Telecom-churn

Problem Statement:

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, customer retention has now become even more important than customer acquisition.

For many incumbent operators, retaining high profitable customers is the number one business goal. To reduce customer churn, telecom companies need to predict which customers are at high risk of churn. In this project, you will analyze customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn, and identify the main indicators of churn.

In this competition, your goal is to build a machine learning model that is able to predict churning customers based on the features provided for their usage.

Customer behaviour during churn:

Customers usually do not decide to switch to another competitor instantly, but rather over a period of time (this is especially applicable to high-value customers). In churn prediction, we assume that there are three phases of customer lifecycle :

The ‘good’ phase: In this phase, the customer is happy with the service and behaves as usual.

The ‘action’ phase: The customer experience starts to sore in this phase, for e.g. he/she gets a compelling offer from a competitor, faces unjust charges, becomes unhappy with service quality etc. In this phase, the customer usually shows different behaviour than the ‘good’ months. It is crucial to identify high-churn-risk customers in this phase, since some corrective actions can be taken at this point (such as matching the competitor’s offer/improving the service quality etc.)

The ‘churn’ phase: In this phase, the customer is said to have churned. In this case, since you are working over a four-month window, the first two months are the ‘good’ phase, the third month is the ‘action’ phase, while the fourth month (September) is the ‘churn’ phase.

Kaggle Submission

**Weightage - 20%**

Based on your Kaggle Submission, you'll be given an evaluation score and a final rank on the Public leaderboard. This will be used to compute your Kaggle Submission Score, which will be calculated out of 20% of your overall case study marks.

**Note: Make sure your accuracy is greater than the sample submission that is present in the leaderboard**

Internal Submission

**Weightage - 80%**

Your internal submission which will carry 80% is going to be evaluated based on the following rubrics.

|  |  |  |
| --- | --- | --- |
| **Stage** | **Meets expectations** | **Does not meet expectations** |
| Data understanding, preparation & pre-processing (10 %) | Has used imputation techniques to treat missing values    Has identified potentially useful and non-useful attributes using feature importance and impact estimation. | Has not used any imputation techniques to treat missing values  Has not used identified potentially useful attributes using feature importances |
| EDA and Feature Engineering (30 %) | Has Performed correlation analysis between variables and has used plots to identify relationships between variables    Has Performed advanced data analysis, including plotting heatmaps and histograms    Has Performed feature engineering i.e. one or more methods on attributes that lead to the creation of a new potentially useful variable e.g. day from the date    Has applied variable transformation which are methods to turn categorical variable into numerical data and scale transformations to numerical data | Has not performed correlation analysis    Has not performed advanced data analysis Has not performed feature engineering    Has not applied variable transformation |
| Model selection, model building, evaluation & prediction (35 %) | Has tried more than one model for training and evaluation purposes on respective metric    Has used cross-validation and hyperparameter tuning for each model        Choosing the best model based on the fit of the dataset and output variable. The relevant evaluation metrics have been calculated and interpreted correctly as per the business context | Has not tried more than one model for the case study      Has not used cross-validation and hyperparameter tuning for each model      An incorrect model has been chosen at the end with wrong and incomplete inferences about the evaluation metrics. |
| Code readability and conciseness (5%) | The code is well commented and the text is written in detail to explain the thought process.    Efficient, concise code is written. | The code is not commented well / text is not written in detail.    Inefficient/verbose code is written. |