**Project Proposal - ISM6136.003F23 - Info Miners**

**Predicting Customer Churn for a Telecom Company**

**What is churning?**

Churning, in a business context, is when customers or subscribers discontinue using a company's offerings, leading to potential revenue loss.

**A person running through a door

Description automatically generated**

**Business Problem and Motivation:**

Over the past decade, the telecommunications industry has witnessed a rapid increase in the number of users. Global Telecommunication Market size was valued at USD 1805.61 billion in 2022 and is expected to grow to USD 3102.74 billion by 2030. However, in today's business landscape, churn has emerged as a significant challenge. For companies, particularly those operating in the telecom sector where acquiring a new customer can cost up to five times more than retaining an existing one, attrition of customers can result in substantial financial repercussions. and this high customer acquisition cost is a considerable challenge for businesses.

Reducing churn and retaining valuable customers is not only crucial for sustaining revenue but also for ensuring long-term profitability and business growth. If a company knows which customer is about to churn, they can proactively implement retention strategies, such as personalized offers, improved customer support, and targeted engagement to mitigate the financial impact of churn and secure their long-term profitability and business growth.

**Objective and Goals:**

The primary objective of this project is to develop a predictive model to effectively identify customers at risk of churning. Our key goals include:

* Building an accurate classification model to predict customer churn, enabling the company to proactively address potential churners.
* Conducting a comprehensive analysis of the dataset to reveal patterns and insights related to the factors and behaviors influencing customer to churn. This analysis will support the development of precise and focused retention strategies.

By achieving these goals, we aim to equip the telecom company with the knowledge needed to effectively identify and implement strategies to address the challenge of customer churn.

**Real World Impact**:

According to Forbes, the customer acquisition cost in 2020 was reported to be as high as $325. Retaining a customer, on the other hand, often costs significantly less. By focusing on customer retention through the churn prediction model, companies can improve cost savings.

In a scenario with 10,000 customers facing a risk of churning, the implementation of a churn prediction model can significantly enhance cost savings. The potential cost savings achievable through the utilization of this model are:

Cost of acquiring 10,000 customers = $325 x 10,000 = $3,250,000

Cost of retaining customers = $50 x 10,000 = $500,000

Cost Savings = Cost of acquiring new customers - Cost of retaining customers

Cost Savings = $3,250,000 - $500,000 = $2,750,000

This calculation illustrates the substantial cost savings associated with the churn prediction model. By retaining 10,000 customers, the company could save up to $2,750,000.

**Details of the Dataset:**

The dataset has been sourced from an Iranian telecom company's database over a 12-month period. The specific details about who collected the data and the time period are not public.

Link to the dataset - <https://archive.ics.uci.edu/dataset/563/iranian+churn+dataset>

The dataset comprises 3150 rows, each representing an individual customer, and includes 13 columns. These columns encompass various attributes, such as call failures, SMS frequency, complaint counts, distinct call numbers, subscription length, age group, charge amounts, service type, usage duration, customer status, frequency of use, and Customer Value.

All attributes, except for the Churn column, contain aggregated data from the initial nine months. The Churn column, serving as our target variable, reflects the customer's status at the end of the 12-month period, with a designated three-month interval for implementing retention strategies based on the behavior observed during the first 9 months.

**Previous Work:**

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014, November 1). *Improved churn prediction in telecommunication industry using data mining techniques*. Applied Soft Computing; Elsevier BV. https://doi.org/10.1016/j.asoc.2014.08.041

In this study, the authors investigated four prominent classification techniques using the dataset, revealing that the Artificial Neural Network (ANN) outperformed K-Nearest Neighbors (KNN), Decision Tree (DT), and Support Vector Machine (SVM), achieving recall accuracy exceeding 95%. However, it's worth noting that the study did not address the issue of data imbalance, which is a significant aspect to consider in real-world applications of classification techniques.

**Methodology and Approach:**

We will implement supervised classification algorithms discussed in class using Python in Jupyter notebooks, along with libraries including pandas, NumPy, and scikit-learn.

We will incorporate the following stages into our project implementation:

* Exploratory Data Analysis
* Preprocessing the dataset.
* Balancing the dataset if required.
* Training and testing a variety of classification methods.
* Choosing the best model based on a metric suitable for the business case.
* Hyper-Parameter tuning for the chosen model.

**Choice of Metric:**

The possible outcomes of predictions in this business case are -

* True Positive (TP): This represents instances where the model correctly predicts customers who are at risk of churning.
* True Negative (TN): This reflects instances where the model correctly predicts customers who are not at risk of churning.
* False Positive (FP): This occurs when the model incorrectly predicts customers as being at risk of churning when they are not. These are false alarms or instances where intervention may be unnecessary.
* False Negative (FN): This indicates situations where the model incorrectly predicts customers as not being at risk of churning when they are. This is a critical area of concern as it may lead to the loss of customers and future revenue for the company.

In this business case where customer churn is a significant concern, minimizing false negatives (FN) is more important than minimizing false positives (FP), therefore prioritizing recall as the metric is an appropriate choice. Maximizing recall ensures that the model is effective at identifying as many customers at risk of churning as possible. This approach aligns with the initial goal of retaining customers and reducing churn, as it places a strong emphasis on capturing all potential churn cases.

Considering the cost implications of false positives (FP) for the company (spending on promotion strategies for customers who are not likely to churn), we will also consider the F2 score along with recall in our evaluation, as it provides a weighted balance between precision and recall, giving higher importance to recall while still considering precision. This approach allows us to better address the cost of FP while prioritizing recall.

**Info Miners - ISM6136.003F23**

Prajay Reddy Minuka

Rakesh Chitrala

Kusuma Yalamarthi

Srija Thota