<https://www.kaggle.com/competitions/telecom-churn-case-study-hackathon-c58/overview>

**Overview**

**Description**

**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.

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.

**Evaluation**

**Goal**

It is your job to predict if a customer will churn, given the ~170 columns containing customer behavior, usage patterns, payment patterns, and other features that might be relevant. Your target variable is "churn\_probability"

**Metric**

Submissions are evaluated on [Classification Accuracy](https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score) between the value of the predicted value and the actual value of churn for each of the customers.



The public leaderboard is going to rank your submission against other users while the competition is active, however, once the competition is ended, the final ranks will be calculated on the private leaderboard.

**Submission file format**

The file should contain a header and have the following format (CSV):

id,churn\_probability

70005,0.0

70006,1.0

70007,0.0

etc.

A sample file can be found attached in the Data section of this competition.

**FAQs**

**How do I contact support?**

For this competition, any questions, concerns, or technical difficulties will be addressed at the Discussion forum on UpGrad Learn platform. Any discussion posted regarding support on this competition page will have to be redirected to the same portal.

**What’s the difference between a private and public leaderboard?**

The Kaggle leaderboard has a public and private component to prevent participants from “overfitting” to the leaderboard. If your model is “overfit” to a dataset then it is not generalizable outside of the dataset you trained it on. This means that your model would have low accuracy on another sample of data taken from a similar dataset.

**Public Leaderboard:** For all participants, the same 70% of predictions from the test set are assigned to the public leaderboard. The score you see on the public leaderboard reflects your model’s accuracy on this portion of the test set.

**Private Leaderboard:** The other 30% of predictions from the test set are assigned to the private leaderboard. The private leaderboard is not visible to participants until the competition has concluded. At the end of a competition, we will reveal the private leaderboard so you can see your score on the other 30% of the test data. The scores on the private leaderboard are used to determine the competition winners.

**What are kernels?**

Kaggle Kernels is a cloud computational environment that enables reproducible and collaborative analysis. Kernels support scripts in R and Python, Jupyter Notebooks, and RMarkdown reports. Go to the Kernels tab to view all of the publicly shared code on this competition.

**Data**

**Dataset Description**

**File descriptions:**

* **train.csv**: Contains 172 columns. The primary key that represents each customer is id. The target variable that you need to predict is churn\_probability which contains a value of 0 or 1. This data is what you are going to use for EDA, cleaning, feature engineering, model building, model evaluation, model selection, and finally model training.
* **test.csv**: Contains 171 columns, doesn't contain the target variable churn\_probability. You will apply all the necessary preprocessing steps to get this data into the right format and then use the model trained using the *train.csv* file to make predictions with this. This is unseen data! Only the competition hosts know the actual values of the target feature for this data and therefore your submissions will be evaluated on how well your model does with this dataset.
* **sample.csv**: This contains the format in which you need to submit the solutions to Kaggle. The id column in this dataset exactly the same as the id column in *test.csv*. You will make your predictions on the *test.csv* data and store them in a submission file that has the same format as this file. Check the **Overview>Evaluation** tab for mode details.
* **data\_dictionary.csv**: This contains the definitions for the various acronyms that you will need to understand each variable. For example, the variable total\_og\_mou\_7, contains the acronyms total, og, mou, and 7, which can be translated as the total outgoing minutes of voice calls made by the user in month of July.

**Data Definitions**

The definitions are also listed down below:

* CIRCLE\_ID : Telecom circle area to which the customer belongs to
* LOC : Local calls - within same telecom circle
* STD : STD calls - outside the calling circle
* IC : Incoming calls
* OG : Outgoing calls
* T2T : Operator T to T, i.e. within same operator (mobile to mobile)
* T2M : Operator T to other operator mobile
* T2O : Operator T to other operator fixed line
* T2F : Operator T to fixed lines of T
* T2C : Operator T to it’s own call center
* ARPU : Average revenue per user
* MOU : Minutes of usage - voice calls
* AON : Age on network - number of days the customer is using the operator T network
* ONNET : All kind of calls within the same operator network
* OFFNET : All kind of calls outside the operator T network
* ROAM : Indicates that customer is in roaming zone during the call
* SPL : Special calls
* ISD : ISD calls
* RECH : Recharge
* NUM : Number
* AMT : Amount in local currency
* MAX : Maximum
* DATA : Mobile internet
* 3G : 3G network
* AV : Average
* VOL : Mobile internet usage volume (in MB)
* 2G : 2G network
* PCK : Prepaid service schemes called - PACKS
* NIGHT : Scheme to use during specific night hours only
* MONTHLY : Service schemes with validity equivalent to a month
* SACHET : Service schemes with validity smaller than a month
* \*.6 : KPI for the month of June
* \*.7 : KPI for the month of July
* \*.8 : KPI for the month of August
* FB\_USER : Service scheme to avail services of Facebook and similar social networking sites
* VBC : Volume based cost - when no specific scheme is not purchased and paid as per usage