[F24] MACHINE LEARNING ASSIGNMENT 1

Due Date: 4th October 2024 (23:59)
Submission Format: Jupyter Notebook(s) and report (PDF).
Data: [Regression and Classification]

ON-TIME ORDER PREPARATION

Timely preparation of orders is critical in industries such as food delivery, e-commerce, and manufacturing. Accurate prediction of preparation time can help optimize resources, reduce delays, and enhance customer satisfaction. This assignment will challenge you to build both regression and classification models to address these issues.

Expected outcomes:

- Data analysis : Data preprocessing and visualization (10 points)
- Feature selection and engineering: Selection of features and extraction form timestamps (15 points)
- Prediction of order preparation time - **Regression task** (15 points)
- Classification whether an order will be prepared on time or late - **Classification task** (10 points)
- Comparison of the selected machine learning models' performance using the appropriate evaluation metrics. (10 points)
- Hyperparameter optimization / tuning (10 points)
- For classification task : Balancing of data to improve model performance (5 points)
- Report and source code (25 points)

DATASET

The dataset is stored in a database, and you will need to assemble the required data by querying the relevant tables. Basic knowledge of MySQL will be sufficient to extract the data for your analysis and model training. You can find detailed information about the structure of the data tables and their data types in the provided documentation.

TASK 1: REGRESSION TASK - PREDICTING PREPARATION TIME

Your first task is to predict the Actual Preparation Time (in minutes) using the provided features. This task will require you to:

(1) **Analyze the dataset**: Perform exploratory data analysis (EDA) to understand the relationships between features and the actual preparation time.
(2) **Preprocess the data**: Handle missing data, perform feature engineering if necessary, and split the data into training (90%) and testing sets.
(3) **Build a regression model**: Train a regression model (e.g., linear regression, support vector regression, neural network, etc.) to predict the actual preparation time. A minimum of 3 machine learning algorithms should be used for the prediction. One of the algorithms should have regularization.
(4) **Evaluate the model**: Use evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared to assess the performance of your models.
(5) **Interpret the results**: Identify the most significant features affecting preparation time and discuss the accuracy and reliability of your predictions.

TASK 2: CLASSIFICATION TASK - PREDICTING ON-TIME COMPLETION

Your second task is to classify whether an order will be prepared on time or late. This task will require you to:

(1) **Analyze the dataset**: Perform exploratory data analysis (EDA) to understand the relationships between features and the actual preparation time.
(2) **Preprocess the data**: Handle missing data, perform feature engineering if necessary, and split the data into training (90%) and testing (10%) sets.
(3) **Remove outliers and balance the dataset**: Implement outlier detection techniques (e.g., Z-score, IQR, or isolation forests) to identify and handle these outliers. Balance the dataset using one of the following techniques: [**imbalanced-learn**,1, 2, 3, 4]
(4) **Train a classification model**: Use a classification algorithm (e.g., logistic regression, SVM, neural network, etc.) to predict if an order will be prepared on time or not using planned time and factual time (if absolute difference between "planned_prep_minutes" and "order_ready_time" - "order_prep_start" is greater than 5 minutes then the order is not prepared on time).
(5) **Evaluate the model**: Assess the performance using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.
(6) **Interpret the results**: Understand the factors contributing to the on-time or late preparation of orders and suggest improvements.

BONUS TASK: OUTLIER DETECTION & FUTURE ORDERS FORECASTING

For this bonus task, you will extend your analysis by developing a model to forecast the number of new orders (for next *n* minutes) for a group of stores in a region. Additionally, you will identify and handle outliers in the dataset to improve the robustness of your model. Jupyter Notebook for this task should be separate.

Subtasks

- **Outlier Detection** (5 points): Analyze the dataset for outliers that could negatively impact the performance of your model. Implement outlier detection techniques for time-series data to identify and handle these outliers [1, 2, 3].
- **Forecasting Future orders** (5 points): Build a forecasting model to predict the number of new orders for each region over a specified future time period (i.e for next *n* minutes). Where *n* is a hyper-parameter.

REPORT AND SOURCE CODE

After performing the comparison of the machine learning models, the results should be presented in a form of a report. The implementation should be in python and PyTorch should be used for neural networks. Your submission should contain

- A well documented Jupyter Notebook (with all cells outputs)
- PDF report

Your report should contain

- **Motivation** : explanation (written in your own words) of the importance of the two tasks that you solved from the perspective of food delivering.
- **Data** : Brief description of the both Regression and Classification data (features and predictor)
- **Exploratory data analysis** : What are the insights from data exploration and did these insights help in feature selection or model design?
- **Task**: The definition of the learning tasks in terms of machine learning, i.e. estimating function
- **Input Format**: If you used an alternative data input format, explain it.
- **Comparison of selected ML models**: Describe which model is better in each task based on the cross-validation performance. Does the model overfit? underfit? How did you avoid both? Use graphs and tables to document the results of your experiments.
- **Results interpretation**: Understand the factors contributing to the prediction power of the ML models and suggest improvements.

The report should be submitted in PDF format. The report should not be more than 2 pages using Association for Computing Machinery (ACM) - SIGPLAN Proceedings template. (if specified report format is not used, 5 points will be deducted from the overall grade) **Template Link**