Prediction of Rental Bike Count Using Linear Regression,Decision Tree Regression,Ridge Regression and Random Forest Regression Techniques

Karthik Navin
MSc. Data Science
Coventry University

***Abstract*—This paper mainly focuses on predicting the count or number of bikes rented for each hour from the information that is available in the given dataset. The Bike sharing dataset contains the hourly count of rental bikes for two years (years 2011 and 2012) with the weather and seasonal information. The counts are predicted using regression machine learning models such as Decision tree, Linear regression, Ridge regression and Random Forest Regression techniques. The models are analyzed to find the best model that can accurately predict the target variable which is the demand for rented bikes.**

Keywords*—Regression, Machine learning, Python, Decision tree, Linear regression, Ridge regression, Random Forest Regression, Data preprocessing, Feature selection

I. Introduction

The bike rental system provides the public with an efficient and convenient mode of ttransportation. Since this is a popular method of transport the analysis of this data can be very useful. Customers can easily rent a bike from one location and return it to another location of his choice. This analysis can help in understanding the demand for bikes which can be used by bike rental organizations.

This count or demand can be predicted using different regression techniques. Regression models like linear regression,ridge regression,random forest regression and decision tree regression are used here. The best model is then selected by comparing the models with their corresponding performance metrics.

This research is useful for both data analysts and the bike sharing sector. We may learn which models are best for properly anticipating bike sharing demand by analysing the performance of the different regression models used. The outcomes of this paper can help bike sharing system operators efficiently allocate resources thereby improving customer happiness and increasing overall system efficiency.

II. Problem

The regression model designed is used in finding the total number of bike rentals from the given features and also finding which model predicts this target variable correctly. In this code, the target variable is the "cnt" variable, which is the total number of bike rentals rented out. The goal of this code is to create and test four regression models that is used to predict the total number of bike rentals based on the given data/features present in the dataset. The code reads a CSV file containing bike sharing data present in the ‘hour.csv’ file. It conducts certain data preparation processes, such as removing extraneous attributes such as 'immediate', 'dteday', 'casual', and 'registered'. The remaining characteristics are then scaled using the StandardScaler built-in function. The algorithm divides the data into training and testing data after pre-processing. Linear Regression, Decision Tree Regression, Ridge Regression and Random Forest Regression are the four regression models that it trains and assesses.

III. Dataset

Dataset was obtained from the UCI Machine Learning repository. It contains 17379 instances each containing 17 attributes with no missing values.

TABLE 1. DATASET FEATURES

No.	**Feature Name**	**Feature Description**	**Type**
1	Instant	No	numeric
2	dteday	Date	Date Time
3	season	Season (1:winter, 2:spring, 3:summer, 4:fall)	Numeric
4	yr	Year (0: 2011, 1:2012)	Numeric
5	mnth	Month	Numeric
6	hr	Time of rental	Numeric
7	holiday	Holiday	Numeric
8	weekday	Weekday	Numeric
9	workingday	Working day	Numeric
10	weathersit	Weather	Numeric
11	temp	Temperature	Numeric
12	atemp	Atmospheric Temperature	Numeric
13	hum	Humidity	Numeric
14	windspeed	Windspeed	Numeric
15	casual	Casual	Numeric
16	registered	Registered	Numeric
17	cnt	Count	Numeric

IV. Data Preparation

A. Scaling

Scaling is done so that the features or columns are compared to the same scale. Some features may have higher numerical value with different units that can interfere with the model prediction. Scaling is usually done using two methods - normalization and standardization.

Standardization is completed using python Standard Scaler class from scikit-learn library in the given code.

B. *Categorical Values*

Since we are performing regression on the given dataset, it is mandatory to ensure that all categorical features are converted to their respective numerical value without compromising the meaning of that value.This can be done by converting categorical values like seasons to numeric values from 1 to 4 ,with each number corresponding to the season it is referring to(season (1:winter, 2:spring, 3:summer, 4:fall)).

V. Feature Selection

This process of dropping or removing certain features is crucial for the regression model to work as expected. The below features have been removed from the dataset:

1.instant: This feature represents the ID of each record in the dataset. It does not provide any meaningful information to the target variable and provides no contribution to model prediction.

2.dteday: This feature represents the date of renting out each bike. Since this column contains the raw date of each data entry, the model can have overfitting issues that can directly affect the performance of each prediction.

Date features can have valuable information and is usually expected to extract relevant date columns that capture the underlying patterns rather than using the raw value of date itself.

3.casual and registered: These features represent the counts of casual and registered bike rentals.

The model is expected to find the total number of bike rentals ('cnt'), it is not necessary to include the casual and registered rentals count as separate features. These features are also removed to avoid information about the target variable 'cnt' in the model. By including the counts of casual and registered rentals as separate features, the model would have direct access to information that is part of the target variable itself. This could lead to overfitting and artificially inflating the model's performance.

VI. Feature Extraction

Here the different features available in the dataset are analyzed and compared to the target variable which is ‘cnt’.

A bar plot is generated using the seaborn library (sns) to visualize the relationship between the hour of the day (hr) and the count of bike rentals (cnt) based on the data in dataset.

From the graph a linear relationship exists between the target variable and the ‘hr’ column. This means that an increase in the ‘hr’ feature contributes to an increase in the target variable.

.

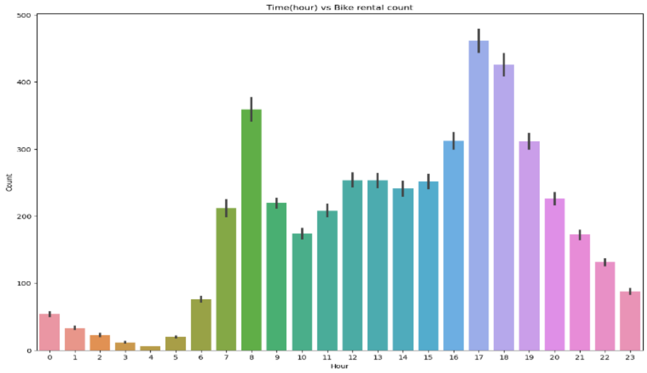


Fig. 1 Feature analysis- hr(time) vs Bike rental count

From the above graph the demand for rented bikes is more around the time from 5 and 6. This data is also useful for data analysis.

The below graph shows the demand for rented bikes is more for year 2 than year 1 i.e. The demand is higher in year 2012 than year 2011. This information shows that the demand increases as time increases.

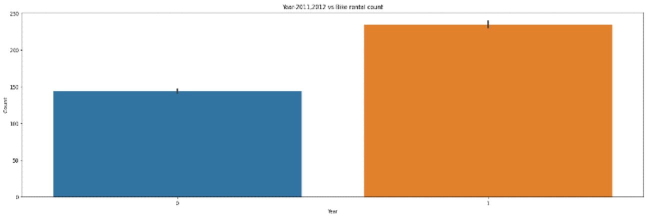


Fig. 2 Feature analysis- year(2011,2012) vs Bike rental count

The below graph shows the bar plot demand for rented bikes for different seasons and the demand is more during summer season.

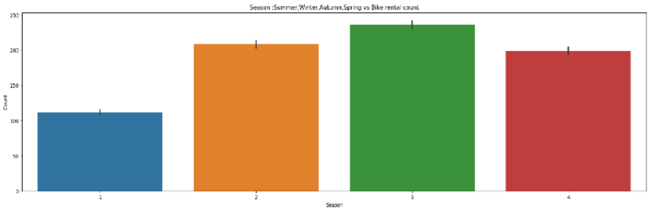


Fig. 3 Feature analysis- Seasons vs Bike rental count

The below graph shows the bar plot demand for rented bikes for different temperatures. From this graph the demand increases as the temperature increases showing a linear relationship between demand and temperature.

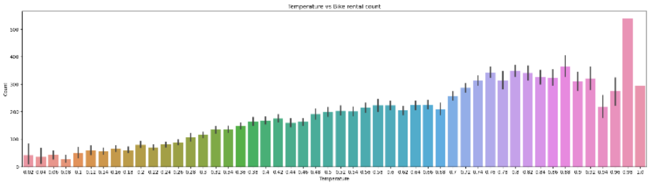


Fig. 4 Feature analysis- Temperature vs Bike rental count

VII. Machine Learning Regression Techniques

Considering the numerous features/factors that can influence the bike sharing demand such as weather conditions, time of day and day of the week, a model that can analyze the various factors available is required. Through this analysis, we hope to understand the factors that contribute to bike sharing demand.

We want to find the best regression model for accurately predicting the total count of bikes rented out by a customer. This can be done by comparing their corresponding model evaluation metrics.

1.Linear Regression:

Linear regression is the most basic and commonly used statistical modeling technique that aims to establish a linear relationship between the input features and the target variable. Since this model is basic it can cause the performance to be inferior to other regression models. It assumes a linear combination of the input features to predict the target variable. Here we can apply linear regression to the Bike Sharing Dataset and can be used to predict bike rental count/demand.

2.Ridge Regression:

Ridge regression is a linear regression model in which the coefficients are estimated using an estimator called the ridge estimator. It augments the conventional least squares objective function with a penalty component, which helps in decreasing the influence of features with high correlation with each other. Ridge regression is a more refined version of a linear regression model.

We will employ ridge regression on the Bike Sharing Dataset to examine its effectiveness in predicting bike rental demand.

3.Decision Tree Regression:

Decision Trees are used for regression and can be useful in datasets where the connection between the variables is discovered to be non-linear.

It is a model with three types of nodes in a tree structure. The Root Node is the starting node that represents the full sample and may be subsequently divided into sub nodes. Interior Nodes represent data set characteristics, whereas branches represent decision rules. Finally, the outcome is represented by the Leaf Nodes. This method is quite beneficial for dealing with decision-making issues.

4.Random Forest Regression:

Random forests, which are tree predictors, are used in regression machine learning. Each tree in the forest is based on a vector that was randomly selected and has the same distribution. The generalization error converges to a limit as the number of trees in the forest grows. The generalization error of the forest is determined by the strength of each individual tree and the association between them. The error rates attained by randomly picking features to divide each node are equivalent to Adaboost but more tolerant to noise. Internal estimates are used to track inaccuracy, strength, and correlation, revealing the effect of increasing the number of characteristics used for splitting. Furthermore, these internal estimations allow for the quantification of variable relevance.

VIII. Conclusion and Model Analysis

By analyzing the R-squared value and the Mean Square Value for all the models, it is seen that the Random Forest Regression model predicts the target variable with the highest accuracy and the lowest error compared to the other models.

TABLE 2. REGRESSION MODEL PERFORMANCE

Model	R Score	MSE
Linear Regression	0.40	19925.40
Ridge Regression	0.40	19925.46
Decision Tree Regression	0.89	3506.95
Random Forest Regression	0.94	1970.57

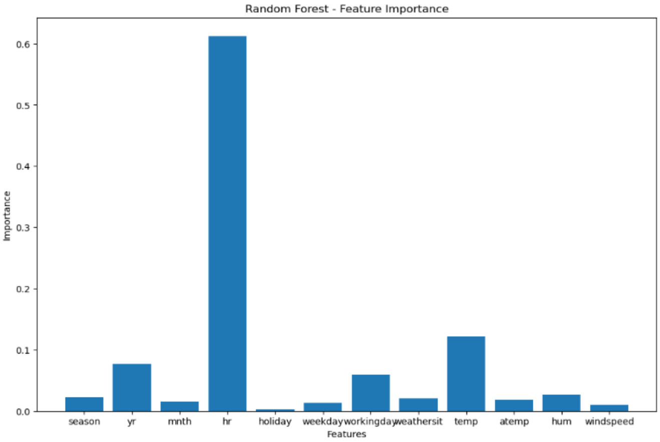
Linear Regression and Ridge regression have similar R-square values and mean square error. The r square value for both the models is 40% which means that both the models can predict only 40% of the target variable. The mean square error is also high and similar for both the models. This means that both the models will have high error in the predicted data.

The identical results between Linear Regression and Ridge Regression might be since Ridge Regression is a regularization approach that adds a penalty term to Linear Regression's model. When compared to the basic Linear Regression model, the regularization effect of Ridge Regression may not have a significant influence on model performance.

The Decision Tree Regression model has an r value of 0.89 and MSE of 3506.95 which is significantly better than Linear Regression and Ridge regression models.

The Random Forest Regression model has the highest R score - 0.94 which shows that it is the best model out of all the four models used.

Correspondingly, the Mean Square Error is also the lowest for Random Forest Regression model. This shows that this model predicts the data with 94 percent target data and the error while predicting the target feature will also be less.



The above graph shows how the different features affect the target variable prediction for random forest regression model.

The feature ‘hr’ is the main feature that is used by this model and the change in this feature contributes more to the change in the predicted target variable. This means that the demand for rented bikes is highly dependent on the time of day.

The features ‘hr’ is followed by ‘temp’ and ‘yr’. From the analysis carried out during feature extraction, the demand for rented bikes is highest when the time is around 5pm – 6 pm, year is 2012 and temperature is high.

IX. References

[1] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[2] UCI Machine Learning Repository. (n.d.). UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset.

[3] Huang, J. C., Ko, K. M., Shu, M. H., & Hsu, B. M. (2020). Application and comparison of several machine learning algorithms and their integration models in regression problems. Neural Computing and Applications, 32, 5461-5469.

X. APPENDIX

Source Code: Both source code and csv file is available in the GitHub link given below.

GitHub link: https://github.com/KarthikNavin/Machine-learning-Coursework.git