German International University of Applied Sciences
Informatics and Computer Science
Dr. Caroline Sabty
Eng. Nouran Khaled
Eng. Sandra Samuel

Machine Learning, Winter 2024
Project 2

1 Customer Churn Prediction Challenge

1.1 Overview

Welcome to the Customer Churn Prediction Challenge! This competition aims to predict customer churn based on various customer features, helping companies retain customers by identifying potential churn early. Participants will focus on building robust classification models using logistic regression, decision trees, and random forest to maximize predictive accuracy.

1.2 Competition Timeline

- **Team Registration:**
  All team members must enter using the invitation link (mentioned in links section) to be added to the competition.
- **Competition Deadlines:**
  - **Start Date:** 13th Of November.
  - **Final Submission Deadline:** 13th of December at 11:55 PM.

1.3 Team Registration

The same teams will continue from the previous project

1.4 Dataset Description

The dataset includes customer profiles and interactions, with features relevant to churn prediction such as demographics, account information, and customer behavior metrics. The dataset is divided into:

- **Training Set:** Contains labeled data where each customer is marked as churned or retained.
- **Test Set:** Contains the same features without the churn labels. Participants will submit their churn predictions for this dataset.

1.5 Objective

The objective is to build a classification model that maximizes the ROC AUC score, which evaluates the model's ability to discriminate between churned and retained customers. This metric provides an intuitive understanding of the model's performance in handling imbalanced data.

1.6 Evaluation Metric

The competition will be evaluated using ROC AUC (Receiver Operating Characteristic Area Under the Curve), calculated as:

- **ROC AUC:** Measures the modelâs ability to distinguish between churned and non-churned customers across different thresholds, ensuring relevance for real-world application in imbalanced datasets. **This is the metric for uploading your test set on the competition leaderboard.**
- **F1 Score:** Combines precision and recall to provide a balanced measure of the modelâs performance, particularly useful in cases with imbalanced classes. **You need to display the f1 score for all your models.**

1.7 Model Requirements

Participants are expected to experiment with the following models and compare their performances:

- **Logistic Regression:** To assess the baseline performance.
- **Decision Trees:** To evaluate interpretability in feature decision-making.
- **Random Forest:** To enhance predictive power through ensemble learning.

For each model, participants should provide a brief explanation of the chosen hyperparameters and why they were selected. Note: You can try additional models and use their results however you must include the models mentioned above.

1.8 Rules

- **External Data:** No external data is allowed; all models must be trained solely on the provided dataset.
- **Submissions:** Limit of 5 submissions per day to prevent leaderboard chasing.
- **Competition Submission:** All teams must submit at least one submission on Kaggle.
- **Notebook Submission:** All teams must submit the Notebook by maximum 13th of December to Form link mentioned in Links Section, *any late submission will not be considered.*

1.9 Links

Visit the competition page and access the dataset here: competition Link.

Submit the Code Notebook here: form Link