Dynamic Pricing Prediction Using Supervised Regression

Model Benchmarking and Optimization for Bike Rental Pricing

Contents

[Introduction 1](#_heading=h.gjdgxs)

[Supervised Machine Learning - Regression 1](#_heading=h.vffe4bl1iaa2)

[Data Understanding and Preparation 2](#_heading=h.64mmnpz47kzy)

[Machine Learning Implementation 3](#_heading=h.dbzrukh8w41)

[Hyperparameter Optimization 5](#_heading=h.jvogeqrwwoc2)

[Machine Learning Evaluation 6](#_heading=h.oogum3dym0op)

[Conclusion 10](#_heading=h.8ovsh3u8pv1k)

[References 10](#_heading=h.n0256wii3knl)

[Bibliography 10](#_heading=h.wlm8ot9vpa1k)

# Introduction

The objective of the assessment is to analyse a biking rental dataset and implement a machine learning model to improve their pricing strategy, fleet management or customer satisfaction. The action that I have taken is to improve the pricing strategy for the bike rental cost. I have implemented a supervised machine learning regression approach to predict the bike rental cost for dynamic pricing based on the features in the dataset. The machine learning models that I have used are the ridge and random forest regression while applying hyperparameter optimization.

# Supervised Machine Learning - Regression

I have chosen a supervised machine learning approach because the dataset included the target variable bike rental cost to predict its value based on the dataset features.

The main objective in supervised learning is to train the machine learning model on a labeled dataset to make predictions on unseen or future data. (Raschka and Mirjalili, 2019, p. 3) Supervised machine learning is commonly used for classification to predict a categorical label or regression to predict a continuous value. Below is the advantages and disadvantages of implementing supervised machine learning:

Advantage

* There is a clear objective when classifying or predicting a specific target value.
* It offers some degree of interpretability to understand the prediction outcome.
* There is a clear performance metric for machine learning evaluation.

Disadvantages

* It requires a large amount of labeled data to achieve optimal results.
* When using large datasets, it can require significant computational resources for the training.
* The model is dependent on the training data. If the data is biased or incomplete it will have an effect on the model's performance.

# 

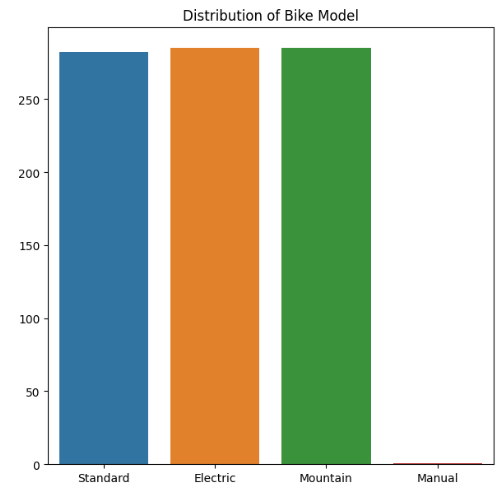
# 

# Data Understanding and Preparation

I have conducted an exploratory data analysis and data preparation for the machine learning implementation. The dataset has 900 rows and 8 columns. The columns of the dataset is the following below:

|  |  |
| --- | --- |
| Columns | 'City', 'Weather Condition', 'Rider Age', 'Bike Model', 'Ride Duration (min)', 'Distance Covered (km)', 'Rider Satisfaction' and Bike Rental Cost ($). |

From inspecting the dataframe, I found several missing values and removed the rows that contained them to provide consistency on the data. There were no duplicate rows or outliers in the dataframe. However I have found that there was only 1 row that contained the bike model “manual”. I have removed this row as there wasn’t enough data for the bike model “manual” compared to the other bike models.



Based on the exploratory data analysis, I have used the following features and target for the machine learning regression implementation below:

|  |  |
| --- | --- |
| Features | 'City', 'Weather Condition', 'Rider Age', 'Bike Model', 'Ride Duration (min)', and 'Distance Covered (km)' |
| Target | Bike Rental Cost ($). |

I have excluded the 'Rider Satisfaction' column, as you won’t have access to this information for dynamic pricing when a user is booking a rental bike. There are also other factors that could influence the rider satisfaction result outside of the data that was provided.

As the ‘City', 'Weather Condition' and 'Bike Model' are categorical data, I have applied the one hot encoding method to convert it to numerical data for the machine learning implementation.

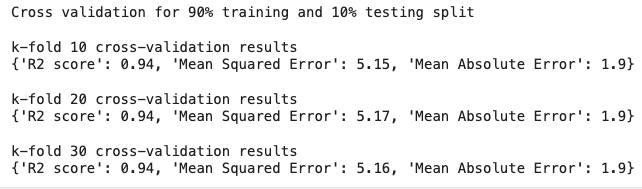
# Machine Learning Implementation

I have implemented both the ridge regression and random forest regression model to predict the bike rental cost. The ridge regression model is a type of linear regression that includes a regularization technique that prevents overfitting (Jolly, 2018, p. 57). The random forest regression model is a type of ensemble learning method that uses multiple decision trees to make predictions (Raschka and Mirjalili, 2019, p. 345). It is used to handle both linear and non-linear data relationships when predicting its final prediction.

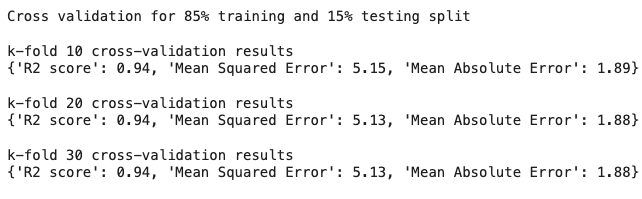
Below I have provided my results for the R2 score, Mean Squared Error (MSE) , and Mean Absolute Error (MAE) when applying k-fold cross validation on three different training and testing splits for the machine learning models below:

**Ridge Regression Model Performance**

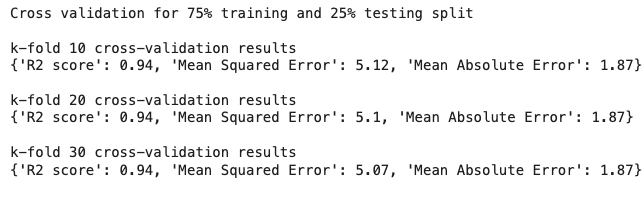
**90% training and 10% testing split**

****

**85% training and 15% testing split**

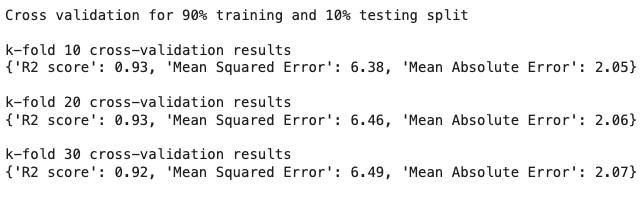
****

**75% training and 25% testing split**

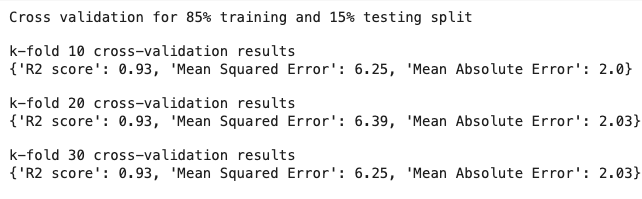


**Random Forest Regression Model Performance**

**90% training and 10% testing split**

****

**85% training and 15% testing split**

****

**75% training and 25% testing split**

****

Based on the above results, the cross-validation across different k-folds produced a similar outcome which indicates that the model's performance is stable.

# Hyperparameter Optimization

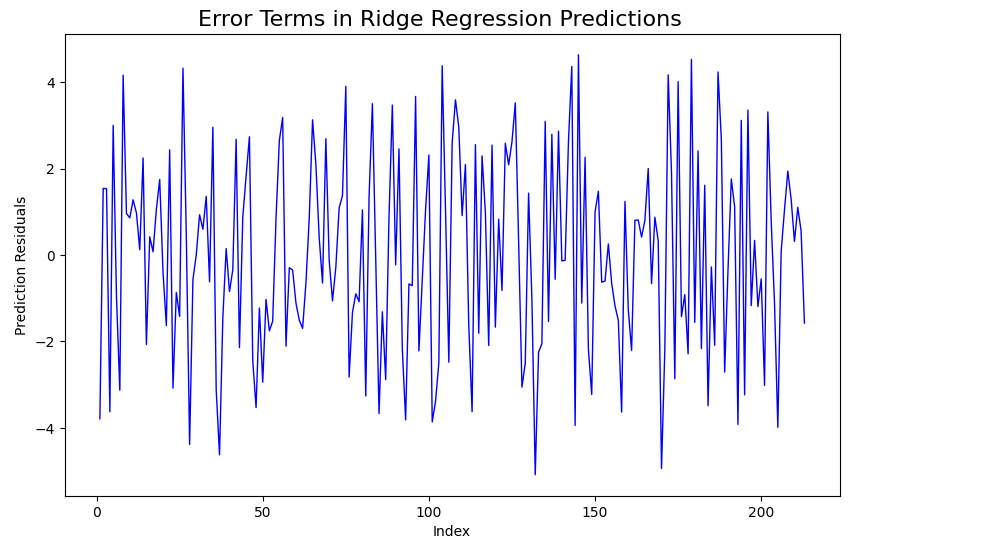
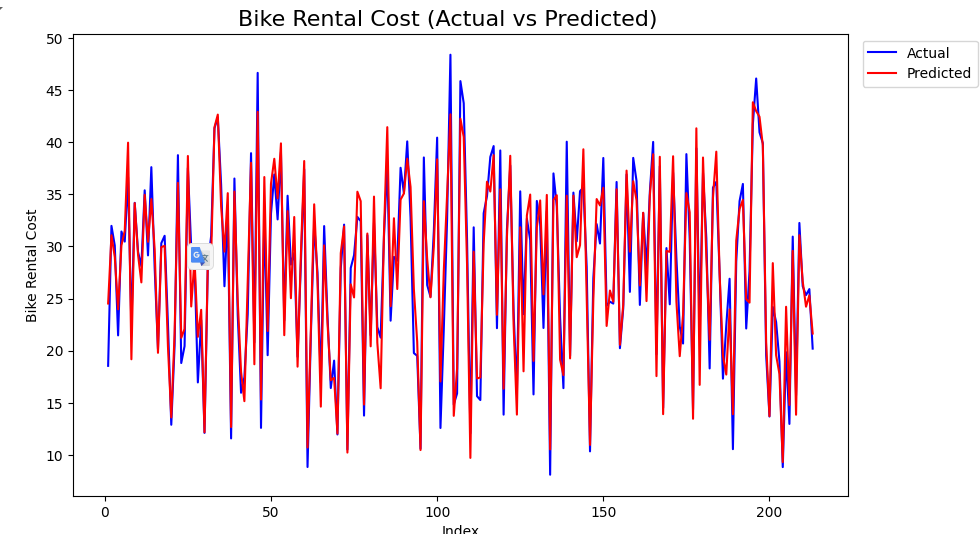
The hyperparameter optimization technique that I have used in this assessment is the GridSearch method. It is used to go through all possible hyperparameter combinations based on the set that you have defined for the machine learning model (Brownlee, 2019). While this approach can be time intensive, it provides the most optimal hyperparameter values for the machine learning model implementation. Below is the best hyperparameter values that I have found for the following machine learning models:

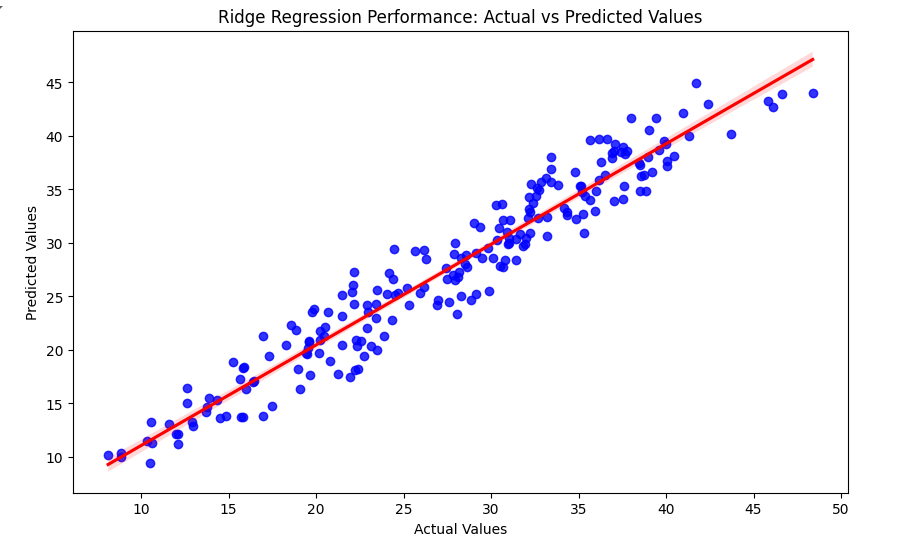
|  |  |
| --- | --- |
|  | **Hyperparameters** |
| **Ridge Regression** | {'alpha': 1, 'fit\_intercept': True, 'solver': 'lsqr'} |
| **Random Forest Regression** | {'max\_depth': 10, 'max\_features': None, 'min\_samples\_leaf': 4, 'min\_samples\_split': 10, 'n\_estimators': 100} |

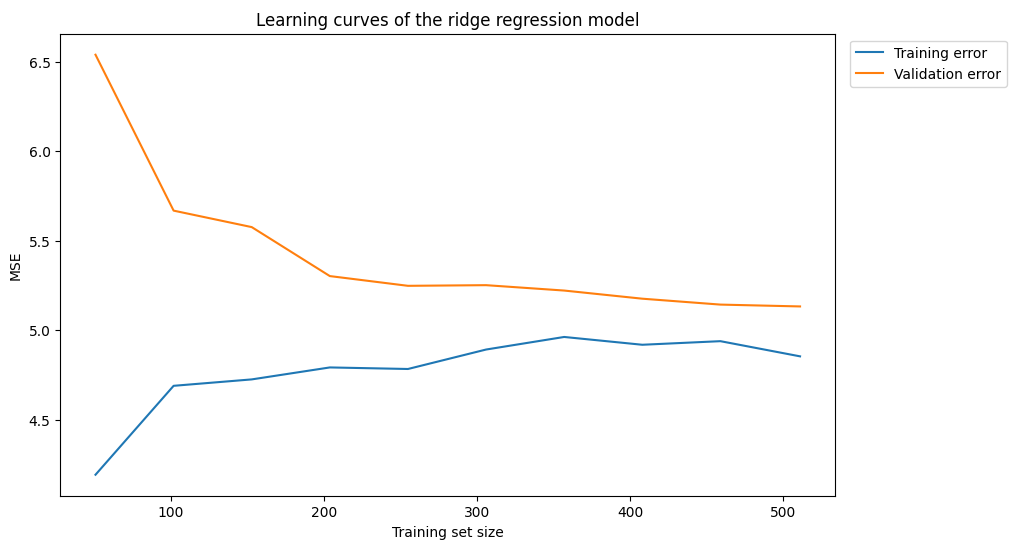
After applying the hyperparameter values to the machine learning model, I have found a slight improvement in terms of its performance.

# Machine Learning Evaluation

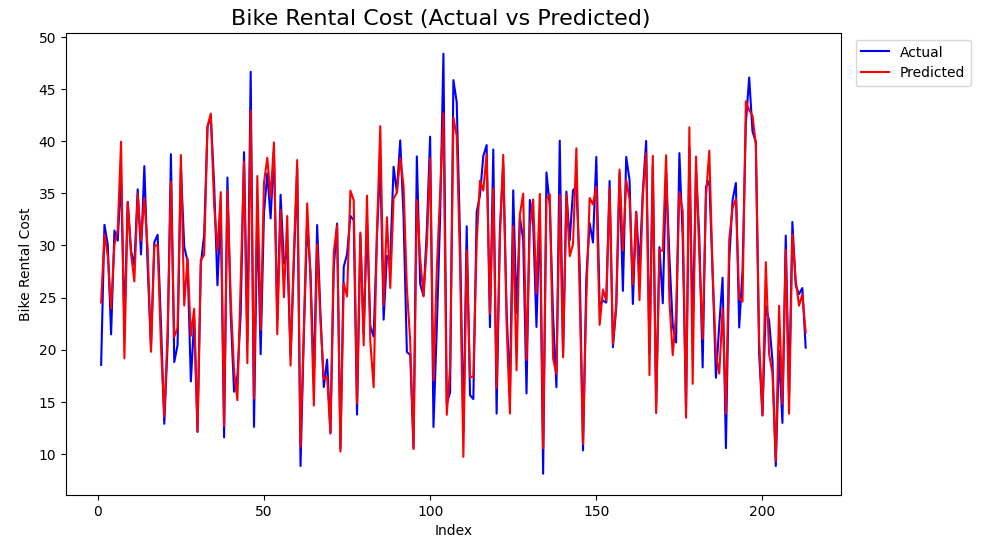
Both Ridge Regression and Random Forest Regression have demonstrated similar performance results in terms of its R2 score, Mean Squared Error (MSE) , and Mean Absolute Error (MAE). It also produced generalization when working on both training and unseen testing data. Below I have provided some visualisation on the performance result of the machine learning models below:

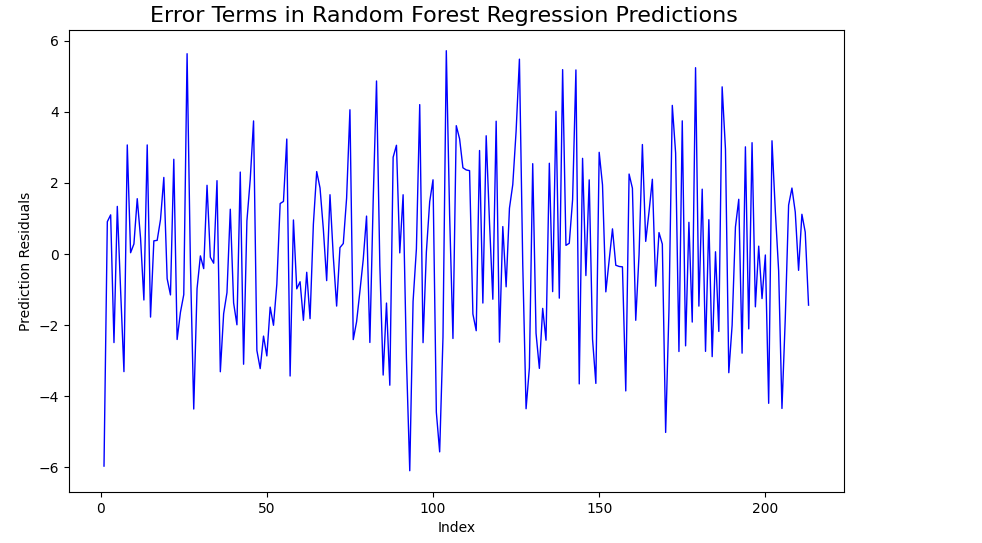
**Ridge Regression Model Performance  
**

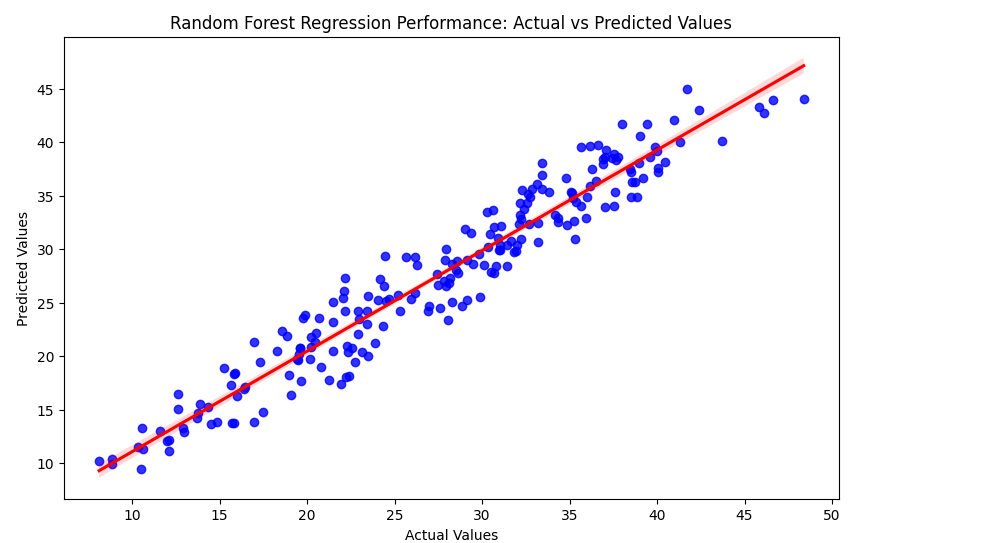


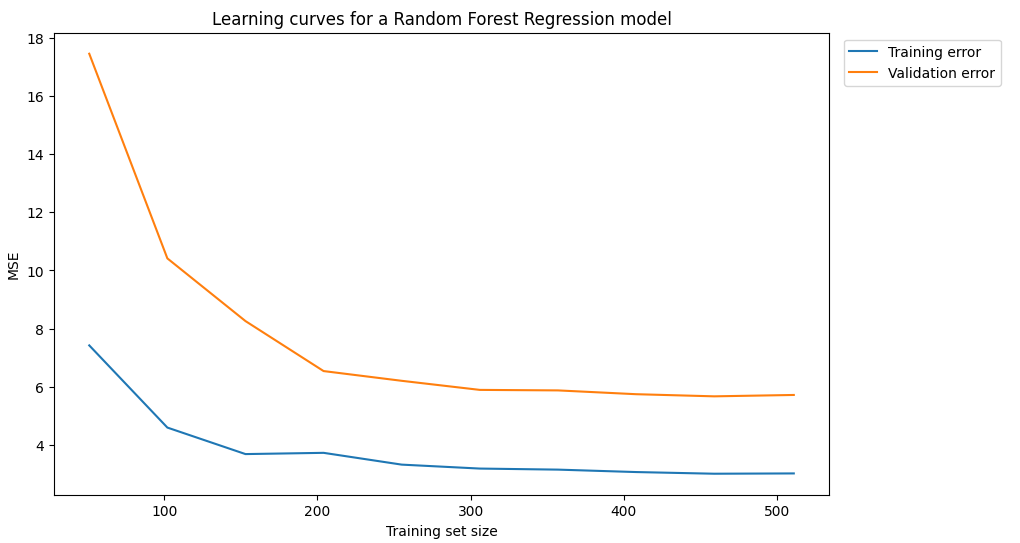


**Random Forest Regression Model Performance**









# Conclusion

Based on the implementation that I have conducted to predict the bike rental cost for dynamic pricing, I have found that the ridge regression model would be the best machine learning model to implement compared to the random forest regression model. While both produced similar results, the ridge regression model was 29 times faster than the random forest regression model. Another reason for choosing the ridge regression model would be the strong linear relationship between the feature and target variable in the dataset.

# 

# References

* Brownlee, J. (2019). *Hyperparameter optimization with random search and grid search*. Machine Learning Mastery. Available at:<https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/> [Accessed 23 December 2024].
* Kevin Jolly (2018) *Machine Learning with Scikit-learn Quick Start Guide : Classification, Regression, and Clustering Techniques in Python*. Birmingham, UK: Packt Publishing. Available at: https://research.ebsco.com/linkprocessor/plink?id=e73a2585-fef9-3744-8789-852897fb6045 (Accessed: 23 December 2024).
* Sebastian Raschka and Vahid Mirjalili (2019) *Python Machine Learning : Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2*. [S.l.]: Packt Publishing. Available at: https://research.ebsco.com/linkprocessor/plink?id=fde49815-c3b7-330b-8084-49a6b9f7d99a (Accessed: 23 December 2024).

# Bibliography

* Brownlee, J. (2019). *Hyperparameter optimization with random search and grid search*. Machine Learning Mastery. Available at:<https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/> [Accessed 23 December 2024].
* DataCamp. (2024) *A Comprehensive Guide to K-Fold Cross Validation*. Available at:<https://www.datacamp.com/tutorial/k-fold-cross-validation> (Accessed: 23 December 2024).
* DataCamp (2022) *Lasso and Ridge Regression in R Tutorial*. Available at:<https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression> (Accessed: 26 December 2024).
* Kevin Jolly (2018) *Machine Learning with Scikit-learn Quick Start Guide : Classification, Regression, and Clustering Techniques in Python*. Birmingham, UK: Packt Publishing. Available at: https://research.ebsco.com/linkprocessor/plink?id=e73a2585-fef9-3744-8789-852897fb6045 (Accessed: 23 December 2024).
* Luís Fernando Torres. (2023). *Evaluation metrics for regression models*. Medium. Available at:<https://medium.com/latinxinai/evaluation-metrics-for-regression-models-03f2143ecec2> (Accessed: 23 December 2024).
* Sebastian Raschka and Vahid Mirjalili (2019) *Python Machine Learning : Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2*. [S.l.]: Packt Publishing. Available at: https://research.ebsco.com/linkprocessor/plink?id=fde49815-c3b7-330b-8084-49a6b9f7d99a (Accessed: 23 December 2024).