

# **Telecom Customer Churn Prediction**

## **Project Proposal Report**

Team 5

**Student Name:** Jaamie Maarsh Joy Martin

**Student Email:** [joymartin.j@northeastern.edu](mailto:joymartin.j@northeastern.edu)

**Percentage of Effort Contributed by Student:** 100%

**Signature of Student:** Jaamie Maarsh Joy Martin

**Submission Date:** 09/19/2024

## **Project Proposal**

As we all know that high attrition is a common aspect in every sector present in the world, and various studies show that the garnering new customers is five times costlier than retaining the old ones. The main idea of this project is to predict customer churn using a various ML model. Customer churn refers to when a customer might leave a service, and identifying this ahead of time allows businesses to take corrective action.

## **Background Section**

### **Bank Customer Churn Prediction based on Random Forest Algorithm - [Link](#)**

**Abstract:** The latest paper by Nini Zhang, Yuzhi Zheng and Chao Duan focuses on predicting bank customer churn using a Random Forest algorithm.

**Approach:** Random forest was taken as a base model to analyze customer churn prediction. Based on the results, are compared with the optimization results of several commonly used classification algorithms.

**Conclusion:** The random forest algorithm has better performance (speed and accuracy) compared to all other classification model. Even though Neural networks also had a high accuracy, since the amount of user data was too small, it was not an ideal in terms of effect.

### **Investigating Decision Tree in Churn Prediction with Class Imbalance - [Link](#)**

**Abstract:** In this paper by Bing Zhu, Guicai Xie, Yuan Yuan and Yiqin Duan, focuses on investigating the nature of decision tree algorithm on imbalanced datasets rather than the churning itself.

**Approach:** The approach was to perform various data level techniques such as pruning, un-pruning and sampling using various methods like SMOTE, ROS and much more.

**Conclusion:** It was found that suggestion sampling method is not actually independent, and their usage should be carefully studied for each task with domain-dependent measure. In the future, we will investigate some related issues on imbalanced datasets when decision tree is used as base model in the ensemble techniques

## **Customer Churn Prediction in Telecommunication Industry Using Machine Learning Classifiers - [Link](#)**

**Abstract:** In this paper by Nurul Izzati Mohammad, Saiful Adli Ismail, Mohd Nazri Kama, Othman Mohd Yusop and Azri Azmi, focuses on predicting telcom customer churn using various machine learning models.

**Approach:** The research team gets a fair idea of the models which has been used prior for this domain and then choose 3 of the models which they feel is the best, logistic regression, artificial neural network and random forest.

**Conclusion:** Based on the experimental result, it was found that every classifier produced good results with high accuracy, but logistic regression showed the best results, having a slight disadvantage in high computational time. The main factors causing the customer to churn was also zeroed-in – total charges, monthly contract and fiber optic internet service.

## **Predicting early user churn in a public digital weight loss intervention - [Link](#)**

**Abstract:** This paper from Robert Jakob, Nils Lepper, Elgar Fleisch, Tobias Kowatsch dealt with high churn rates related with DHIs, as many users disengaged before achieving desired outcomes.

**Approach:** There were three model dimensionality categories (LDM, MDM, HDM) created and was applied on each of the first seven days of app usage. All models inherently performed feature selection (e.g., through regularization), ranking features based on their importance in making predictions. The compatibility was evaluated using various evaluation metrics.

**Conclusion:** Even though, Random Forest (LDL) was the best performing model out of the lot but also had many false positives. In conclusion, the results indicated that churn prediction in DHIs required more prospective studies that were needed to validate the real-world applicability of these models for the prevention of user churn.

## Modeling and Analysis of Telecom User churn Warning

**Abstract:** This paper proposes a machine learning based method for customer churn prediction in Telecom customers.

**Approach:** The research team adds four new fields to the dataset to measure user stickiness (users joined in the last 1-2 months) and randomly chooses 3 of the models which they feel is the best, logistic regression, XGboost and random forest.

**Conclusion:** It was concluded that user retainability can be achieved to provide different service offerings or provide different benefits for different churn situations and the model gives a fair idea for distinct scenario.

### Project Plan and Timeline:

Below is the table format for the plan and milestone for the project:

Week	Task	Milestone
Week 3	Data Collection and Preprocessing	Dataset cleaned and ready for analysis
Week 4	Exploratory Data Analysis (EDA)	Key insights from the dataset
Week 5	Feature Engineering	Identified and engineered relevant features
Week 6	Model Selection	Chose ML models based on problem requirements
Week 7	Model Training	Initial model training completed
Week 8	Model Evaluation (Cross-Validation)	Evaluated models with validation data

Week	Task	Milestone
Week 9	Model Tuning	Hyperparameter tuning for model optimization
Week 10	Final Model Testing on Test Data	Final model tested with test data
Week 11	Performance Analysis	Detailed analysis of model performance metrics
Week 12	Refinement & Improvements	Improve model based on evaluation
Week 13	Final Report Preparation	Final project report drafted
Week 14	Final Presentation Preparation	Final presentation slides prepared