Improvement Of Seoul Bike Sharing Linear Model Using Feature Engineering, Model Selection And Gradient Descent

Ashish Alden Dsouza, Aashish Pathak, Rajat Kumar Sahu, Santosh Susarla

Abstract—**This project presents a Improvement Of Seoul Bike Sharing Linear Model Using Feature Engineering, Model Selection And Gradient Descent. Nowadays, rental bikes are introduced to the public in almost all the top urban cities in the world for enhancement of mobile comfort ability and environmental factors [1]. The most important part is the prediction of bike counts or bike availability at a certain point at an hourly basis for stable supply of rental bikes. The dataset includes weather information like temperature, humidity, wind speed, visibility, dew point, solar radiation, rainfall etc. This project explores the application of Linear Regression models for predicting bike count for meeting the necessary demands. This paper discusses the models for hourly rental bike demand prediction.**

I. INTRODUCTION

Currently the bike renting scheme is well accepted by the public in the top urban parts of the world. Most of these bike rental companies allow people to borrow and return a bike from a common bike rental station. Also some of the bike rental companies give the option of dropping off the bike at whichever end point or destination the user wants [1]. For expanding availability of bicycles for public use, the operators running this service allocate a truck that collects bicycles parked in various stations and relocate them to the original station gradually. These rental bikes obviously come with a GPS tracking system for the user to know at which points on the map the bikes are available and for the company to know where the bikes are dropped off so that they can relocate it to the most demanding stations for the customers. The bike rental has helped see a lot of positive environmental impact with betterment of physical human activity mentally and physically. Which is one of the few reasons why this bike sharing concept is a trend and quite stable business.

Why a Bike Sharing system? A bicycle-sharing system, public bicycle scheme, or bike-share system, is a service in which bicycles are made available for shared use to individuals on a short term basis for a price or free. Advantages of Bike Sharing are-improved Distribution Ease of Installation- the installation of these bikes stations can be easily installed anywhere in the city at a lot lesser price. Powering Stations – There are also several charging stations across the city for charging of powered bikes.

Tracking:

Pedal Assistance – for senior citizens or people with any physical disability this motor powered pedal assistance is a gem as it reduces their load of pedaling

Environmental Friendly- As per the figure mentioned the bike is a clean and green way for a green and pollution free city.

II. Literature Survey

In the bike renting business world a number of studies have been carried out and various different predicting models have been built. Research and studies were based on the different categories of bikes, documentation process, user analysis etc. In every business we may see a downfall in sales, the same has happened in this industry. So real time data was collected using automated computerized devices for gathering the data which can be used to study in the near future. This historical data which is collected is then studied and various predicting models have been built which are discussed in brief below. Some of these famous models have helped in the business in this industry [1].

Vogel & Mattfeld, in the year 2011 collected information which were recorded by bike sharing frameworks to predict the medium-term request. In order to enhance operational planning and

[1]

operational cost. In the first generation, free bikes were given to people. There were different types of bicycle which were different in functioning. Bikes were not locked and there were no particular stations to park it. The second generation, coin deposit systems were introduced. It was important to lock the bicycles, so it was provided with it. There were different types of bicycle and it was parked in a station. The third generation, smart cards were introduced. There were access booths from where one can get different types of bicycle. It was important to lock the bicycles, so it was provided with it and was parked in a station. The fourth generation, E-bicycles were introduced which had GPS tracking systems. Real time availability was there which made bicycles available immediately. Smart card system was followed and parking stations were available. The fifth generation E-bicycles along with a GPS tracking system made the parking easy. It made data management possible which made the planning, exploring, accessing easier than ever. In the meantime a GPS based technique was studied and a model was built which would predict at which bike sharing stations what is the demand. This was researched by Garcia-Palomares, Gutierrez; in the year 2012 [3] .

Erdogan, Battarra, & Calvo, [4] in the year 2015 developed an algorithm which focuses on reaching the destination inventory for a total of bicycles in individual stations. This study has shown ways to deal with enhancing the bike station areas and strategies for stock . Directing the trucks to redistribute bikes considering spatial and temporal variations.

Russell &Norvig; [5] in the year 2016 introduced the concept of artificial neural networks. It has been utilized to forecast a wide variety of regions. Here artificial intelligence and data mining techniques have been used to increase accuracy thereby increasing the cost which was the main drawback of this concept.

Later in the year 2017 had come up with a model which would predict the more accurately and in real time where and how many bikes are required at bike stations. Which will fulfill the client's request and simultaneously reduce the operational cost.Gao & Lee [6] Lee in the year 2019 on the moment based demand in public.

This bike rental and sharing industry is constantly changing with time. The predicting model used in the past becomes exceptionally unpredictable and is influenced by many external factors. In our research we study the issue of predicting and the public rental bike demand in a bike sharing framework utilizing a data mining approach considering weather information and building an OLS model.

Quick Fact-Paul DeMaio [7](paul@metrobike.net) has been involved in bike-sharing since 1996 as an undergraduate student in Copenhagen, Denmark. In 2005, he created Metro Bike, LLC to focus on bike-sharing and bike transportation planning. He has a Bachelor of City Planning from the University Of Virginia School Of Architecture and a Master of Transportation Policy, Operations, and Logistics from the George Mason University School of Public Policy. He is also the author of The Bike-sharing Blog (bike-sharing.com), an international news resource about the field.

In the first generation, free bikes were given to people. There were different types of bicycle which were different in functioning. Bikes were not locked and there were no particular stations to park it. In the second generation, coin deposit systems were introduced. It was important to lock the bicycles, so it was provided with it. There were different types of bicycle and it was parked in a station. In the third generation, smart cards were introduced. There were access booths from where one can get different types of bicycle. It was important to lock the bicycles, so it was provided with it and was parked in a station. In the fourth generation, E-bicycles were introduced which had GPS [8] tracking systems. Real time availability was there which made bicycles available immediately. Smart card system was followed and parking stations were available. In the fifth generation, E-bicycles along with a GPS tracking system made the parking easy. It made data management possible which made the planning, exploring, accessing easier than ever.

Dataset Description

We have sourced the data from UCI Machine Learning Repository.

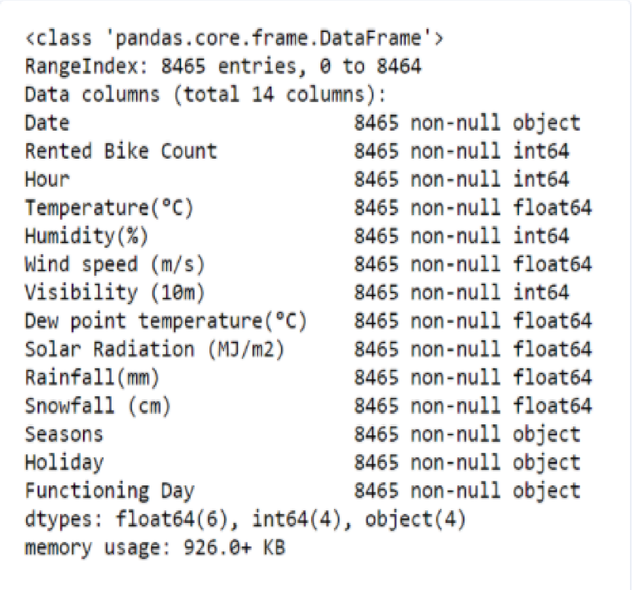
The dataset contains the count of public bikes rented each hour in Seoul Bike Sharing System with the corresponding weather data and holidays information used include weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information. Since any form of transportation mainly depends upon the Climatic conditions, the corresponding weather information such as Temperature, Humidity, Wind speed, Visibility, Dew point temperature, rainfall, and snowfall for each hour is added. The processed data consists of the total count of rental bikes rented at each hour with date/time variable and Weather information.

Methodology

Linear regression - Linear regression Linear regression (LM) in the most simplest method, that is equated with the relationship between the Y attribute of the scalar output and one or even more X attributes of the input quantity. The case of an independent attribute is known as simple linear regression, and the method is called as multiple linear regressions when more than one independent attributes are considered. Data is designed using linear predictor functions in linear regression, and from data, the unknown model parameters are estimated. Usually, Linear regression refers to a system where the conditional mean of Y is an affine function of X, given the value of X. The model is assumed as in Eq. (1). $Y = \beta 0 + \beta 1X + s$ (1) Here $\beta 0$ and $\beta 1$ are two unknown constants representing the intercept and slope, also known as parameters or coefficients, and s is the term of error.

Data understanding & Visualization of the analysis done

Data pre-processing



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8465 entries, 0 to 8464
Data columns (total 14 columns):
Date                         8465 non-null object
Rented Bike Count            8465 non-null int64
Hour                         8465 non-null int64
Temperature(°C)              8465 non-null float64
Humidity(%)                  8465 non-null int64
Wind speed (m/s)             8465 non-null float64
Visibility (10m)             8465 non-null int64
Dew point temperature(°C)    8465 non-null float64
Solar Radiation (MJ/m2)      8465 non-null float64
Rainfall(mm)                 8465 non-null float64
Snowfall (cm)                8465 non-null float64
Seasons                      8465 non-null object
Holiday                      8465 non-null object
Functioning Day              8465 non-null object
dtypes: float64(6), int64(4), object(4)
memory usage: 926.0+ KB
```

Fig 1:- checking the data types

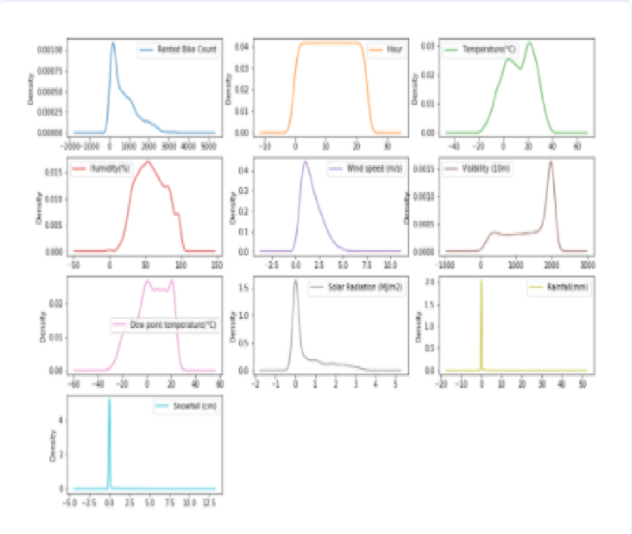


Fig 2: - analyzing different features

After loading the dataset, we checked the shape of the data and found the shape to be (8760, 14) , then we checked the data type of all the variables. We checked the null values if any and found no null values in our data. Also we found the date to be unique so we will drop it later, following which we checked the distribution of the categorical variables. We checked the rented bike count if zero for any rows and found 5 rows having zero bike count. Then we checked the count of functioning day and non functioning day. And found out that there are 8465 functioning days and 408 holidays. Following which we checked the correlation to see if any variables are correlated with each other.

We segregated the categorical and numerical variables and on grouping the Hour and Rented bike count we found the average bikes required every hour. We plotted a KDE plot and we can say that the data is right skewed also maximum values lie between 0 and 1000.Following which we plotted all the numerical variables. After plotting a count plot for the rented bikes across holidays and non-holidays we saw that the bike count is comparatively a lot more on 'No Holiday' than during a holiday.

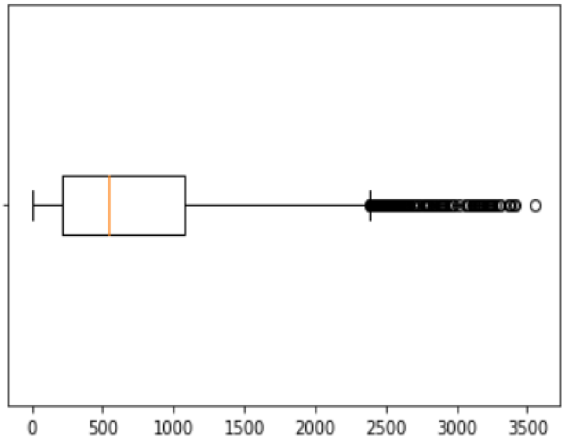


Fig 3: - Box plot showing outliers in data

We plotted a box plot for the rented bike count and found a lot many outliers in the data. Following which we plotted a heat map and found that the dew point temperature and the Temperature have a very high positive correlation.

Box plot is a standard way of presenting the data distribution. And, it is used to see how tightly the data is grouped and to visualize the skewness in the data. As seen from Figure 3, A box plot is composed of Lower whisker, upper whisker, lower quartile, upper quartile and middle quartile (median). Using these four quartiles groupings within the are made. Each group has 25% of data. The middle quartile (median) represents the midpoint of the data, and it is represented by the line, which divides the box into two parts. The green box denotes the 50% of middle values in each data field considered. This range of values from lower to upper quartile limit is known as interquartile range. 75% of values fall below the upper quartile and 25% fall above it. And, the values above the upper whisker are considered as outliers. . In this study, a box plot-based visualization is utilized to study the data better.

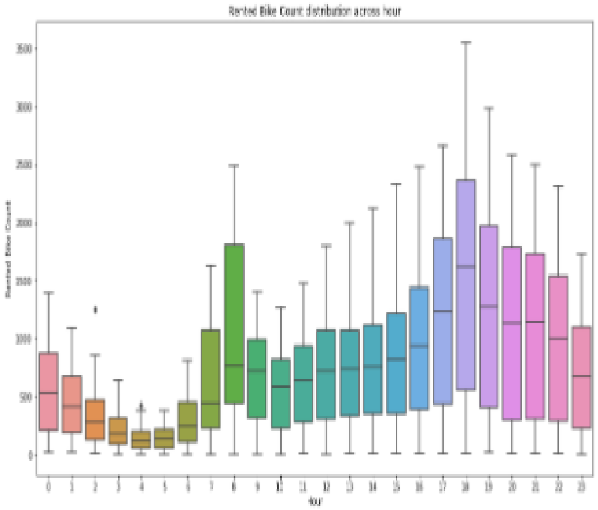


Fig 4:- Bike count across different hours of the day

As per the below box plot, we can see that more number of bikes have been hired on a working day instead of a holiday.

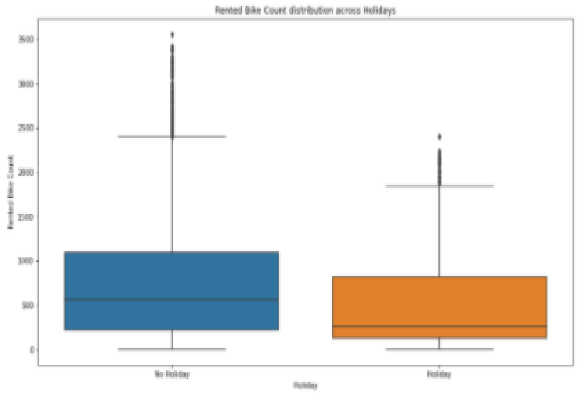


Fig 5 :- Box plot for bike count on no holiday and holiday

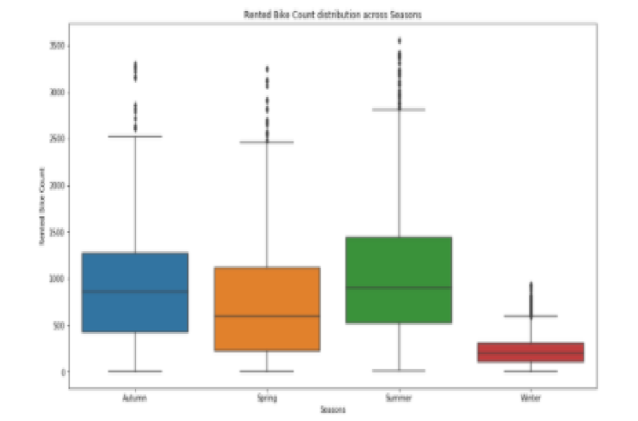


Fig 6: Box plot

As per the above plot we see that in winter least number of bikes have been hired and for the remaining seasons the bike count is almost the same.

Following which we plotted a graph for the count of bikes rented every hour across different seasons. And we can see that Winter has bike counts. We applied One way and found out that the p-value is far less than 0.05 for holiday and non holidays. This indicates there is a significant difference in the mean of the rented bike count across holidays.

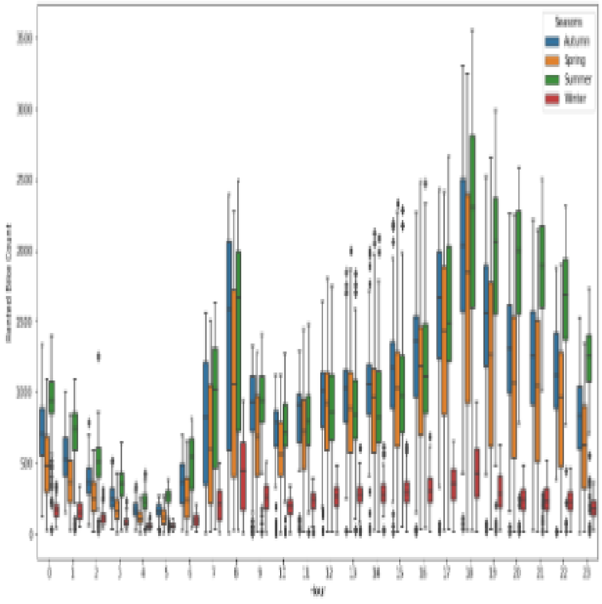


Fig 7: - Combination of different boxplots

Fixing Outliers:

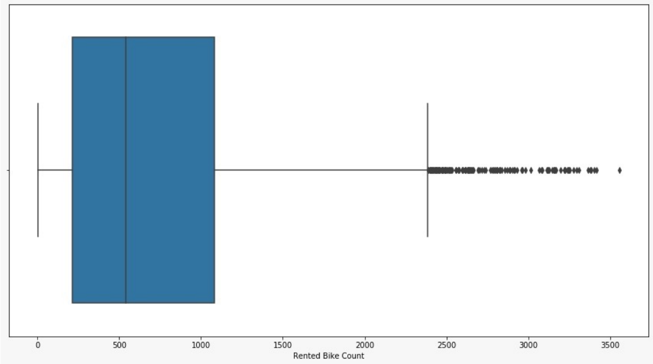


Fig 8:- Fixing outliers using box plot

Previously we saw there are several outliers in our dataset so to remove them we dropped the date as it was insignificant. So as per the plotted graph in the Jupyter notebook we can see that the maximum outliers are present in the rented bike count and for wind speed, rainfall, snowfall are equal to 0, hence we dropped rented bike count and visibility.

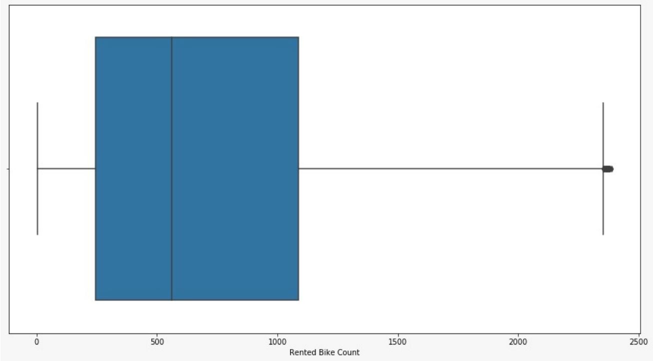
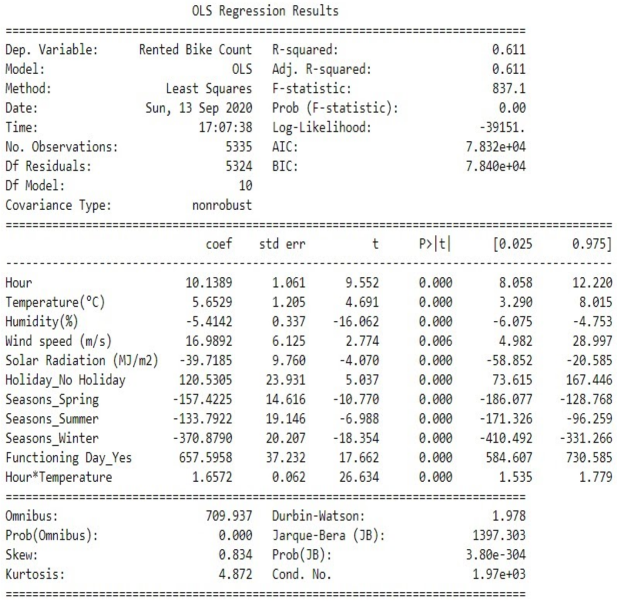


Fig 9: - Box plot after outlier correction

We can see from the above box plot that now the outliers have been reduced after applying the above technique.

Model Building:

A power transform will make the probability distribution of a variable more Gaussian. This is often described as removing a skew in the distribution, although more generally is described as stabilizing the variance of the distribution. Power Transformer is used to make the data distribution more-Gaussian and standardize the result, centering the values on the mean value of 0 and a standard deviation of 1.0.We applied the power transformation technique with power transformation technique, from the r squared and adj r squared values are almost equal hence we can again say that our model is predicting correctly also the RMSE value had slightly reduced. Following is the OLS model we obtained for power transformation technique.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Rented Bike Count   R-squared:                       0.611
Model:                            OLS   Adj. R-squared:                  0.611
Method:                 Least Squares   F-statistic:                     837.1
Date:                Sun, 13 Sep 2020   Prob (F-statistic):               0.00
Time:                        17:07:38   Log-Likelihood:                -39151.
No. Observations:                5335   AIC:                         7.832e+04
Df Residuals:                    5324   BIC:                         7.840e+04
Df Model:                          10
Covariance Type:            nonrobust
==========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Hour                       10.1389      1.061      9.552      0.000       8.058      12.220
Temperature(°C)             5.6529      1.205      4.691      0.000       3.290       8.015
Humidity(%)                -5.4142      0.337    -16.062      0.000      -6.075      -4.753
Wind speed (m/s)           16.9892      6.125      2.774      0.006       4.982      28.997
Solar Radiation (MJ/m2)   -39.7185      9.760     -4.070      0.000     -58.852     -20.585
Holiday_No Holiday        120.5305     23.931      5.037      0.000      73.615     167.446
Seasons_Spring           -157.4225     14.616    -10.770      0.000    -186.077    -128.768
Seasons_Summer           -133.7922     19.146     -6.988      0.000    -171.326     -96.259
Seasons_Winter           -370.8790     20.207    -18.354      0.000    -410.492    -331.266
Functioning Day_Yes       657.5958     37.232     17.662      0.000     584.607     730.585
Hour*Temperature            1.6572      0.062     26.634      0.000       1.535       1.779
==============================================================================
Omnibus:                      709.937   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1397.303
Skew:                           0.834   Prob(JB):                    3.80e-304
Kurtosis:                       4.872   Cond. No.                     1.97e+03
==============================================================================
```

Fig 10:- OLS summary

After power transform we try to fit using the fit function to check the linear regression.

Then we separated the significant features from the significant features by finding the p-value.

We got 'Visibility (10m)' and 'Dew point temperature (°C) as insignificant variables following which we checked how the model predicts with and without transformation technique. Hence, with power transformation technique, from the r squared and adj r squared values are almost equal hence we can again say that our model is predicting correctly also the RMSE value has slightly reduced as compared to the previous models. Going forward scaling the data as it is found in different units according to different features. So we checked from the score card that we see that linreg_model_data_scaled_without_insignificant model predicts correctly but is performing poorly compared to the power transform model. Then we checked for interaction based models and found that Based on the above models comparison with the linreg_model_data_interacted we can see that it is performing better than other models except the power transformation model.

Check for assumptions:

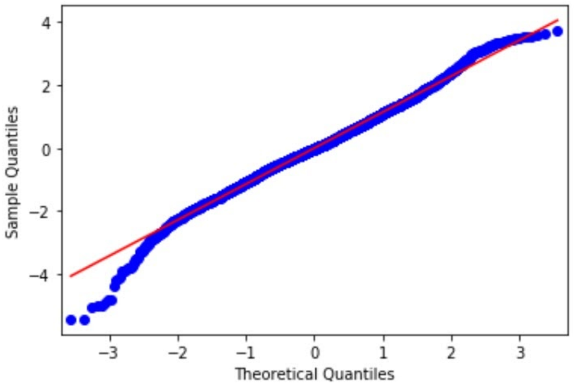


Fig: -11 Line graph showing fitting model

So running the linear regression full model with Power transformation again because it has the highest value for r_square and adj_r_square and lowest RMSE.We checked multicollinearity so as to check that the variables are not dependent on each other and are independent. Also checked the Variance Inflation Factor for all features and found that that the VIF for all the above features is less than 15, hence we conclude that all the features are independent.

We checked the linearity using scatter plots and could see some linearity. Similarity we checked the linearity for all features and saw that none of the plots show specific patterns. We may conclude that the variables are linearly related to the dependent variable. We used a linear rainbow function to find that the data is 8.9 % related; hence linearity exists in this model.

Following which we checked Breusch-Pagan which is the test for detecting heteroskedasticity: The null and alternate hypothesis of Breusch-Pagan test is as follows: H0: The residuals are homoscedastic, H1: The residuals are not homoscedastic.

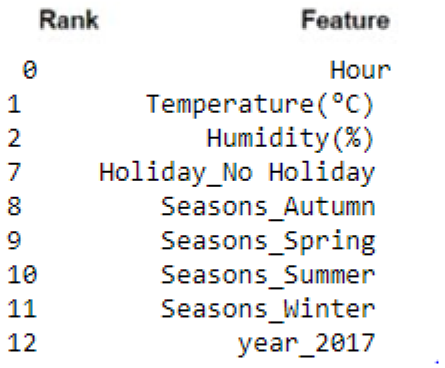
We observe that p-value is less than 0.05 and thus reject the null hypothesis. We conclude that there is heteroskedasticity present in the data. In real life it might not be possible to meet all the assumptions of linear regression. The mean of the residuals is very much closer to zero. Therefore, we can say that linearity is present, we applied jarque_bera method and found that the data is not normally distributed .We applied Ridge, Lasso and Elastic -net technique and can see that the values are closer to each other hence we can say our assumptions are correct. We also checked Linear Regression with Stochastic Gradient descent using GridSearchCV and without and see that the model is not performing well in Stochastic Gradient Descent.

Feature engineering is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms. Feature engineering can be considered as applied machine learning itself. A feature is an attribute or property shared by all of the independent units on which analysis or prediction is to be done. Any attribute could be a feature, as long as it is useful to the model.

The purpose of a feature, other than being an attribute, would be much easier to understand in the context of a problem. A feature is a characteristic that might help when solving the problem

Feature Selection (Statistical significance)- Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features. Selecting the right features gives the following advantages:-· Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.· Improves Accuracy: Less misleading data means modelling accuracy improves.· Reduces Training Time: fewer data points reduce algorithm complexity and algorithms train faster.

We used Forward selection method and RFE-Recursive feature elimination technique and found the top significant features rank wise for which as per the figure below Hour and Temperature stand First and second respectively.



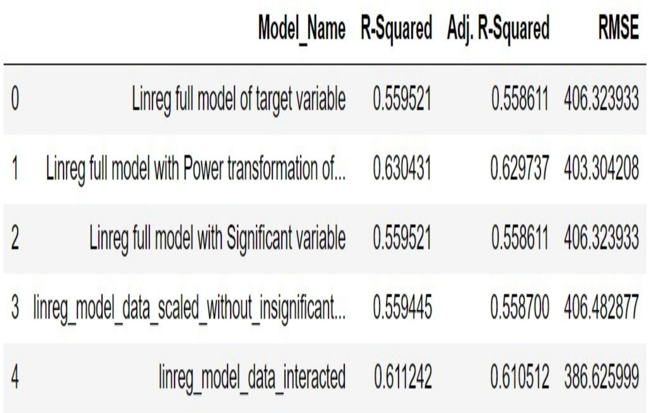
```
Rank              Feature

 0                   Hour
1          Temperature(°C)
2              Humidity(%)
7      Holiday_No Holiday
8          Seasons_Autumn
9          Seasons_Spring
10         Seasons_Summer
11         Seasons_Winter
12              year_2017
```

Fig:-12 Feature ranked as per decreasing priority

Model evaluation – Based on the scorecard in the figure below we exclude 'Linear Regression SGD' and 'Linear Regression SGD' using best parameters models since its output are uninterruptable. Evaluation indices- Multiple evaluating criteria are used for comparing the performance of regression models. The performance evaluation indices used here are: Root Mean Squared Error (RMSE R2 and adjusted R2. RMSE stands for the sample standard deviation of the residuals between the observed and the predicted values. Large errors can be identified using this measure and the fluctuation of model response regarding variance can be evaluated. RMSE metric is a scale dependent evaluation metric, and it produces values with identical units of the measurements. R2 is called as the coefficient of determination, with values ranging from 0 to 1, denoting the goodness of a prediction model fit. A high value of R2 denotes the predicted values exactly fit the observed values. Figure shows the formula written for R2, Adjusted R2 and RMSE.

Table I : Model Building



	Model_Name	R-Squared	Adj. R-Squared	RMSE
0	Linreg full model of target variable	0.559521	0.558611	406.323933
1	Linreg full model with Power transformation of...	0.630431	0.629737	403.304208
2	Linreg full model with Significant variable	0.559521	0.558611	406.323933
3	linreg_model_data_scaled_without_insignificant...	0.559445	0.558700	406.482877
4	linreg_model_data_interacted	0.611242	0.610512	386.625999

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

$$\text{Adjusted } R^2 = 1 - \frac{SS_{residuals}/(n-K)}{SS_{total}/(n-1)}$$

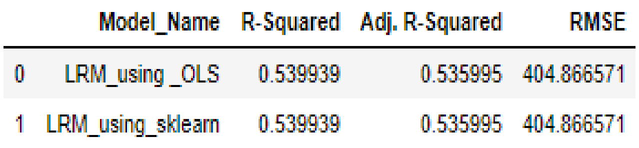
Fig :-13 Calculation For Adjusted R2, RMSE and R2.

1). Model Building Using OLS Method and Sklearn Package [10] :

Ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function.

Upon the completion of EDA we build our first model using OLS method and we obtain an adjusted R square of 53.5% and RMSE of 404.86.

Table II : Model Building Using OLS Method and Sklearn Package.

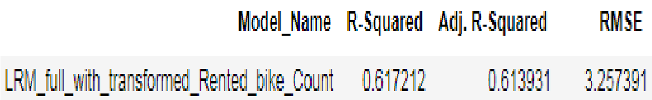


	Model_Name	R-Squared	Adj. R-Squared	RMSE
0	LRM_using _OLS	0.539939	0.535995	404.866571
1	LRM_using_sklearn	0.539939	0.535995	404.866571

2). Model Building Using Transformation Technique [11] :

Here the dependent variable is transformed using some of the transformation techniques like log, power, inverse, square root, box cox transformation and many more. As we apply the transformation we should observe skewness which is very close to zero. The box cox transformation give us a skewness of 0.07159 which is very close when compared to other transformation techniques. Hence we have built our model based on box cox transformation and have seen adjusted R square of 61.39% and RMSE of 3.257390.

Table III : Model Building Using Transformation Technique.

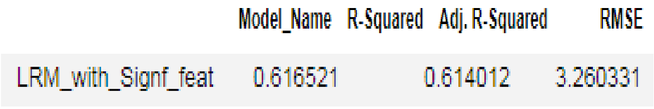


Model_Name	R-Squared	Adj. R-Squared	RMSE
LRM_full_with_transformed_Rented_bike_Count	0.617212	0.613931	3.257391

3). Model Building Using Transformed Model With Significant Features:

From the above model we check the assumptions like multicolinearity. The features like year_2018, Dew point temperature (°C), month and Holiday as the multicolinearity is not satisfied. After dropping these features we see that our model has now got adjusted R square of 61.40% and RMSE of 3.26033

Table IV: Model Building Using Transformed Model With Significant Features.

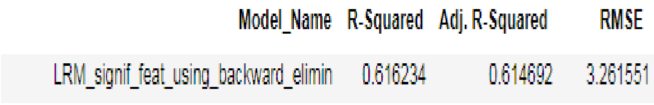


Model_Name	R-Squared	Adj. R-Squared	RMSE
LRM_with_Signf_feat	0.616521	0.614012	3.260331

4). Model Building Using Backward Elimination Method [12] :

Backward elimination is a feature selection technique while building a machine learning model. It is used to remove those features that do not have a significant effect on the dependent variable or prediction of output. The significant features by applying this technique is 'Hour', 'Temperature(°C)', 'Humidity(%)', 'Solar Radiation (MJ/m2)', 'Holiday_No Holiday', 'Seasons_Autumn', 'Seasons_Winter', 'year_2017' and we achieve adjusted R square of 61.46% and RMSE of 3.26155.

Table V : Model Building Using Backward Elimination Method.

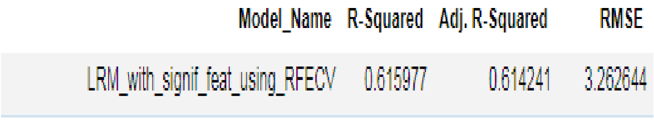


Model_Name	R-Squared	Adj. R-Squared	RMSE
LRM_signif_feat_using_backward_elimin	0.616234	0.614692	3.261551

5). Model Building Using Best Features Using RFE Method (Recursive feature elimination) [12] :

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. After building the model on RFEM we see an adjusted r square of 61.45% and RMSE of 3.2639.

Table VI : Model Building Using Best Features Using RFE Method (Recursive feature elimination).

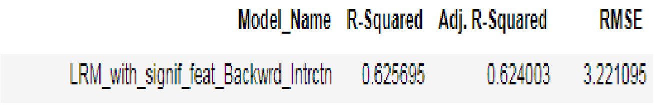


Model_Name	R-Squared	Adj. R-Squared	RMSE
LRM_with_signif_feat_using_RFECV	0.615977	0.614241	3.262644

6). Model Building With Interaction Effect And Backward Elimination [12] :

Interaction effect is nothing but the multiplication of two best features. Here the two features are temperature and hour. When modeling the interaction effect and backward elimination we have adjusted r square of 62.4% and RMSE of 3.2211.

Table VII : Model Building With Interaction Effect and Backward Elimination.

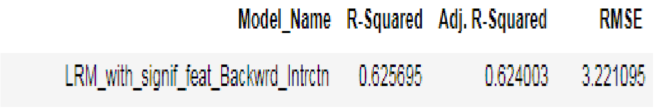


Model_Name	R-Squared	Adj. R-Squared	RMSE
LRM_with_signif_feat_Backwrd_Intrctn	0.625695	0.624003	3.221095

7). Model Building With Interaction Effect And RFEM:

As already explained what interaction effect and RFEM is, we shall go ahead and build a model on this concept. After model building we notice that adjusted r square of 62.37% and RMSE of 3.2221.

Table VIII : Model Building with Interaction Effect and RFEM.

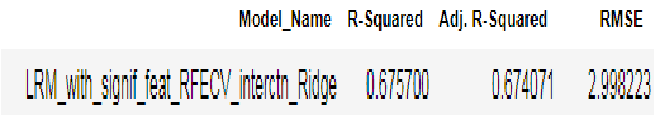


Model_Name	R-Squared	Adj. R-Squared	RMSE
LRM_with_signif_feat_Backwrd_Intrctn	0.625695	0.624003	3.221095

8). Model Building With Interaction Effect And Ridge [13] :

This model solves a regression model where the loss function is the linear least squares function and regularization is given by the l2-norm.After model building we notice that adjusted r square of 67.40% and RMSE of 2.998223.

Table IX : Model Building With Interaction Effect and Ridge.

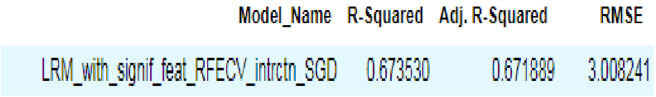


	Model_Name	R-Squared	Adj. R-Squared	RMSE
	LRM_with_signif_feat_RFECV_interctn_Ridge	0.675700	0.674071	2.998223

9). Model Building With Interaction Effect And Stochastic Gradient Descent (SGD) [14] :

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. After model building we notice that adjusted r square of 67.1889% and RMSE of 3.008241.

Table X : Model Building With Interaction Effect And Stochastic Gradient Descent (SGD) **.**

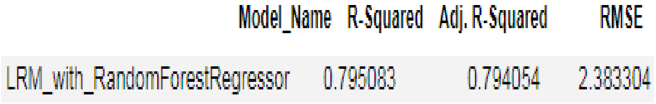


	Model_Name	R-Squared	Adj. R-Squared	RMSE
	LRM_with_signif_feat_RFECV_intrctn_SGD	0.673530	0.671889	3.008241

10). Model Building Using Significant Feature In Random Forest Regression [15] :

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. After model building we notice that adjusted r square of 79.40% and RMSE of 2.3833

Table XI : Model Building Using Significant Feature In Random Forest Regression.

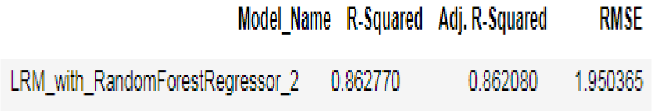


	Model_Name	R-Squared	Adj. R-Squared	RMSE
	LRM_with_RandomForestRegressor	0.795083	0.794054	2.383304

11). Model Building By Using All Variables By Removing Multi-Collinear Variable In Random Forest Regression [15] :

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. After model building we notice that adjusted r square of 89.20% and RMSE of 1.950365.

Table XII : Model Building By Using All Variables By Removing Multi-Collinear Variable In Random Forest Regression.



	Model_Name	R-Squared	Adj. R-Squared	RMSE
	LRM_with_RandomForestRegressor_2	0.862770	0.862080	1.950365

Results and Discussion

So we have made model by using various feature engineering methods. We have made total of 11 models based on increase of Adj-R-Squared value and decrease of RMSE value.

So as we can see from the above Table I to Table XII different types of feature engineering model build and used to predict R-Squared, Adj.R-Squared and RMSE of the model.
The **Table I** shows representation and sequence of the first four model.

The **Table II** shows to base model which includes OLs module and the sklearn package model and we get
For OLS Model:

R-squared: 0.53993
Adj. R-squared: 0.535995
RMSE: 404.867

For Sklearn Model:
R-squared: 0.539939
Adj. R-squared: 0.535995
RMSE: 404.867.

The **Table III** shows Model Build With Significant Features and we get
Form Table III :
R-squared: 0.613931
Adj. R-squared: 0.617212
RMSE: 3.257.

The **Table IV** shows Model Building Using Transformed Model With Significant Features and we get

Form Table IV :
R-squared: 0.616521
Adj. R-squared: 0.614012
RMSE: 3.26033.

The **Table V** shows Model Building Using Backward Elimination Method and we get

Form Table V :
R-squared: 0.616234
Adj. R-squared: 0.614692
RMSE: 3.26155.

The **Table VI** shows Model Building Using Best Features Using RFE Method (Recursive feature elimination) and we get

Form Table VI :
R-squared: 0.615977
Adj. R-squared: 0.614241
RMSE: 3.26264.

The **Table VII** shows Model Building With Interaction Effect And Backward Elimination and we get

Form Table VII :
R-squared: 0.625695
Adj. R-squared: 0.624003
RMSE: 3.2211.

The **Table VIII** shows Model Building with Interaction Effect and RFEM and we get

Form Table VIII :
R-squared: 0.625615
Adj. R-squared: 0.623923
RMSE: 3.22144.

The **Table IX** shows Model Building with Interaction Effect and Ridge and we get

Form Table IX :
R-squared: 0.675700
Adj. R-squared: 0.674071
RMSE: 2.99822.

The **Table X** shows Model Building with Interaction Effect and Stochastic Gradient Descent (SGD) and we get

Form Table X:
R-squared: 0.674683
Adj. R-squared: 0.673048
RMSE: 3.00292.

The **Table XI** shows Model Building Using Significant Feature in Random Forest Regression and we get

Form Table XI :
R-squared: 0.795083
Adj. R-squared: 0.794054
RMSE: 2.3833.

The **Table XII** shows Model Building by Using All Variables by Removing Multi-Collinear Variable in Random Forest Regression and we get

Form Table XII :
R-squared: 0.862770
Adj. R-squared: 0.862080
RMSE:1.950365

Now Plotting the above results in graph and seeing the model with high Adj. R-squared and R-squared and also having very low RMSE value.

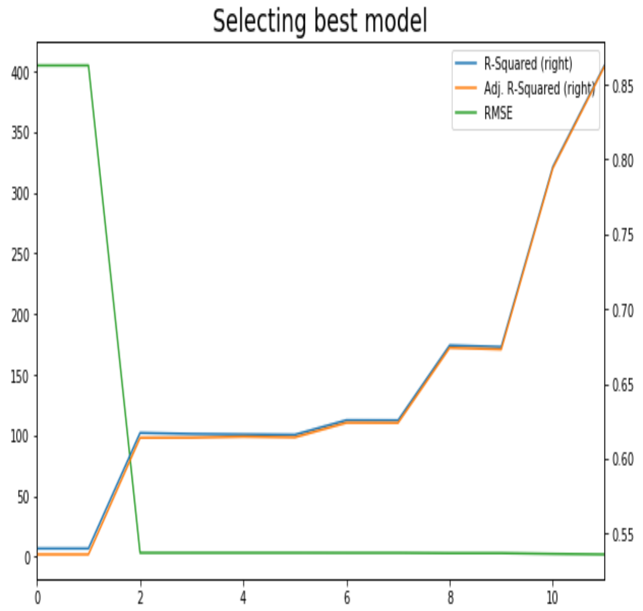
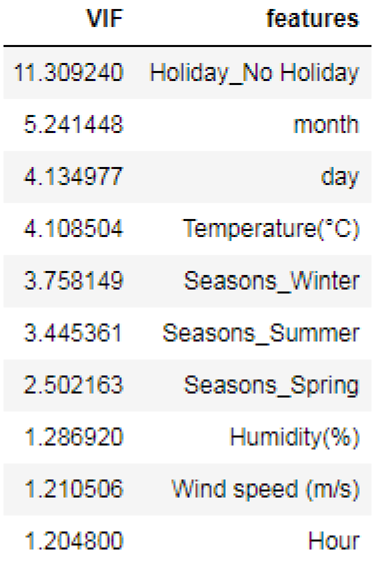


Fig: - 14 Model Comparison

So based on the above graph we can see that the Model 11 that is Using All Variables And Removed Multi-Collinear Between Variable In Random Forest Regression performs best among all the model and has the highest Adj. R-squared and R-squared and also having very low RMSE value.

So we can now see the variable used on this model and Variance Inflation Factor.

Table XIII: Variable Used



VIF	features
11.309240	Holiday_No Holiday
5.241448	month
4.134977	day
4.108504	Temperature(°C)
3.758149	Seasons_Winter
3.445361	Seasons_Summer
2.502163	Seasons_Spring
1.286920	Humidity(%)
1.210506	Wind speed (m/s)
1.204800	Hour

Base on the Model 11 we can see the Y-pred and Y-test value in graph to see the linearity.

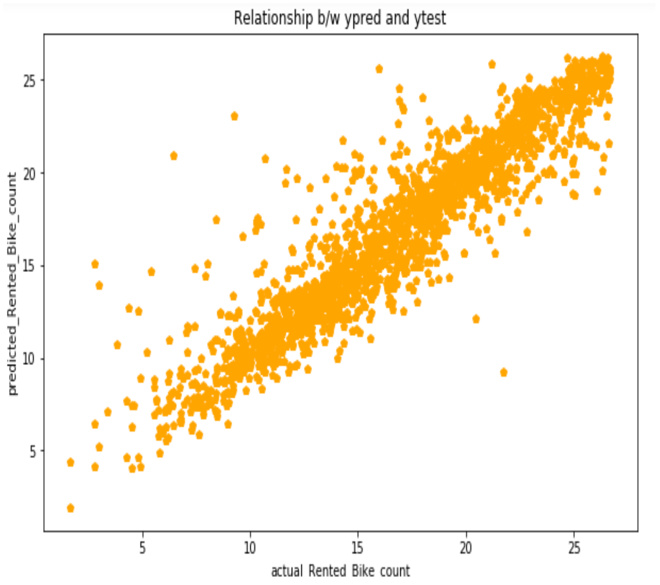


Fig: - 15 Actual vs Predict.

So as we can from the above graph that the linearity between Y-test and Y-pred almost has a linear relation and we can say that the predication is almost linear for the Y-test values

Scope for further advanced techniques

Future work will focus on district wise rental bike demand prediction by considering seasonal changes. The future of bike-sharing is clear: there will be a lot more of it. Gilles Vesco, Vice President of Greater Lyon, quotes his mayor when saying, “There are two types of mayors in the world: those who have bike-sharing and those who want bike sharing.” This certainly seems to be the case as each bike-sharing program creates more interest in this form of transit—call it a virtuous cycle. As the price of fuel rises, traffic congestion worsens, populations grow, and a greater world-wide consciousness arises around climate change, it will be necessary for leaders around the world to find new modes of transport and better adapt existing modes to move people in more environmentally sound, efficient, and economically feasible ways. Bike-sharing is evolving rapidly to fit the needs of the 21st century.

Acknowledgment

We would like to sincerely thank our project mentors Mrs.Sudha B.G and Mr. Senthil Kumar for their valuable time to review our paper and insights.

References

1. https://www.researchgate.net/publication/251714314_Understanding_Bike-Sharing_Systems_using_Data_Mining_Exploring_Activity_Patterns
2. https://www.tandfonline.com/doi/full/10.1080/22797254.2020.1725789
3. https://www.researchgate.net/publication/256972521_Optimizing_the_location_of_stations_in_bike-sharing_programs_A_GIS_approach
4. https://researchportal.bath.ac.uk/en/publications/an-exact-algorithm-for-the-static-rebalancing-problem-arising-in-

5. https://www.tandfonline.com/doi/full/10.1080/22797254.2020.1725789
6. https://www.researchgate.net/publication/339266153_A_rule-based_model_for_Seoul_Bike_sharing_demand_prediction_using_weather_data
7. https://www.sciencedirect.com/science/article/abs/pii/S0360835218306260
8. https://www.tandfonline.com/doi/full/10.1080/22797254.2020.1725789
9. https://mobisoftinfotech.com/resources/blog/list-of-popular-bike-rental-startups-in-india/
10. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
11. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.boxcox.html
12. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html
13. http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html
14. https://scikit-learn.org/stable/modules/sgd.html
15. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html