**Business Analytics for Strategic Decision Making**

**Project : Telecom Customer Churn**

**Problem Statement:**

One of the key industries in developed nations is now telecommunications.

The technical progress and the increasing number of operators raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies.

To increase revenue, three key strategies have been put forward:

1) Acquire new customers

2) Upsell the existing customers

3) Increase the retention period of customers.

The third strategy is the most profitable one, which keeps the existing customer costs much less than finding a new one. It is also much simpler than the upselling strategy, according to a comparison of these strategies that took the value of return on investment of each into account. To implement the third strategy, businesses must reduce the risk of customer churn, also referred to as "the customer movement from one provider to another”.

**Objective:**

In service industries with intense competition, customer churn is a major concern. Predicting which customers are likely to leave the company is a potentially large added revenue source if done early on. After much research, it is confirmed that machine learning technology is highly efficient to predict this situation.

**Dataset Description:**

1. Internet

|  |  |
| --- | --- |
| **Variables** | **Description** |
| Customer ID | A unique ID that finds each customer |
| Multiple Lines | Indicates if the customer subscribes to multiple telephone lines with the company: Yes, no |
| Internet Service | Indicates if the customer subscribes to Internet service with the company: no, DSL, Fiber Optic,Cable |
| Online Security | Indicates if the customer subscribes to an added online security service provided by the company: Yes, no |
| Online Backup | Indicates if the customer subscribes to an added online backup service provided by the company: Yes, no |
| Device Protection Plan | Indicates if the customer subscribes to an added device protection plan for their Internet equipment provided by the company: Yes, no |
| Premium Tech Support | Indicates if the customer subscribes to an added technical support plan from the company with reduced wait times: Yes, no |
| Streaming TV | Indicates if the customer uses their Internet service to stream television programming from a third-party provider: Yes, no |
| Streaming Movies | Indicates if the customer uses their Internet service to stream movies from a third-party provider: Yes, no |

1. Customer

|  |  |
| --- | --- |
| **Variables** | **Description** |
| Customer ID | A unique ID that finds each customer |
| Gender | Male or female |
| Senior Citizen | 1: Yes, 0: No |
| Partner | Yes, no |
| Dependents | Yes, no |

1. Churn

|  |  |
| --- | --- |
| **Variables** | **Description** |
| Phone Service | Indicates if the customer subscribes to home phone service with the company: Yes, no |
| Tenure | Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above |
| Contract | Indicates the customer’s current contract type: Month-to-Month, One Year, or Two Years |
| Paperless Billing | Indicates if the customer has chosen paperless billing: Yes, no |
| Payment Method | Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, or Mailed Check |
| Monthly Charge | Indicates the customer’s current total monthly charge for all their services from the company |
| Total Charges | Indicates the customer’s total charges, calculated to the end of the quarter specified above |
| Churn | Yes,no |

**Note:**

* Download the “internet.csv”, “customer.csv”, and “churn.csv” using the link given in the telecom customer churn project statement.

**Steps to perform:**

1. Import the data files
2. Check the structure of the data files and show the common column
3. Create combined data by merging the three datafiles
4. Do a preliminary statistical summary of this combined data
5. Identify all categorical columns and convert them to factor type
6. Check for missing and trivial values
7. Remove the rows with missing values
8. Check for outliers in the numerical variables using a box plot and if any, remove them
9. Study the overall distribution of the Churn variable
10. Study the distribution of categorical variables
11. Remove unwanted variables, like Customer ID
12. In the variables related to internet service, there are three categories: Yes,No, and No Internet Service. Transform them to binary variables with **No** as no and No internet service and **Yes** otherwise. Do the same with the variable “Multiline”
13. Follow the pointers below for model building and prediction:

a. Divide the data into train and test as 80 – 2

b. Use the train data to develop a logistic regression model with cross-validation

c. Check the summary of the Logit model and calculate the marginal effects of the dependent variables

d. Evaluate the model using a test dataset with a confusion matrix and ROCR curve and discuss the results

e. Plot the fitted values from logit regression and the linear probability model and compare the results. Discuss the dependent and independent variables for regression and their relationship

f. Compare the model performance of the logit model and linear probability model with classification accuracy using threshold values of zero