ISB46703 Principles of Artificial Intelligence
Universiti Kuala Lumpur
October 2023
F.M.N

Mini Project

Weight: 15%
Due: 6 February 2024

1 Introduction

The objective of this group assignment is to apply regression on a bike rental dataset to predict number of bike rentals. This project will cover basic exploratory data analysis, data visualisation, data pre-processing and model evaluation for the regression model.

1.1 Goal

Your final goal is to predict the number of bike rentals.

2 Dataset

Our dataset comes from a popular machine learning repository that hosts open source datasets for educational and research purposes, the UCI Machine Learning Repository. The dataset can be downloaded from here.

2.1 Bike Rentals

The dataset contains information about bike rentals, including various features such as date, hour of the day, temperature, humidity, windspeed, visibility, dew point temperature, solar radiation, rainfall, snowfall, seasons, holiday status, and functional day status. The description of the dataset can be found here

3 Objective

Your task is to build a machine learning model to predict the number of bike rentals based on the given features.

4 Instructions

4.1 Exploratory Data Analysis [10 points]

1. Load the dataset and perform initial data exploration.
2. Examine the distribution of the target variable, `Rented Bike count`.

4.2 Data Visualisation [10 points]

1. Create visualisations to explore relationships between features and the claim status.
2. Use appropriate plots and charts to illustrate the patterns and trends in the data.
3. Analyse the impact of different features on the likelihood of filing an insurance claim.

4.3 Data Preprocessing [10 points]

1. Explore and handle any outliers in the numerical features.
2. Perform necessary pre-processing steps such as handling missing (NaN) values if any, encoding categorical variables (Seasons, Holiday, Functional Day), and scaling numeric features by computing its z-score:

$$z = \frac{x - \mu}{\delta}$$

3. Split the dataset into training and testing sets (70 : 30 ratio) using `train_test_split()` function.

4.4 Modelling [5 points]

1. Import `LinearRegression()` library from `sklearn.linear_model`
2. Fit the training data.

4.5 Model Evaluation [10 points]

1. Predict the number of bike rentals using the trained model on testing data.
2. Evaluate the performance of the model using regression metric *Root-Mean Square Error* (RMSE).
3. Interpret the results and discuss the model's effectiveness in predicting number of bike rentals.

4.6 Model Improvement [10 points]

1. Repeat Section 4.3, but before standardising (step 2) transform feature `Windspeed` using $\sqrt{x}$.
2. Now, instead of standardising, normalise the data:

$$d'_i = \frac{d_i - min_D}{max_D - min_D}$$

   where $D = [d_1, d_2, d_3, ..., d_i]$.

3. Normalisation can be performed using `sklearn.preprocessing.MinMaxScaler`.

4. Repeat Section 4.5 using new transformed and normalised data and compare the RMSE with previous model.

5. Interpret the results and discuss the impact of transformation in predicting bike rentals.

4.7 Conclusion and Discussion [5 points]

1. Summarise the findings of the assignment, including the initial model's performance and the impact data transformation.

2. Discuss the limitations of the experiment and suggest potential improvements.

5 Groupwork Policy

Each group member should actively participate in all stages of the assignment, including exploratory data analysis, data visualisation, data pre-processing, model creation, evaluation and conclusion. Separate peer reviews will be conducted to evaluate each member's contribution, providing a fair assessment. Collaboration, division of tasks, and effective communication within the group are essential for successful completion of the assignment.

6 Presentation

Each group will present their findings, including an overview of the dataset, exploratory data analysis, data visualisation, data pre-processing steps, model creation, evaluation results, and final conclusions. Visual aids, such as graphs and plots, are encouraged to enhance the clarity of the presentation.

Presentation date: **Tuesday 6 February 2024, 10:30am - 12:30pm**

7 Submission

Only one submission per–group. Please submit the following materials:

- Jupyter Notebook or code files used for data analysis and modeling via VLE.
- Presentation slides summarising the key findings and insights from the assignment.