**Bike Sales Prediction**

**In Seoul, South Korea**

1. **Main Objective**

Attemp to predict sales using regression techniques with the focus on prediction task.

In this report, the main goal is predicting rented bike in spring and summer seasons.

1. **Data Definition and Description**

Seoul bike sales is an open public datasets provided by University of California Irvine in their machine learning repository. In brief explanation, the dataset is about importancy to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

The dataset contain 8760 observations and 14 attributes (variables).

Date : year-month-day

Rented Bike count : Count of bikes rented at each hour

Hour : Hour of he day

Temperature : Temperature in Celsius

Humidity : in percentage

Windspeed : m/s

Visibility : 10m

Dew point temperature : Celsius

Solar radiation : MJ/m2

Rainfall : mm

Snowfall : cm

Seasons : Winter, Spring, Summer, Autumn

Holiday : Holiday/No holiday

Functional Day : NoFunc(Non Functional Hours), Fun(Functional hours)

1. **Pre-processing and Feature Engineering**

* Inspect the columns and its data type
* Inspect null value
* Inspect duplicate
* Fix the column data type and names
* Create custom function to get column statistic for detecting and outlier treatment
* Inspect outliers and treat the outliers with reasonable value
* Inspect the independent variable correlation strength to the target label
* Due to prediction main objective, variable transformation will be performed
* Wind speed\_ms transformed with log1p
* Solar Radiation\_mjm2 transformed with log1p
* Rainfaal transformed with log1p
* Humidity trasnformed with boxcox1p
* Feature selection for modeling; drop date,snowfall,winter-autumn seasons feature
* Create machine learning pipeline for further transformation;

1. Perform One hot encoder to categorical variable
2. Standard Scaling
3. **Model Selection and Training Results**

To achieve best prediction accuracy, we will conduct three different techinques which are Linear Regression, Polynomial Regression, and Elastic Net Regularization. The model training conducted with ;

* Pipeline technique
* cross-validation with 5 k-fold
* The model training use train-test split with 10 % hold out test set
* Random state: 42
* Mean absolute error, root mean square error and R squared metrics
* Grid search for searching best parameter

**Modeling**

1. Base Linear Regression Model

Summary performance of base linear regression on hold-out test set ;

mean of the y hat: 884.6565977742448

Linear regression MAE score: 359.78258749649007

Linear regression RMSE score: 479.84186908708415

Linear regression R2 score: 0.49984074925764777

1. Polynomial Regression

Summary performance of Polynomial regression on hold-out test set ;

To search parameters {polynomial degree : 2,4}

Best degree : 2

mean of the y hat: 884.6565977742448

Polynomial regression MAE score: 293.4903259236082

Polynomial regression RMSE score: 390.16239709328227

Polynomial regression R2 score: 0.6693238693048882

1. L3 (Elastic Net) Regularization

Summary performance of Elastic net regression on hold-out test set ;

To search parameters {polynomial degree: 2,4 ; lasso regularization ratio : 0, 0.25, 0.5, 0.75, 1 ; regularization strength (alpha): 0.01,0.5,1,3,5,10}

Best result : polynomial degree 4, l1\_ratio 0.75, alpha 0.01

mean of the y hat: 884.6565977742448

Polynomial regression MAE score: 212.96796168019475

Polynomial regression RMSE score: 306.1322599677754

Polynomial regression R2 score: 0.7964222918698888

1. **Model Selection, Conclusion, and Summary**

Based on the performance metrics it is highly **recommended to choose Elastic Net regression** with the best parameter result which is

* Elastic Net regression
* L1 ratio: 0.75
* Alpha: 0.01
* Polynomial degree: 4
* R squared score : 80%

From now on, this model will be mentioned as choosen model.

Due to prediction accuracy main objective, model training will focus on how to create a model with high prediction accuracy. With polynomial degree of 4 and the attribute of 11 independent variables **the choosen model** will generate **near to black box model**, which mean even the model engineer does not know what happened inside the algorithm.

Based on the parameters used and other model comparison, we can conclude that below conlusion will drive model accuracy;

1. Adding regularization will boost the test accuracy
2. The chosen model use 75 percent on LASSO regularization and using Ridge for the rest of it.
3. Strong regularization, with alpha 0.01
4. Need more complexity to capture actual data behavior, polynomial degree 4
5. **Further Development**

With dataset open-sourcity and broader techniques, any researcher can expand this report or conduct different research such as implementing linear regression for the interpretability purpose, time-series analysis, exploratory data analysis on this dataset, and many more. As the author, i would suggest the data need more bike specific feature that possibly drive the sales such as color of the bike, bike type, price, etc.

**References**

<https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/0fafd523-aabb-41c7-86e6-97c5191d8ee4/view?access_token=feb62114edb9bae699554d726666cbbdd416125cc85f8034aa0221ea924ec949>

<https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand>