Marax AI
Churn Prediction Assignment

Problem Statement:

The objective of this assignment is to explore data, create data pipeline and build a model for churn prediction on a gaming app ‘TagPro’ using the play log data. TagPro is a casual CTF (capture-the-flag) game on a web browser. A churned user is someone who has stopped playing the game for some amount of time. In this assignment you need to extract relevant information, clean and pre-process it to create input features which can be used to model a machine learning algorithm.

Requirements and guidelines:

1. **Data exploration report**
    a. **Definition of churn:** Explore the play log data and come up with a problem definition for a machine learning model.
        - Formalise churn as a classification or a regression problem.
        - Explore possibilities of other churn definitions.

    b. **Indicators for churn prediction model:** Use the provided data to come up with features and preprocessing it for the model.
        - Analyze patterns in the data to identify key events which indicate churn.
        - Follow a reasonable method to extract features from the play log.
2. **Data processing pipeline:**
   Create a pipeline for data extraction and processing.
        - Input: List of unique user identifiers.
        - Output: User identifiers with their corresponding feature vectors.

   It should be in a form that output can be consumed by a machine learning model. Following are some of the modules that can be implemented:

    a. **Feature extraction:** Use the insights derived from explorations in the previous step to extract relevant features related to every user.

    b. **Data preprocessing:** Perform transformations on extracted features, preprocess and clean it.
3. **Prediction model:** Machine Learning model which is trained and tested on the given data to predict churn.
    - Explore possible model architectures based on time-series nature of data and problem definition. Use of RNN/LSTM can be a good strategy for handling time-series data.
    - Define proper metrics to compare models
    - Perform hyper-parameter tuning to get the best model and keep a track of model performance as you tune them.

Bonus:

- Implementing the pipeline using Distributed Computing tools like Spark will earn you extra scores.

Deliverables:

- **Data Exploration Report**
  Code and visualizations around the data analysis for definition of churn and feature extraction.

- **Data Processing Pipeline Code**
  Code involved with feature extraction, preprocessing and standardizing for specific model.

- **Model Code**
  Code for the model implemented and a method to reproduce the output.

- **Design Documentation**
  Design specification document, this must help us understand your approach in its entirety and contain any future improvements you intend to make

It is recommended to use Python as the programming language for the assignment, Jupyter Notebooks for presenting data analysis. Clean coding practices should be followed with proper comments and all the outputs should be reproducible.

Resources:
In order to assist you in completing the task we provide a link to detailed description and one of the approaches used to solve this problem.

- Paper https://doi.org/10.1371/journal.pone.0180735
- Data Description https://tagpro.eu/?science

It is recommended to build your own understanding of the problem and propose a better solution.

All the best!