CodeClause

Allocated Project

Task 1 - Churn Prediction in Telecom Industry using Logistic Regression

Inference

Introduction

Customer Churn Prediction

Customer attrition or churn, is when customers stop doing business with a company. It can have a significant impact on a company's revenue and it's crucial for businesses to find out the reasons why customers are leaving and take steps to reduce the number of customers leaving. One way to do this is by identifying customer segments that are at risk of leaving, and implementing retention strategies to keep them. Also, by using data and machine learning techniques, companies can predict which customers are likely to leave in the future and take actions to keep them before they decide to leave.

We are going to build a basic model for predicting customer churn using the Telecommunication Customer Churn dataset. We are using some classification algorithm to model customers who have left, using Python tools such as pandas for data manipulation and matplotlib for visualizations.

Let's get started.

Steps Involved to Predict Customer Churn

- Importing Libraries
- Loading Dataset
- Exploratory Data Analysis
- Outliers using IQR method
- Cleaning and Transforming Data
    - One-hot Encoding
    - Rearranging Columns
    - Feature Scaling
    - Feature Selection
- Prediction using Logistic Regression

We have 2 types of features in the dataset: categorical (two or more values and without any order) and numerical. Most of the feature names are self-explanatory, except for:

- Married: Whether the customer has a partner or not (Yes, No).
- Tenure: Number of months the customer has stayed with the company.
- MonthlyCharges: The amount charged to the customer monthly.
- TotalCharges: The total amount charged to the customer.

There are 126 customers in the dataset and 15 features without Customer's ID (non-informative) and Churn column (target variable). Most of the categorical features have 4 or less unique values.

Feature distribution

We plot distributions for numerical and categorical features to check for outliers and compare feature distributions with target variables.

Numerical features distribution

Numeric summarizing techniques (mean, standard deviation, etc.) don't show us spikes, shapes of distributions and it is hard to observe outliers with it. That is the reason we use histograms.

When we look at distributions of numerical features in relation to the target variable. We can observe that the greater TotalCharges and tenure are the less is the probability of churn.

Categorical feature distribution

To analyze categorical features, we use bar charts. We observe that Senior citizens and customers without phone service are less represented in the data.

The next step is to look at categorical features in relation to the target variable. We do this only for contract features. Users who have a month-to-month contract are more likely to churn than users with long term contracts.

Target variable distribution

Target variable distribution shows that we are dealing with an imbalanced problem as there are many more non-churned as compared to churned users. The model would achieve high accuracy as it would mostly predict the majority class - users who didn't churn in our example.

Few things we can do to minimize the influence of imbalanced dataset:

- Resample data,
- Collect more samples,
- Use precision and recall as accuracy metrics.

Outliers Analysis with IQR Method

No outliers in tenure

No outliers in MonthlyCharges

Insights from the bar plot of correlation matrix of churn

HIGH Churn can be seen in the case of payment methods(electronic check),internet service(fiber optic),paperless billing,monthly charges and streaming movies.

LOW Churn is seen in the case of tenure,one year contracts,long-term contracts,total charges and tech support.

Factors such as gender, phone service availability, and streaming T.V. have almost NO impact on churn.