Advanced Data Science Coding Challenge

Scenario

You are working for a company that has developed a new wellness app. The company wants to understand how certain features of the app and user characteristics are affecting the daily engagement time on the app (measured in minutes).

You have been provided with a dataset that contains the following features:

'day_of_week': Day of the week represented as an integer (0 - Monday, 1 - Tuesday, ..., 6 - Sunday).

'promotion_value': A daily promotional offer value. This is a number between 0 and 100 indicating the strength of the promotion.

'happiness_index': A self-reported user happiness index on a scale from 1 to 5, where 1 is very sad and 5 is very happy.

'daily_active_minutes': The target variable, i.e., daily active minutes on the app.

The data was collected over three years and your task is to build a model that predicts the daily active minutes based on the day of the week, the promotional offer value, and the user's happiness index.

Tasks

- Data Loading: Load the 'user_engagement.csv' dataset into a pandas DataFrame.
- Exploratory Data Analysis (EDA): Conduct an exploratory data analysis to understand the structure and the distribution of the data. Generate summary statistics and visualize the data as you see fit.
- Data Preparation: Create a feature matrix X which includes 'day_of_week', 'promotion_value', 'happiness_index', and a target vector y with 'daily_active_minutes'. You might also need to handle missing values and outliers if necessary.
- Data Splitting: Split the data into a training set (70%) and a test set (30%). Ensure that you use a random seed for reproducibility of results.
- Model Building: Train at least two different types of machine learning models on the training data. You may consider models such as linear regression, decision tree, random forest, or any other model you think would be suitable for this regression task.
- Model Evaluation: Evaluate each model's performance on the test data. Use appropriate metrics for regression tasks such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2).
- Model Selection: Based on the evaluation metrics, choose the model that you believe is the best for this task. Justify your model choice in the coding challenge form

- Prediction Function: Write a function *predict_active_minutes(day_of_week, promotion_value, happiness_index)* that takes these inputs and returns the predicted number of active minutes for that day.

Important: Please include comments in your code explaining your thought process and decisions.

Submission

Submit your solution as a Python (.py) script or a Jupyter notebook (.ipynb) in coding challenge form. Your solution should be able to be run as a standalone script or notebook with minimal setup.

Questions

Please reach out to Adi (adi@neumile.com) if you have any questions.