Junior Machine Learning Developer Coding Assignment

Assignment Title: Ethereum Blockchain Transaction Analysis and Prediction

Objective:

The goal of this assignment is to analyze a dataset of Ethereum blockchain transactions and develop a machine learning model to predict the gas price of future transactions. This task will help you understand the basic processes involved in data preprocessing, exploratory data analysis, feature engineering, model building, and evaluation.

Ethereum Blockchain Transactions:

https://www.kaggle.com/datasets/bigquery/ethereum-blockchain?select=transactions

Dataset:

Use the Ethereum Blockchain Transactions dataset from Kaggle. Download the dataset and follow the instructions below.

Tasks and Deliverables:

1. **Data Exploration and Preprocessing (30 points)**
    - Load the dataset and understand its structure.
    - Perform basic data cleaning, including handling missing values and removing duplicates.
    - Generate summary statistics and visualize the distributions of key features.
    - Examine the correlations between features.
    - Deliverable: A Jupyter notebook with code and markdown explaining your findings and preprocessing steps.
2. **Feature Engineering (20 points)**
    - Create new features that may be useful for predicting gas prices, such as transaction time of day, day of the week, or transaction volume.
    - Normalize or standardize features if necessary.
    - Deliverable: An updated Jupyter notebook with the new features added and explanations for why these features might help.
3. **Model Building and Training (30 points)**
    - Split the dataset into training and test sets.
    - Train at least three different models (one of these models to be a neural network) to predict the gas price.
    - Use cross-validation to tune hyperparameters for each model.
    - Deliverable: A Jupyter notebook with the model training code, hyperparameter tuning, and cross-validation results.
4. **Model Evaluation and Selection (20 points)**
    - Evaluate the performance of each model using appropriate metrics
    - Compare the models and select the best one based on performance metrics.
    - Deliverable: A Jupyter notebook with the evaluation results and your reasoning for selecting the best model.

5. **Documentation and Presentation (Extra 10 points)**
    - Create a final report summarizing your approach, findings, and results.
    - Include visualizations that help explain your process and results.
    - Deliverable: A PDF report generated from your Jupyter notebook, or a presentation deck.

Submission Guidelines:

- Submit your Jupyter notebook(s) with clear, well-documented code and explanations.
- Ensure all plots and visualizations are properly labeled.
- Include your final report or presentation summarizing the work done.
- The deadline for submission is 02.08.2024.

Evaluation Criteria:

- **Code Quality:** Clarity, organization, and use of best practices.
- **Completeness:** Completion of all specified tasks and deliverables.
- **Documentation:** Clarity and comprehensiveness of explanations and justifications.
- **Results:** Accuracy and reliability of the model predictions.

Additional Resources:

- Pandas Documentation
- Scikit-learn Documentation
- Matplotlib Documentation
- Ethereum blockchain transactions: https://www.kaggle.com/datasets/bigquery/ethereum-blockchain?select=transactions