Bike Sharing Demand Prediction

V. Bhavya Reddy

Abstract—**From a business perspective, predicting the number of rented bikes is a crucial part of the revenue generation process because having excess bikes results in resource waste (bike maintenance, bike parking, and security costs), and having fewer bikes leads to revenue loss (losing customers due to non-availability). By estimating the bikes to be rented, the company can work more efficiently.**

The goal of this project is to provide a model to predict stable supply of bike rentals to predict demand at any hour. This dataset contains hourly bike rental counts along with weather information and dates spanning an entire year.

Index Terms—**Linear Regression, Lasso, Rented Bike Count, Ridge, Random Forest Regression, Decision Tree.**

I. Problem Statement

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important 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.

A. Data Description

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

B. Attribute Information:

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - Humidity as
- 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)

Data consists of 14 columns variables and 10 of them are Numeric Columns.

II. Introduction

This project is approached in the following way:

We begin by reviewing the data set overview, in which we briefly analyze the observations and note several characteristics of numeric and categorical. In addition, we find the number of NULL values - none were found.

The next step is Exploratory Data Analysis (EDA) on the dataset, where we analyze numeric and categorical features through bar charts, box plots, countplots, etc. We also use heatmaps to analyze correlation, and VIF calculations are carried out to check for multicollinearity.

In feature Engineering, we eliminate the outliers and drop unnecessary columns, and modify the dataset. Data encoding followed with train-test split are performed on the data set.

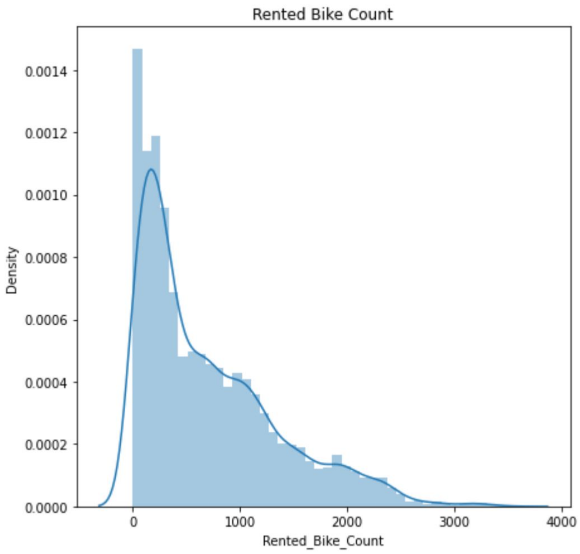
Using train data we train our models using Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, and Random Forest Regression. So, a total of five regression models were used to predict the rented bikes per hour: Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, and Random Forest Regression. We found that the Random Forest Model had the best/lowest RMSE of all the models. Ridge Regression and Lasso Regression models did not provide any improvement over Linear Regression. In order to gain a better understanding of the data, several plots are drawn (MSE, R2 score, Actual vs predicted, important features, etc.).

III. Exploratory Data Analysis

Before we do the modelling, let’s look at the important features and their effect on the bike rent count.

A. Rented Bike Count

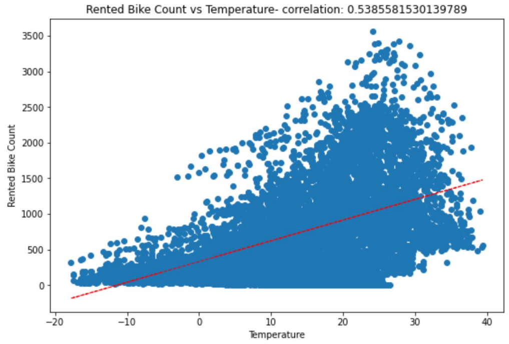
Signifies the count of bikes rented at each hour. The target column (Rented Bike Count per hour) ranges between 0 and 3556 over the 1 year span. Mean of Rented Bike Count = 704.6, with median and 75% percentile = 504.5 and 1065.25, respectively. This suggests that the ’Rented Bike Count’ distribution is more denser at lower values. This is expected as out of 24 hours, we would expect the bike demand/usage to be high for maximum of maybe 6 hours or so.

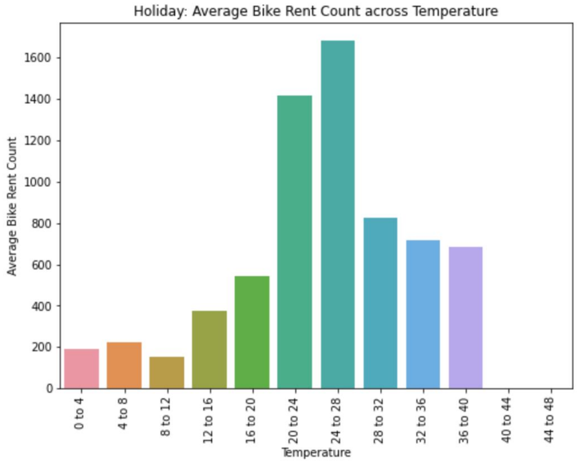


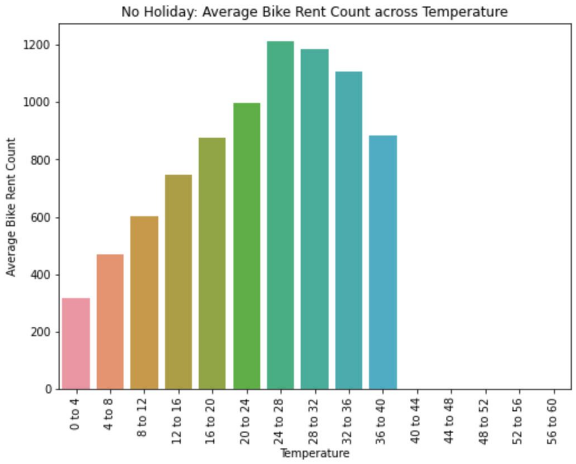
B. Temperature

People generally prefer to bike at moderate to high temperatures. We see highest rental counts between 20 to 32 degree Celsius. However, at a very high temperature the bikes rented count will decrease.

Temperature will turn out be the most important feature while training the model.







C. Humidity - m/s, Rainfall - mm, Snowfall - cm

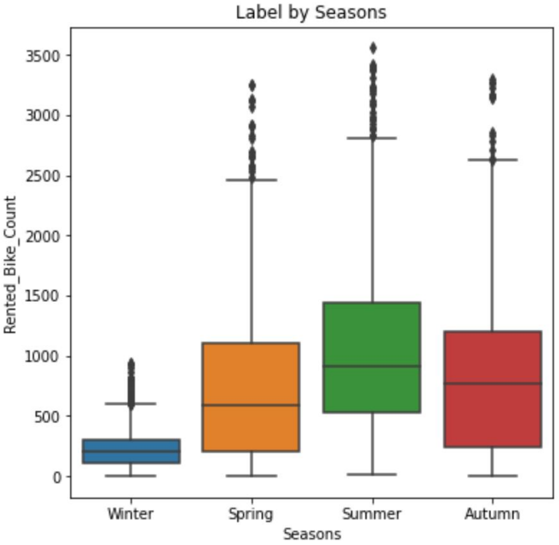
These have a negative correlation with bike rents.

D. Windspeed - m/s, Visibility - 10m, Solar radiation - MJ/m2

These have slight positive correlation with bike rents.

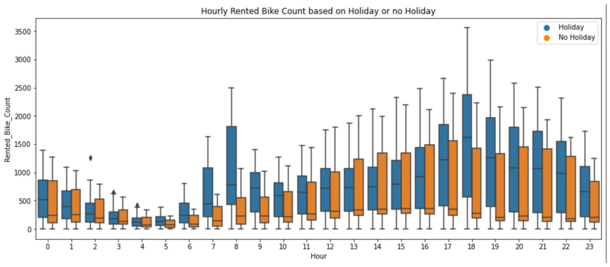
E. Seasons

We see the highest number bike rentals in Summer Season and the lowest in Winter season. We notice many outliers in each season.

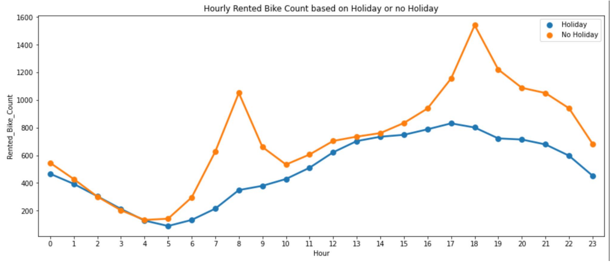


F. Hour

Very few number of outliers can be seen in the seaborn box plot across hours.

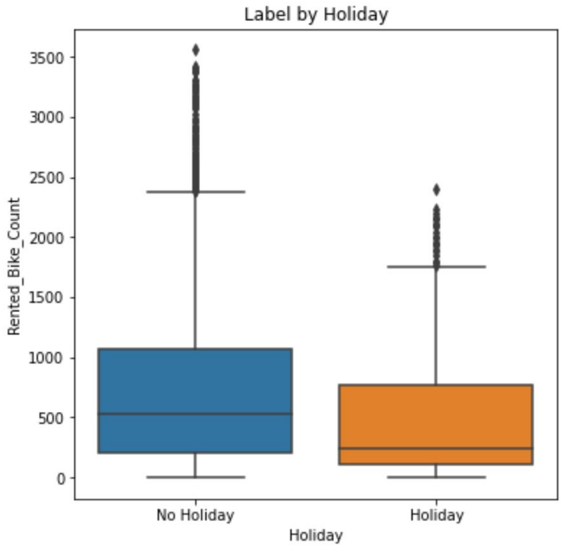


Higher reservations can be seen at around 8am and 6pm (office hours) and very low reservations at very early in the morning. No Holiday: There is a peak in the rentals at around 8am and another at around 6pm. Holiday: There is more or less a uniform rentals across the day with a peak at during 1pm to 5pm. These correspond to probably tourists.



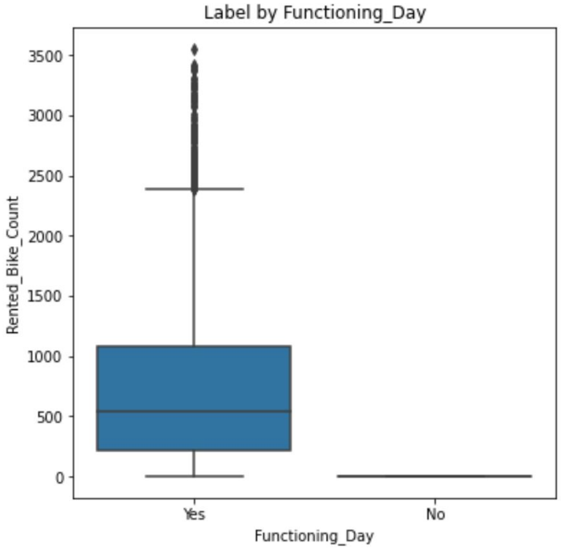
G. Holiday

More bikes are rented on working days.



H. Functional Day

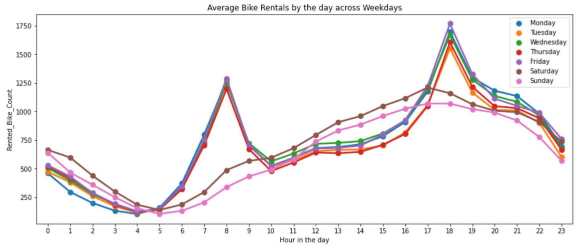
No bikes are rented on non-functional day.



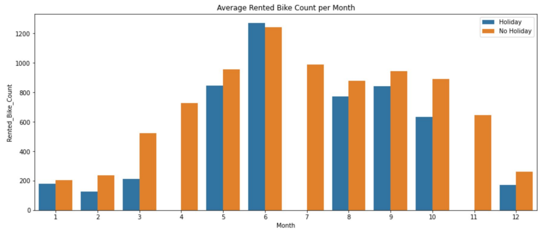
I. Date

From the Date feature we extract the month, day of month, and day of week; out of which only month seems to have a decent effect on the target variable(rented bike count).

Monday to Friday seem to have similar an effect on the rented bike count and on Saturday, Sunday the effect on the rented bike count is slightly different.

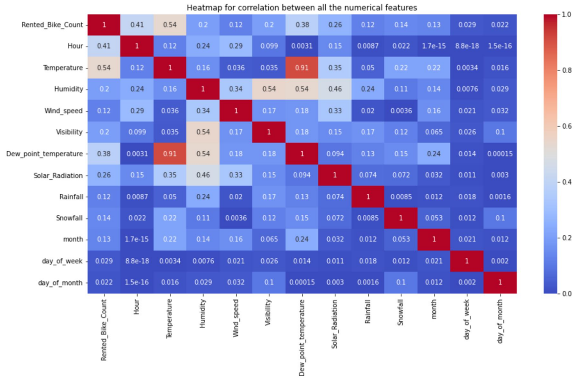


we can see that we have highest bike rents in June. No holidays in 4th, 7th, and 11th month.



IV. Correlation Analysis

Below is a heatmap plot of the correlation between all the numerical columns.



We can infer the following from the above heatmap:

- Temperature and Dew point temperature are highly correlated.
- We see a positive correlation between Rented Bike Count and Temperature (as seen in the scatter plot). This is probably only true for the range of temperatures provided.
- There is a negative correlation between Rented Bike Count and Humidity.
- The more the humidity, the less people prefer to ride. Not a great amount of correlation between humidity and temperature.
- Rented Bike Count has a weak dependence on day_of_month, day_of_week.

V. Data Cleaning

We have earlier mentioned that we have no ’NaN’ values in the provided dataset.

A. Functioning Day

A non-functioning day had no bikes rented, and this feature tells us whether any bikes are rented or not. This feature does not have an impact on the bikes rented for every hour, so it shall be removed. On the non-functioning day, no bikes will be available, so we will remove those observations as well.

B. Removing Outliers

There were a lot of outliers in holidays, and a very few in hour features. Therefore, we shall remove them by computing zscores.

VI. Feature Engineering

Few features were removed, modified, and added to the provided dataset. Below is the feature engineering carried on the provided dataset.

- The Date column which contained the date in ‘yyyy-mm-dd’ format was split into individual [‘month’, ‘day_of_month’, ‘day_of_week’] categorical columns.
- Drop ’Seasons’ column: This is because month column has a direct mapping with season (Winter:December, January, and February, Summer: June to August, Autumn: September to November and Spring: March to May).
- Drop ’Functioning Day’ column: The Functioning Day column had information about the bikes if they are rented.
- Drop ’Date’ column: Intuitively, there is should be no dependency on date. Hence drop this column.
- Drop ’Dew point temperature’ column: Temperature and Dew point temperature are very highly correlated and essentially indicate the same thing. Hence retain only the temperature column.
- One-Hot Encoding of categorical feature set:
  - Hour: Split Hour column to Hour_0, Hour_1, ... Hour_22. Drop Hour and Hour_23 columns since it is a function of the rest of the retained Hour columns.
  - Month: Split month column to month_1, month_2, ..., month_12. Drop month_12 and month columns since they are a function of the rest of the retained month columns.

VII. Modelling

After the feature engineering, we split the data into Train set and Test set. Using MSE, RMSE, R2score, Adjusted R2 as our evaluation metric, we compare various models and select the regression algorithm based on the lowest RMSE on the Test data.

A. Train/Test Split

Data is randomly spilt into 80% of train data and 20% of test data.

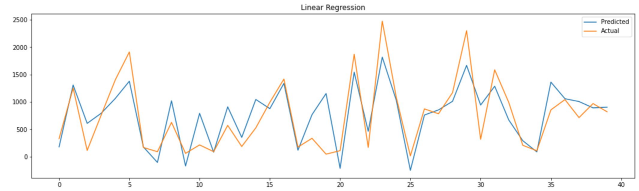
B. Linear Regression

For Linear Regression, we use the entire set of features obtained via OneHotEncoding. The various performance metrics are given in the below table.

MSE on train data	114837.751
MSE on test data	119490.275
R2 score	0.685
Adjusted R2	0.677

RMSE on the test data and training data are almost the same. Linear Regression model is definitely not an overfit model.

Below plots the actual bikes rented value vs. the model predicted value for a few data points in the test data set.



Overall a decent initial model.

C. Lasso Regression

For Lasso Regression too, we use the same dataset as Linear regression.

We tune the hyperparameter α = 0.2 with GridSearchCV of 5-fold cross validation to train the model with training dataset.

The various performance metrics are given in the below table.

MSE on train data	114842.704
MSE on test data	119547.598
R2 score	0.685
Adjusted R2	0.677

The metrics from linear regression and lasso regression are not different.

D. Ridge Regression

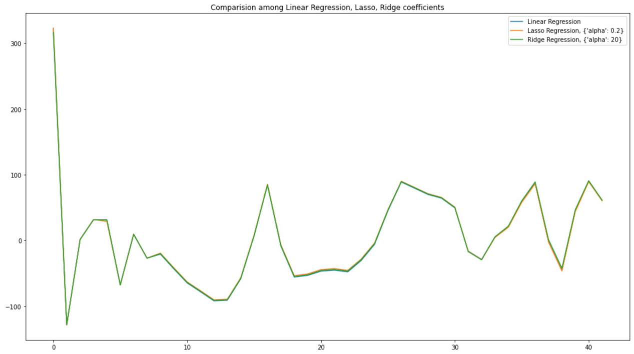
We tune the hyperparameter α = 20 with GridSearchCV of 5-fold cross validation to train the model with training dataset.

The various performance metrics are given in the below table.

MSE on train data	114843.321
MSE on test data	119420.024
R2 score	0.685
Adjusted R2	0.677

The metrics from linear regression, lasso regression and Ridge regression are not in anyway different.

coefficient comparison is shown in the figure below:



Here, we shall notice that there is no difference by regularising Linear Regression.

E. Decision Tree Regression

We tune the hyperparameter min_samples_leaf to 21 and train the model.

The various performance metrics are given in the below table.

MSE on train data	57900.108
MSE on test data	85226.86
R2 score	0.758
Adjusted R2	0.752

The model seems to be overfitted.

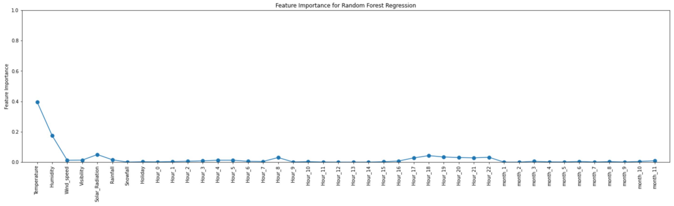
F. Random Forest Regression

Best parameters for Random Forest Regression Model: 'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 0.0001, 'n_estimators': 5000

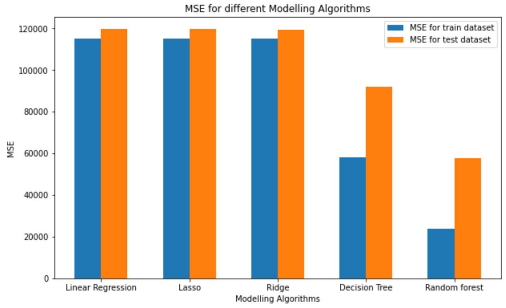
The various performance metrics are given in the below table.

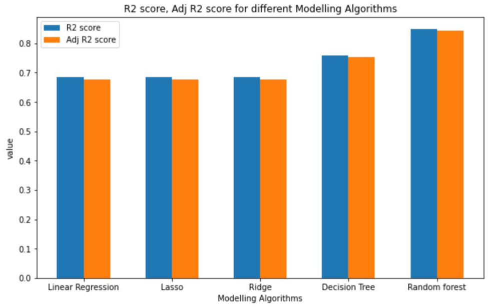
MSE on train data	23857.452
MSE on test data	57705.602
R2 score	0.848
Adjusted R2	0.844

The model seems to be overfitted model. Below is a feature importance plot for RF2 model.



Below we summarize various metric analysis for all the earlier models through bar charts.





CONCLUSION

Data Exploration Conclusion:

- Temperature: People generally prefer to bike at moderate to high temperatures. We see highest rental counts between 20 to 32 degree Celsius.
- Humidity: With increasing humidity, we see decrease in the number of bike rental count.
- Hour: Bike rental count is mostly correlated with the time of the day. As indicated above, the count reaches a high point during peak hours on a no holiday and is mostly uniform during the day on a non-holiday.
- Temperature, Windspeed, Visibility, Solar radiation: They have a positive correlation with bike rents.
- Rainfall, Snowfall: They have a negative correlation with bike rents.
- Seasons: We see highest number bike rentals in Summer and the lowest in Winter season.

Modeling Conclusion:

- We use 5 Regression Models to predict the hourly rented bike count - Linear Regression, Lasso, Ridge, Decision Tree, Random Forest.
- Among all the 5 models, Random Forest Model has the best metric analysis.
- Lasso or Ridge regularisation did not provide any improvement to the regular linear regression.