Regression Based Prediction Model for Bike Sharing Resource Management System

https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Thilini Deshika Sudam Kalpage Hashan Eranga

Abstract

Cities are adopting innovative mobility solutions to address smart city concerns such as carbon emission reduction, multimodal urban transportation, and pandemic risk mitigation. Emphasizing the use of shared forms of transportation, such as bike-sharing networks. This work asks a research issue and conducts a comprehensive literature assessment, concentrating on the contributions of machine learning techniques used to bike-sharing systems to improve city transportation. After going through the data cleaning tried to see whether there was any missing data within the dataset. This was taken to acknowledge before the process to make sure all the data are there in the dataset. Then descriptive statistical analysis was done on the dataset to convert fields into numerical. After all the data pre-processes are done. Develop some data visualizations on the dataset for identifying relationships between each feature. For that, barcharts, histograms, boxplots and heap diagrams were used. Above visualization approaches to refrain from the features which provide similar behaviours to the model. The model was trained using machine learning algorithms; Linear Regression, Lasso, Ridge, Decision Tree, Random Forest, Gradient Boosting estimator.

Introdution

Up to date bike sharing systems are able to automate the system and replace the traditional bike rental systems, From user registration, subscription and all the activities are connected. This system enables an efficient service for the clients while reducing a larger amount of carbon footprint from the world. Since there are about over 500 bike sharing programs around the world. This plays an important role in traffic,nature and for the health of the living beings.

Problem

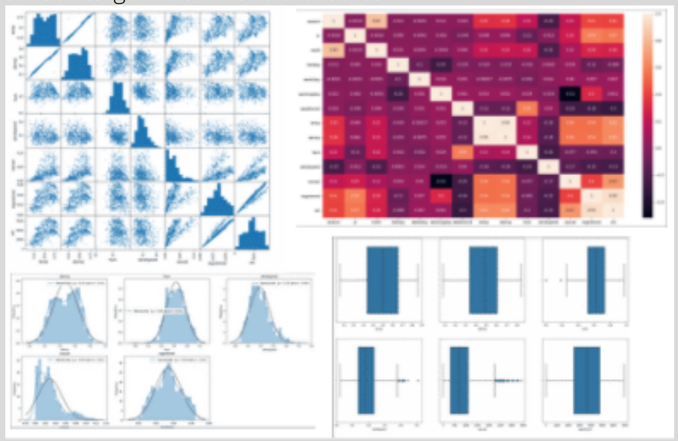
Although having a fully automated system may not resolve all the requirements. This system is capable of solving most of the requirements of the client phase. But in order for more financial and sustainable optimization, the data collected during the system operates should help to develop the system more efficiently. Currently, one of the main problems and challenges in the management of Bike Sharing Systems is to assure that users will be able to find available bicycles and parking slots in a station independently of the time. However, users' behavior causes bicycles to be asymmetrically distributed. Therefore a Rebalancing System is needed to maintain the adequate number of bikes at each station, in order to satisfy the demand. Rebalancing is a costly operation in terms of logistics and its inefficiency represents the main cause of dissatisfaction within customers. The design of algorithms that help and advise operators to redistribute bicycles accurately and efficiently through the cities is a very important step in terms of sustainability of these systems.

Methodology

For Data Wrangling it was checked whether there are null values present in the dataset and was removed from the dataset. The ‘dteday’ column was updated with the date of the ‘dteday’ column. Data Preprocessing includes normalizing numerical columns; ‘registered’ and ‘casual’. Then, one hot encoding was applied for the categorical features. The skewness of the 'casual' column was reduced using cubic root transformation. The Data Cleaning was done by removing outliers. Feature selection was done based on correlations values between attributes and target column.

Data Visualization

Distribution Plots were used to visualize the target column. Numeric features were visualized in histograms, while categorical features were visualized in bar charts. Scatter plots with correlation statistics were used to visualize the intersection of each numeric feature and label values. Box plots were used to compare the target column with categorical features. Scatter matrix and Heatmap were used to visualize the relationships in between attributes and also between target column and attributes.



Split Data

The dataset was split as 70%-30% into the training set and the test set.

Train Regression Model

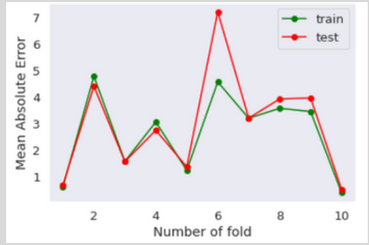
The model was trained for six regression algorithms. Linear Regression, Lasso and Ridge Regression are linear algorithms used. A tree-based algorithm, Decision Tree Regression was also used to train the model. As ensemble algorithms, the Random Forest model and Gradient Boosting estimator were used for model training.

Hyperparameter Tuning

The best hyperparameter combination was found to optimize the R2 metric. Hyperparameter tuning was done for Lasso and Ridge Regression, Decision Tree Regression, Random Forest model and Gradient Boosting estimator algorithms.

Results

The model was evaluated using three evaluation metrics. They are Root Mean Square Error (RMSE), Coefficient of Determination (R-squared: R2) Score and Explained Variance Score (EVS). Checked for overfitting cross-validated inputs with 10 folds with mean square error, No clear overfitting was noticed. Results were checked for several algorithms



	Root Mean Square Error (RMSE)	Coefficient of Determination (*R-squared* : R^2)	Explained variance score
Linear Algorithms			
Linear Regression	9.71115372445517e-05	0.9999999999999977	0.9999999999999978
Lasso	1.2870824931787181	0.9999995986099108	0.9999995992026574
Ridge	15.963534949971496	0.9999382535471568	0.999939794450767
Tree-Based Algorithms			
Decision Tree Regression	187.43520092409148	0.9914875183052737	0.9914947995081638
Ensemble Algorithms			
Random Forest Regression	139.91934150041263	0.9952563855432207	0.9953723044667285

Conclusion

The results for above models show that further improvement is needed. In general, usage of the raw data could be explored by adjusting features and assumptions about the inclusion and exclusion of certain parameters.We built a predictive model for casual bike rental volume using machine learning algorithms. This study suggests that it is possible to develop a reproducible and transportable predictive instrument for casual bike rental volume. Regression algorithms may be improved by using weights (weighted linear regression) and other available algorithms like linear, ensemble or tree algorithms.

References

https://www.jigsawacademy.com/popular-regression-algorithms-ml/

https://scikit-learn.org/stable/supervised_learning.html#supervised-learning