Artificial Intelligence Techniques (CSY3025)

Deadline: 18th March 2024 (11:59)

Assignment Title: Advanced Predictive Modeling on the California Housing Dataset

Objective:

Students will develop a predictive model to estimate median house values in California districts. They are free to choose any algorithm they believe is best suited for the task. The project will involve data exploration, preprocessing, feature engineering, model selection and justification, implementation, and a detailed comparison of their chosen model against a baseline model (Linear Regression). The findings and methodologies will be documented in a comprehensive technical report.

Baseline Model:

Before embarking on model selection and implementation, it is crucial to understand the performance benchmark set by the baseline model. A simple Linear Regression model has been applied to the California Housing Dataset, yielding the following results:

- **Mean Squared Error (MSE):** 0.5559
- **R-squared Score:** 0.5758

These metrics serve as the baseline performance indicators. Your objective is to develop a model that not only surpasses these figures but also provides insightful analysis and justification for the chosen methodology.

The R-squared score is particularly suited to regression models like Linear Regression as it represents the proportion of variance in the dependent variable that is predictable from the independent variables. It is a measure of the model's accuracy in explaining the observed variation.

Performance Metrics for Your Model:

Depending on the algorithm you select, the choice of performance metrics might vary. While MSE and R-squared are common for regression tasks, other metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or even more complex ones might be more appropriate based on the specific characteristics and objectives of your model. It is crucial to justify the selection of performance metrics in your technical report, explaining how they provide a meaningful evaluation of your model's performance in the context of the task.

Dataset:

The dataset for this assignment is the California Housing Dataset. It can be accessed through popular data science libraries like Scikit-Learn, or directly from public repositories.

- **Scikit-Learn Documentation for California Housing Dataset**: Scikit-Learn Dataset Description
- **Direct Dataset Link**: https://www.kaggle.com/datasets/camnugent/california-housing-prices

Part 1: Project Solution (Practical Implementation)

Tasks:

1. **Data Exploration and Preprocessing**:
    - Perform initial data exploration to understand the dataset's features and target variable.
    - Clean the data by handling missing values, removing outliers, and normalizing data if necessary.
2. **Feature Engineering**:
    - Create new features that could improve the predictive performance of the models.
3. **Model Selection**:
    - Research and select algorithm other than Linear Regression. Justify the choice based on the dataset's characteristics and preliminary analysis.
4. **Model Implementation and Baseline Comparison**:
    - Implement the chosen model and Linear Regression as a baseline for comparison.
    - Train and evaluate both models using appropriate metrics (e.g., MSE, RMSE, R^2 score).

Part 2: Project Technical Report

Report Structure:

1. **Introduction**:
    - Overview of the dataset and the prediction task.

- Objectives and scope of the project.

2. **Data Exploration and Preprocessing**:
    - Insights from initial data analysis, including visualizations.
    - Detailed description of data cleaning and preprocessing steps.
3. **Feature Engineering**:
    - Explanation of new features created and their expected impact on the models.
4. **Model Selection and Justification**:
    - Comprehensive justification for the chosen algorithm.
    - Theoretical background of the selected model and its applicability to the dataset.
5. **Model Implementation and Evaluation**:
    - Description of the implementation process for both the chosen model and the baseline model.
    - Comparative analysis of their performances with detailed explanations and visualizations.
6. **Conclusion and Future Work**:
    - Summary of findings, including the effectiveness of the chosen model compared to the baseline.
    - Recommendations for future improvements or areas of research.
7. **References**:
    - Citations for all datasets, libraries, and external resources used in the project.

Deliverables:

- Source code for the entire project, including data preprocessing, feature engineering, model implementation, and evaluation.
- A technical report in PDF format that covers all the sections outlined above, emphasizing the rationale behind model selection and the insights gained from the comparative analysis.

- Viva/Demo of 5 min in Week 7 class 11th March 2024

Evaluation Criteria:

- Creativity and effectiveness in feature engineering and data preprocessing.
- Solid justification for the choice of algorithm.
- Rigor in model implementation and evaluation.
- Depth of analysis and insightfulness in the technical report.

Academic Integrity

Check UON Academic Integrity and Misconduct policy

Avoid plagiarism, collusion, contract cheating.

All content (including the source code) borrowed from the Internet and ChatGPT-like AI tools must be correctly referenced in the submitted work.