier in order to classify the data into '0' and '1' categories. This classification was based on all the features including ratios of services used, time since purchase, the number of services assigned, the number of services bought, Customer size, Segment except for the P1_Ratio since it was already used to divide the data into the categories. Each classification algorithm consisted of 4 iterations: The test data in 1st iteration was removed and training set was used as total data for 2nd iteration and similarly, the test data in 2nd iteration was removed and training set was used as the total dataset for the third iteration and for 4th iteration, the whole dataset which was given the actual value as '0' was used as the test data. So, in each iteration, the data which had actual value as '0' had at least been tested once (i.e. in the 4th iteration) and maximum of twice (i.e. if present in 1st, 2nd or 3rd dataset as test data). Continuing this across the 4 classification algorithms, each observation with value '0' had at least been tested 4 times (once in 4th iteration of each algorithm) and a maximum of 8 times. These datasets were merged consisting of the predicted values for all the data for the 16 iterations (4 for 4 algorithms). The observations which had actual value as '0' but had been predicted as '1' in a given number of iterations (ex: 3 for P_1 Ratios) were assumed to be potential targets i.e. they were assumed to be customers which were at present in the '0' category but had the potential of moving into the '1' category and hence could be targeted for possible changes.

The next step was to predict the possible changes for these observations which was done using Regression. The K-Nearest Neighbors Regression algorithm was used to predict the number of changes possible in these observations by finding the 5 closest observations in the ‘1’ data and taking their mean value as the predicted ratio possible for them. This gave us the possible value for P_1 Ratio for these observations which was closest to their current value but would ensure that they move up to the ‘1’ category. The difference in the actual ratios and the predicted ratios (i.e. predicted by the KNN model) was the possible change for the observations which when multiplied by the individual’s Customer Size gave the number of services that could be used more by the customer and hence add to the company’s revenue.

Thank You for this opportunity.

Team Members:

Sanchit Goyal, Ankit Tiwari, Kasina Jyoti Swaroop, Astitva, Aditya, Sachin S Singh, Shubham Jain