Bike Sharing Demand Prediction

Parth Sharma-Cohort Amsot

Abstract:

A Bike-sharing system is a shared transport service in which bikes are made available for shared use to individuals on a short-term basis for a price. Many bike share systems allow people to borrow a bike from a "dock" and return it at another dock belonging to the same system.

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**.**

The first segment of this work is dedicated towards the exploratory data analysis i.e. understanding the pattern lying beneath and second part is dedicated towards applying different models and selecting the appropriate one.

1. Problem Statement

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**.**

2. Introduction

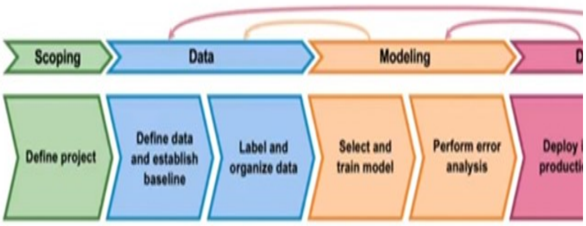
Bike-renting systems allow anyone to hire bike from one of the city's numerous automated rental stations, ride it for a short distance, and then return it to any station in the city. Many cities across the world have recently implemented similar systems .First country to implement this model is the Portland, Oregon, following their example bike rental systems are widespread throughout the world .The ability of a rental bike system to meet the variable demand for bicycles and make it available at the time of peak of its demand. This is accomplished by a repositioning operation that involves taking bicycles from some stations and transporting them to other stations with the help of a specialized fleet of trucks. From the economical point of view, this process is highly intensive and expensive for bike sharing companies. Therefore another optimal solution is needed. To solve such a problem machine learning methods will used**.**

To what extent we can predict bike count required for the stable supply of rental bikes using machine learning. Since the goal is to find the best model that will output highest accuracy on data set, therefore the first sub-question that comes to mind is: "Which machine learning algorithm gives the highest prediction accuracy.

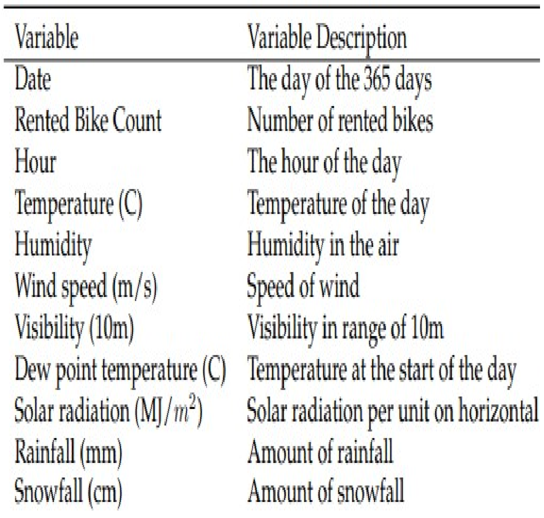
In addition, there are many features in the data such as temperature and time which can be used to predict bike counts. It would be interesting to see which feature is more importance for prediction. Therefore, a second question that comes to mind is “Which features are the most important for bike count prediction”.

3. Steps and Methods Involved:

First step of any project would be peeking into what data is provided. Therefore, we will start with learning about the dataset and then move forward to data pre-processing, EDA and then will move to apply different ML models on the data set.



- **Dataset:**
  Our data set contain the following attributes that are defined over 24 hours of the day. Date, Rented Bike Count, Hour, Temperature(°C), Humidity (%), Wind Speed (m/s), Visibility (10m), Dew Point Temperature(°C), Solar Radiation (MJ/m2), Rainfall (mm), Snowfall (cm), Seasons, Holidays, Functioning Day with 8760 records.



Variable	Variable Description
Date	The day of the 365 days
Rented Bike Count	Number of rented bikes
Hour	The hour of the day
Temperature (C)	Temperature of the day
Humidity	Humidity in the air
Wind speed (m/s)	Speed of wind
Visibility (10m)	Visibility in range of 10m
Dew point temperature (C)	Temperature at the start of the day
Solar radiation (MJ/m^2)	Solar radiation per unit on horizontal
Rainfall (mm)	Amount of rainfall
Snowfall (cm)	Amount of snowfall

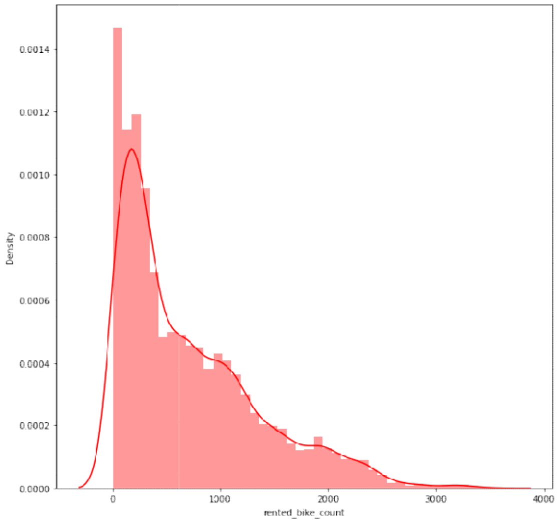
- **Data Pre-Processing And Exploratory Data Analysis:**
  In this section, methods for data exploration and pre-processing will be presented so that we can make our data fit for passing into different ML models.
  The majority of the real-world datasets are highly susceptible to missing, inconsistent, and noisy data due to their heterogeneous origin. Applying algorithms on this noisy data would not give quality results as they would fail to identify patterns effectively. Data Processing is, therefore an important step or stage to improve the overall data quality.Duplicate or missing values may give an incorrect view of the overall statistics of data. Outliers and inconsistent data points often tend to disturb the model’s overall learning, leading to false predictions.
  Major Tasks in Data Pre-processing:
  1. Data cleaning
  2. Data integration
  3. Data reduction
  4. Data transformation

EDA or Exploratory Data Analysis is the critical process of performing the initial investigation on the data to find the anomalies in our data and shape it such that it is useful for taking some insights to solve our purpose. There are certain step that we follow initially we will clean our data and make it free from anomalies such as Nan values, missing values and such values that could hinder the accuacy of our analysis.

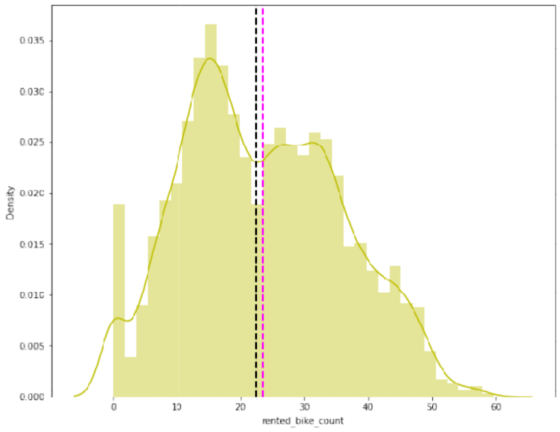
The ultimate aim of the step is to provide our ML models the best possible and clean data, for that we need to treat the outliers; duplicate values must also be taken into account. Luckily, in our dataset there were absolute zero missing values and duplicate values. Therefore, our data was clean in this regard.

It is crucial to handle categorical data during the pre-processing phase, as in the data having categorical features, namely 'hour', 'seasons', 'holiday', 'functioning day', 'month'. Machine learning can’t deal with categorical data. Hence, we must be converting into numerical ones. One-hot encoding, creating dummy variables and many other methods can be used for categorical variables. Based on current work, dummy variables were created for each of the features.

The next step is to verify that the data is normal. Target variable distribution seems to not follow a normal distribution. Thus, data was transformed to closely resemble a normal distribution.



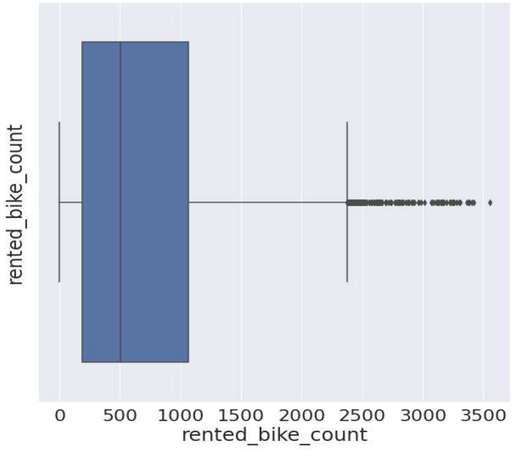
BEFORE TRANSFORMATION



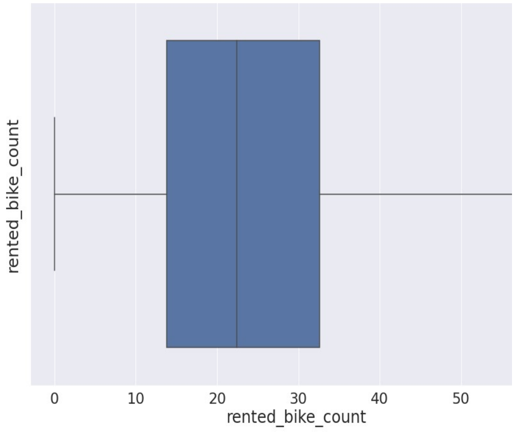
AFTER TRANSFORMATION

We have applied here square-root method to normalize the dependent variable. Initially, Log-transformed was applied but due the presence of ‘0’ at some instances. Therefore, mathematical transformer is used.

The outliers were also got treated as soon as we applied the transform.



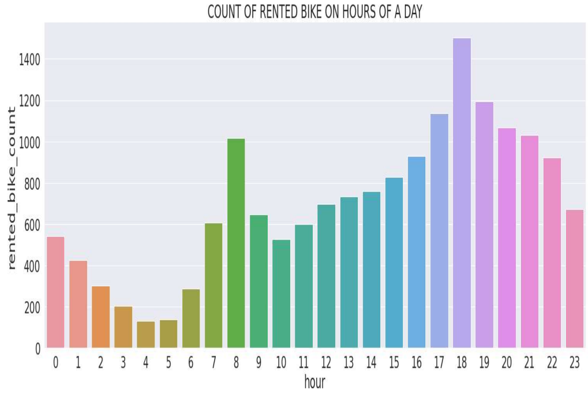
BEFORE TRANSFORMATION



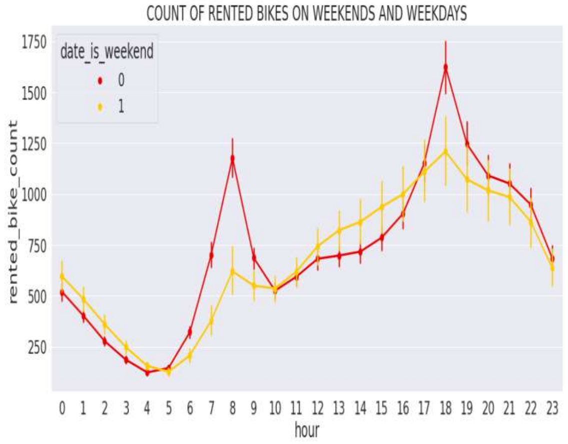
AFTER TRANSFORMATION

- **Visualisation:**
  Data visualization is the most important step while doing the analysis. It is more impressive, interesting and understanding when we represent our study or analysis with the help of colours and graphics. Using visualization elements like graphs, charts, maps, etc., it becomes easier to understand the underlying structure, trends, patterns and relationships among variables within the dataset.
  The libraries use for the plotting and visualisation are **Matplotlib** and **Seabon**.
  Some of the plots are show below:

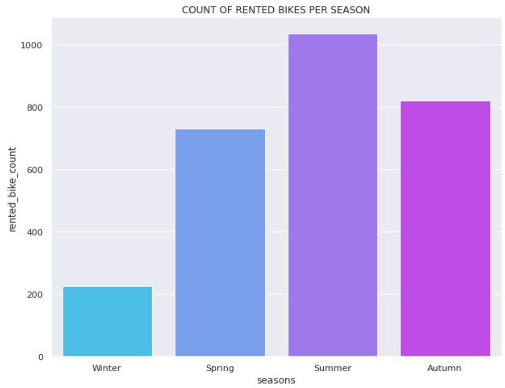
- **Count Of Rented Bike On Hours Of a Day:**

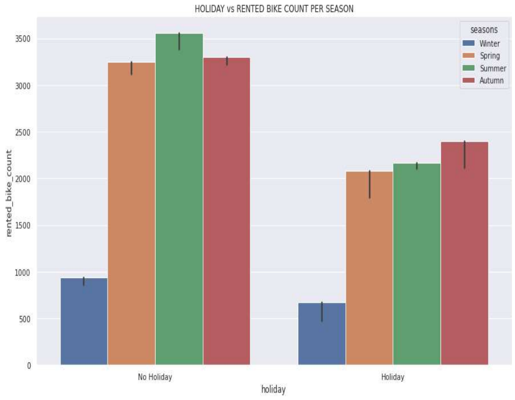


- **Rented Bike Demand On Weekdays And Weekends:**

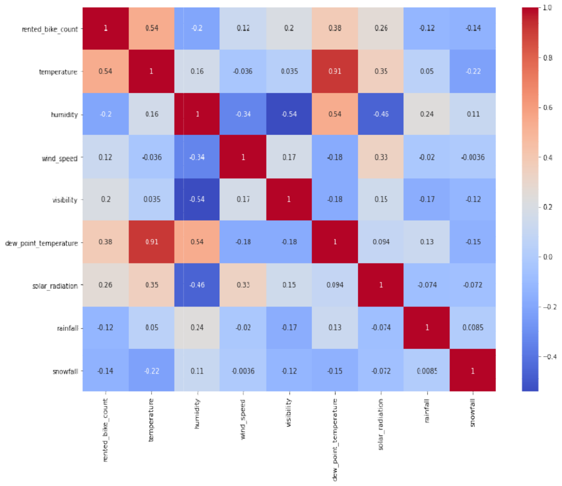


- **Some more Visualisations:**



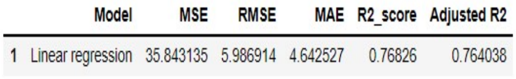


- **Correlation Matrix:**
  On the target variable line the most positively correlated variables are: Temperature, Wind Speed, Dew Point Temperature, Solar Radiation, and Visibility.
  Most negatively correlated variables are Humidity, Rainfall, and Snowfall.



➢ MODELS:

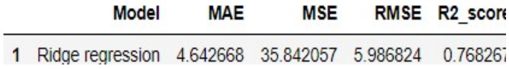
- **Linear Regression**:
  In the world of machine learning, linear regression is one of the most simple and widely used models. Linear regression assumes that the dependent variable is/are linearly correlated to the independent features in the dataset. Linear regression fits a linear model with coefficients for each feature to minimize the mean square error. Below shown the metrics that we got, it represents the goodness of fit of a model.



	Model	MSE	RMSE	MAE	R2_score	Adjusted R2
1	Linear regression	35.843135	5.986914	4.642527	0.76826	0.764038

Since, R2 we got is 0.768. Therefore, we can say that our model gave us good results.

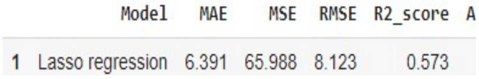
- **Ridge Regression**:
  Ridge is popular regularization techniques which are used to prevent over fitting. Ridge regression applies an L2 penalty, which penalized coefficients with higher weights. Below shown the metrics that we got, it represents the goodness of fit of a model. Below shown the metrics that we got, it represents the goodness of fit of a model.



	Model	MAE	MSE	RMSE	R2_scor
1	Ridge regression	4.642668	35.842057	5.986824	0.76826

Since, R2 we got is 0.768. Therefore, we can say that our model gave us good results.

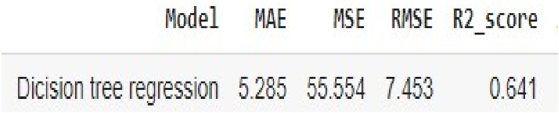
- **Lasso Regression**:
  Lasso regression employs the L1 normalization technique which penalizes less relevant features in the dataset by setting coefficients zero and hence eliminates them. As a result, you get of feature selection and a more efficient model. Below shown the metrics that we got, it represents the goodness of fit of a model.



	Model	MAE	MSE	RMSE	R2_score	A
1	Lasso regression	6.391	65.988	8.123	0.573	

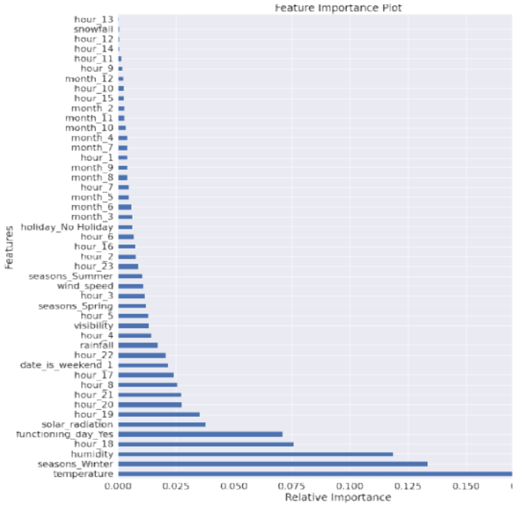
Since, R2 we got is 0.573. Therefore, we can say that our model gave us poor results.
It was expected from both Ridge and Lasso Regressor to provide better results than the base line model but both can't stand on our expectations.

- **DECISION TREE:**
  Decision trees are statistical models that measure a target value using a collection of binary rules. A decision tree generates an approximation by "asking" the data a series of questions between 2 nodes, each of which narrows the range of possible values until the model is positive to make a single prediction. The model determines the order of the questions as well as their content. Below shown the metrics that we got, it represents the goodness of fit of a model.



Model	MAE	MSE	RMSE	R2_score
Dicision tree regression	5.285	55.554	7.453	0.641

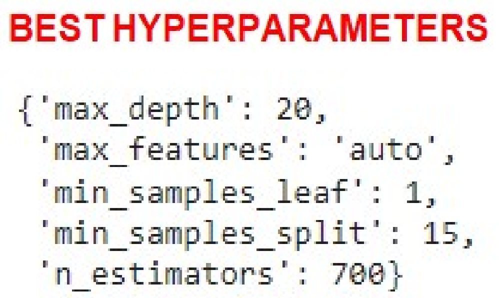
Since, R2 we got is 0.641. Therefore, we can say that our model gave us fairly Ok results but not that good. Shown below the feature importance according to the decision tree model. Temperature tops the list as it posses the maximum relative importance score. Feature that is least apart from those are one hot encoded is snowfall.



- **Random Forest tuning with Randomized Search CV:**

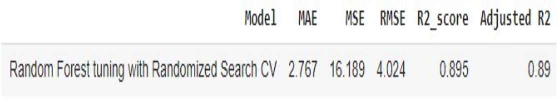
  Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. The random forest algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. Also, after hyper parameter tuning we get a set of hyper parameters that is provided by the model to get us the best results. Here, with random forest we have used Randomized Search CV.

  Now, we will have a look over the hyper parameter that RSCV has suggested and it was observed that it has increased the model accuracy.


BEST HYPERPARAMETERS

```
{'max_depth': 20,
 'max_features': 'auto',
 'min_samples_leaf': 1,
 'min_samples_split': 15,
 'n_estimators': 700}
```

Below shown the metrics that we have got after applying the optimized model.



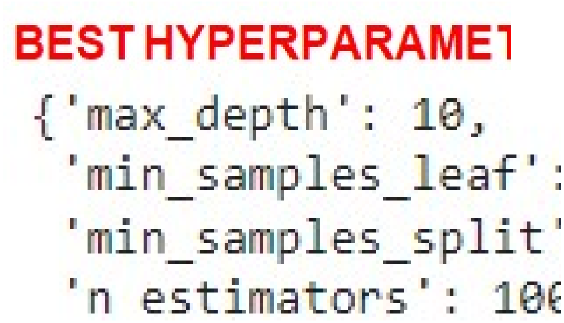
Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Random Forest tuning with Randomized Search CV	2.767	16.189	4.024	0.895	0.89

Since, R2 we got is 0.895. Therefore, we can say that our model gave us better results .But there could be even better R2 score that we can get and let just go and check for other model also.

- **Gradient Boosting Regressor with Grid Search CV:**

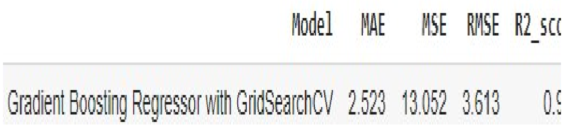
  Gradient Boosting is a machine learning algorithm, used for both classification and regression problems. It works on the principle that many weak learners can together make a more accurate predictor. Gradient boosting works by building simpler (weak) prediction models sequentially where each model tries to predict the error left over by the previous model. As already discussed that our aim is to improve the model in any way possible. One

important factor in the performances of these models are their hyper parameters, once we set appropriate values for these hyper parameters, the performance of a model can improve significantly. Thus, just like Randomized Search CV, Grid Search CV is one such method.
Now, we will have a look over the hyper parameter that GSCV has suggested and it was observed that it has increased the model accuracy.



```
BEST HYPERPARAMET
{'max_depth': 10,
  'min_samples_leaf':
  'min_samples_split'
  'n estimators': 10
```

Below shown the metrics that we have got after applying the optimized model.



	Model	MAE	MSE	RMSE	R2_sc
Gradient Boosting Regressor with GridSearchCV		2.523	13.052	3.613	0.9

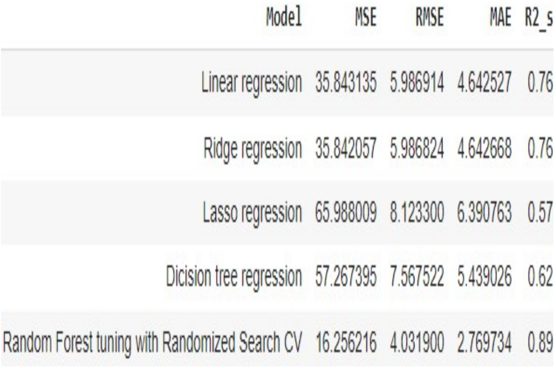
It is clearly visible that so far the best R2 score that we have got is 0.916 and that is through Gradient Boosting Regressor with Grid Search CV.

4. Observations:

- There was a high correlation between the dependent variables specifically dew point temperature and temperature due to which the VIF was going very high.
- Temperature, Wind Speed, Dew Point Temperature, Solar Radiation, Visibility are positively correlated with the target variable.
- Humidity, Rainfall, Snowfall are negatively correlated with the target variable.
- As seen in the comparison table Linear Regression as well as Ridge Regression gave us good comparatively results.
- But the Lasso Regression gave worst metrics.
- Decision Tree gave the results that were good in comparison with our traditional models.
- Random Forest tuning with Randomized Search CV, gave us better results but the exceptionally best result was obtained from Gradient Boosting Regressor with Grid Search CV.
- Temperature came out to be the most important feature out of all from the list.

5. Conclusion:

Since Gradient Boosting Regressor with Grid Search CV gave us the best results i.e. R2 score to be .916.Therefore, we can deploy it for our predictions. Also, As this data is time dependent, the values for variables will not always be consistent. Therefore, we need constantly keep checking for the models. The complete comparison sheet has been shown below that has the value of all the metrics that has been used during the analysis.



Model	MSE	RMSE	MAE	R2_s
Linear regression	35.843135	5.986914	4.642527	0.76
Ridge regression	35.842057	5.986824	4.642668	0.76
Lasso regression	65.988009	8.123300	6.390763	0.57
Dicision tree regression	57.267395	7.567522	5.439026	0.62
Random Forest tuning with Randomized Search CV	16.256216	4.031900	2.769734	0.89

5. References:

1. Towards Data Science.
2. Tutorials point.
3. Analytics Vidhya.
4. Wikipedia.
5. Stackover flow.
6. machinelearningmastery.com