ISYE 6740 – Spring 2023
Project Proposal

Team Member Names: Wenyu Sui

Project Title: Telecom Customer Churn Prediction

I. Problem Statement

The industry of telecommunications is competitive. In this industry, customers are able to make their selections between different service providers. They may actively switch from one phone carrier to another because of price differences, service qualities and other factors. Typically, phone carriers experience an average annual customer churn rate of 15-25%. [1] In 2020, this number was 21%. [2] Since acquiring a new customer normally costs 5-10 times more than retaining an existing customer [1], it is extremely important for phone carriers to improve their client retention rate.

One important issue involved in retaining the customers is to detect those customers who are likely to change service providers in the near future. The phone carriers could reach out to these customers and retain them by offering special discounts or improving services based on their feedback, if the phone carriers could predict that these customers would churn.

The task of this project is to develop a machine learning model which will help phone carriers to predict their customer churns. The model will take each customer's temporal behavioral data as input and predict if the customer will churn in the future. We will develop a few models using different algorithms, use cross validations to measure their performance, and select one model that performs best.

II. Data Source

The dataset used in this project is obtained from the following Kaggle competition:
https://www.kaggle.com/competitions/telecom-churn-case-study-hackathon-c52/overview
We are going to use the train.csv dataset to perform our analysis. [1]

This dataset contains the temporal behavioral data of telecom customers that we can use to predict if a customer will churn. It includes 172 columns in total. The target variable $churn\_probability$ contains a value of 0 or 1, where 0 represents "not churn" and 1 represents "churn" – this value is what our machine learning model needs to predict.

The following data definition table contains the various abbreviations that we need in order to understand the meaning of each variable. For example, the column $onnet\_mou\_8$ contains "*onnet*", "*mou*" and "*8*". Therefore, it represents the minutes of usage of voice calls within the same operator network in August.

Acronyms	Description
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 mobiles
T2O	Operator T to other operator fixed line
T2F	Operator T to fixed lines of T
T2C	Operator T to its 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	G network
AV	Average
VOL	Mobile internet usage volume in MB
2G	G 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

III. Proposed Methodology

1. Preprocessing

Here are the steps that we plan to take to process the data:

1. In the original dataset, some columns are meaningless for analysis. These fields need to be removed from the data. For example, the columns $date\_of\_last\_rech\_6$, $date\_of\_last\_rech\_7$ and $date\_of\_last\_rech\_8$ simply indicate the date of last day in each month. These columns need to be removed since they don't provide any useful information about the sampled customers.

2. Remove / impute columns with missing values. Some columns in the dataset may include a few missing values while some others may include a lot. For each column, if the missing values are less than 15% of the total samples, we plan to impute the missing values by using the mean value of the rest of the data in the same column. If the missing values are more than 15%, the column will be removed.

3. Split training and testing data. Since the test dataset provided by the original data source [1] doesn't include the target variable $churn\_probability$, we need to create our own test dataset. Here we are going to use 80% of the train data provided by the original data source as our training dataset, and the rest 20% as our testing dataset. The data will be split by using the function $model\_selection.train\_test\_split$ from Python package $sklearn$. [3]

4. Principal component analysis. The dataset contains 172 columns in total, and many of the columns are correlated to each other. Therefore, we may use principal component analysis to reduce dimensions of the data and eliminate collinearity.

2. Classification Algorithms

This project tries to resolve a classification problem. Therefore, we are going to implement some of the following models and compare their performance: Logistic Regression, Neural Network, Random Forest, Support Vector Machine, Boosting Algorithm, etc.

One of the challenges is that the original dataset is very imbalanced. It has 70,000 samples in total, but only 7,132 of them have 1 in column $churn\_probability$ and all the others have 0. We need to be cautious since imbalanced dataset may cause poor performance for some of the algorithms mentioned above. Some steps may need to be taken to correct the issues caused by the imbalanced dataset. For example, we may consider implementing a rare event correction when we implement Logistic Regression. [4] [5]

3. Performance Evaluation

To avoid any issue caused by overfitting, we will use K-fold cross validation to train the models and compare their performance. We will use Classification Accuracy as the most important performance metric. The formula is shown as below:

$$Classification\ Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

Another metric that we may consider is Sensitivity, which is defined below:

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

In this problem, we are more concerned about finding the customers who will churn, so we would like the model to correctly identify as many churned customers as possible. Therefore, it's worthwhile to be a little "conservative" and let the model generate more positive predictions (indicating the customer will churn), even though we may also have more false positive predictions and lower classification accuracy. Another reason why we need to consider Sensitivity is that our dataset is imbalanced, with too many zeros (not churn) and too few ones (churn) in column $churn\_probability$. In this situation, a model can easily achieve a very high Classification Accuracy (but very low Sensitivity) in cross validation by predicting all the input datapoints to be 0.

Finally, the model that achieves the highest classification accuracy and a reasonably high sensitivity in the K-fold cross validation will be selected as our final model. We will then use the explanatory variables from the testing dataset as inputs and make predictions using this model. The performance of the final model will be measured by the $Classification\ Accuracy$ calculated from the prediction results using the testing dataset.

References

[1] T. Jain, "Telecom Churn Case Study Hackathon - Predict churning customers for a Telecom company based on temporal behaviour," Upgrad, 12 Sep 2023. [Online]. Available: https://www.kaggle.com/competitions/telecom-churn-case-study-hackathon-c52/overview. [Accessed 17 Sep 2023].

[2] Statista Research Department, "Customer churn rate in the United States in 2020, by industry," statista, 6 Jul 2022. [Online]. Available: https://www.statista.com/statistics/816735/customer-churn-rate-by-industry-us/. [Accessed 17 Sep 2023].

[3] scikitlearn, "sklearn.model_selection.train_test_split," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html. [Accessed 14 Oct 2023].

[4] amoeba, "Does an unbalanced sample matter when doing logistic regression?," 6 Nov 2018. [Online]. Available: https://stats.stackexchange.com/questions/6067/does-an-unbalanced-sample-matter-when-doing-logistic-regression. [Accessed 17 Sep 2023].

[5] G. K. a. L. Zeng, "Logistic Regression in Rare Events Data," *Political Analysis,* no. 9, p. 137–163, 2001.