**Project Proposal**

**Predicting and Explaining Customer Churn in Telecommunications Using Machine Learning**

The telecom industry is very competitive, making customer retention a top priority. One big challenge is managing customer churn, which refers to customers leaving a service provider for a competitor. Churn significantly impacts profitability, as acquiring new customers is often far more expensive than retaining existing ones. This project aims to build machine learning models to predict customer churn using the publicly available **Telco Customer Churn dataset**. Our approach includes data preprocessing, feature engineering, and applying classification algorithms such as **Logistic Regression, Random Forest, and Decision Trees**. To address the dataset’s class imbalance, we will use a sampling method called SMOTE. Expected outcomes include a reliable churn prediction model, identification of key factors driving churn, and actionable insights that can help telecom companies improve customer loyalty.

**Research Questions**

This project will investigate the problem of customer churn by addressing the following questions:

1. Is it possible to develop a reliable model to predict high-risk churn customers?
2. What customer characteristics most strongly predict churn, based on insights from our predictive model?
3. How does our model's performance compare to the results reported in previous research by Wu et al. (2021)?

**Methodology**

Dataset

This project uses the Telco Customer Churn dataset, which is publicly available on Kaggle and was originally provided by IBM. The dataset contains 7,043 customer records and includes 21 attributes related to customer demographics, account details, and subscribed services. The target variable, **Churn**, is binary and tells us whether a customer has left the service provider. Our goal is to build a model that can identify patterns and customer characteristics that are linked to a higher risk of leaving.

Tools and Compute Resources

The analysis will be carried out using Python, either in a Jupyter Notebook or in Google Colab. Core libraries used in the project will include pandas for data manipulation, NumPy for numerical operations, and Matplotlib and seaborn for data visualization. For building and evaluating machine learning models, we will use scikit-learn. Based on the size of the dataset, we anticipate that standard personal laptops should be sufficient for most tasks. However, access to Carleton provided computing resources may be helpful, especially if model complexity or processing requirements increase during the project.

**Analysis Stages**

Our approach will involve three key stages:

Data Exploration and Preprocessing: We will begin with exploratory data analysis (EDA) to understand the distribution of variables and explore relationships between features and the target variable, *Churn*. This stage includes data cleaning, handling missing values, and preparing the dataset for modeling.

Feature Engineering and Selection**:** To improve model performance, we will perform feature engineering steps such as one-hot encoding to convert categorical variables into numerical form and standardization to scale numerical features. We will also explore feature selection techniques to reduce dimensionality.

Predictive Modeling and Evaluation**:** In the final stage, we will build and train multiple classification models to predict customer churn. Since churned customers represent a minority class, we will address the class imbalance using SMOTE (Synthetic Minority Over-sampling Technique). Model performance will be evaluated using metrics appropriate for imbalanced classification problems, including Precision, Recall, F1-Score, and ROC AUC. Model comparison will be done using 5-fold stratified cross-validation, implemented using pipelines to ensure that all preprocessing steps are applied correctly within each fold and to **avoid data leakage**.

**Expected Results**

By the end of this project, we expect to deliver the following outcomes:

A Predictive Model**:** A well-documented machine learning model capable of predicting customer churn, with a specific focus on identifying high-risk (likely-to-churn) customers. Because the dataset is imbalanced, we will evaluate model performance using metrics such as Precision, Recall, F1-Score, and ROC AUC -- rather than relying on accuracy alone.

Actionable Insights on Churn Drivers**:** A list and analysis of the most predictive features related to churn. These insights will highlight areas where the business could intervene, such as customers on specific contract types or service combinations.

A Comparison with Prior Research: We will compare our model's performance and findings with those presented in *Wu et al. (2021)*, who developed a telecom churn prediction model using machine learning techniques. This comparison will help validate our results and provide some context for evaluating the effectiveness of our approach.

**All documents and source code for this project will be made available at:**

https://github.com/ChurnProject/DATA-5000-Project/tree/main

**References**

blastchar. (2018). Telco Customer Churn [Data set]. Kaggle. https://www.kaggle.com/datasets/blastchar/telco-customer-churn

McKinsey & Company. (2021). *How telecom companies can win in the digital revolution*.

Wu, S., Yau, W., Ong, T., & Chin, C. (2021). *Integrated Churn Prediction and Customer Segmentation Framework for Telco Business.* IEEE Access. Retrieved from - <https://www.researchgate.net/publication/350930550_Integrated_Churn_Prediction_and_Customer_Segmentation_Framework_for_Telco_Business>