Bike Sharing Demand Prediction

Kirtesh Verma, Pravin Bejjo, and

Sahil Pardeshi

Data Science Trainees,

Almabetter, Nashik

➢ Introduction:

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.

➢ Problem Statement:

Maximize: The availability of bikes to the customer.

Minimize: Minimise the time of waiting to get a bike on rent.

The main goal of the project is to:

Finding factors and cause those influence shortage of bike and time delay of availing bike on rent. Using the data provided, this paper aims to analyse the data to determine what variables are correlated with customer churn, if any. Hourly count of bike for rent will also be predicted.

➢ DATASET PREPARATION:

The bike sharing demand prediction dataset from rented bike provider company from Seoul contains 14 features and 8760 observations of a complete year I.e., from 1.12.2017 to 31.11.2018. Below Table shows the data features.

Data-set description

FEATURE	TYPE
Date: year-month-day	Date
Rented Bike Count	Int64
Hour	Int64
Temperature(°C)	Float64
Humidity (%)	Int64
Wind speed (m/s)	Float64
Visibility (10m)	Int64
Dew Point temperature (°C)	Float64
Solar Radiation (MJ/m2)	Float64
Rainfall (mm)	Float64
Snowfall(cm)	Float64
Seasons	Object
Holiday	Object
Functioning day	Object

➢ Feature Breakdown

Date: The date of the day, during 365 days from 01/12/2017 to 30/11/2018, formatting in DD/MM/YYYY, we need to convert into date-time format.

Rented Bike Count: Number of rented bikes per hour which our dependent variable and we need to predict that

Hour: The hour of the day, starting from 0-23 it's in a digital time format

Temperature (°C): Temperature of the weather in Celsius and it varies from -17*°C to 39.4°C*.

Humidity (%): Availability of Humidity in the air during the booking and ranges from 0 to 98%.

Wind speed (m/s): Speed of the wind while booking and ranges from 0 to 7.4m/s.

Visibility (10m): Visibility to the eyes during driving in "m" and ranges from 27m to 2000m.

Dew point temperature (°C): Temperature
At the beginning of the day and it ranges from -30.6°C to 27.2°C.

Solar Radiation (MJ/m2): Sun contribution or solar radiation during ride booking which varies from 0 to 3.5 MJ/m2.

Rainfall (mm): The amount of rainfall during bike booking which ranges from 0 to 35mm.

Snowfall (cm): Amount of snowing in cm during the booking in cm and ranges from 0 to 8.8 cm.

Seasons: Seasons of the year and total there are 4 distinct seasons I.e., summer, autumn, spring and winter.

Holiday: If the day is holiday period or not and there are 2 types of data that is holiday and no holiday

Functioning Day: If the day is a Functioning Day or not and it contains object data type yes and no.

➢ Steps Involved

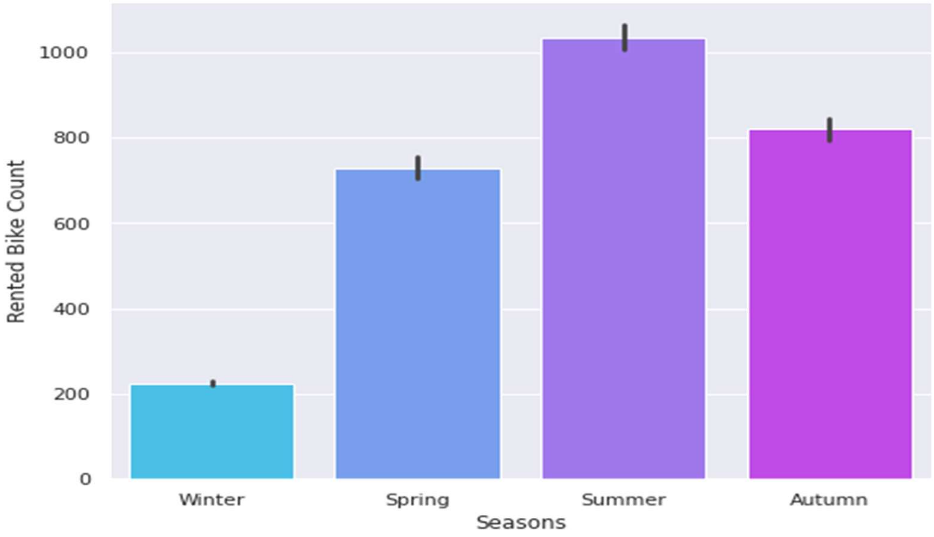
I. Performing EDA (Exploratory data analysis)

1) Exploring head and tail of the data to get insights on the given data.
2) Looking for null values and removing them if it affects the performance of the model.
3) Converting the data into appropriate data types to create a regression model.
4) Creating data frames which help in drawing insights from the dataset.
5) Creating more columns in our dataset which would be helpful for creating model.
6) Encoding the string type data to better fit our regression model.
7) Calculating inter-quartile range and filtering our data.
8) Extracting correlation heatmap and calculating VIF to remove correlated and multicollinear variables.

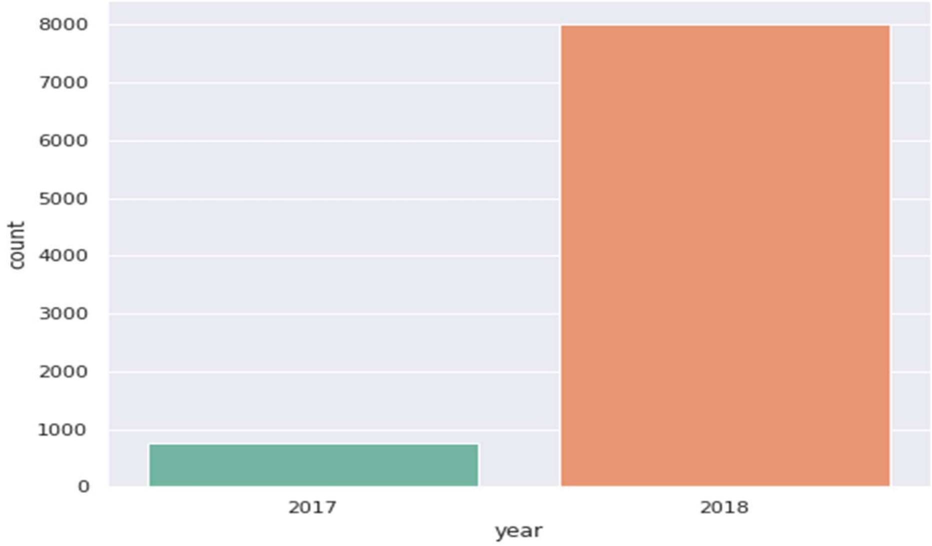
II. Drawing conclusion from the data

Plotting necessary graphs which provides relevant information on our data like:

1) Most bikes have been rented in the summer season
2) Least bike rent count is in the winter season.
3) autumn and spring seasons have almost equal amounts of bike rent count.



4) Most of the bikes have been rented in the year 2018.

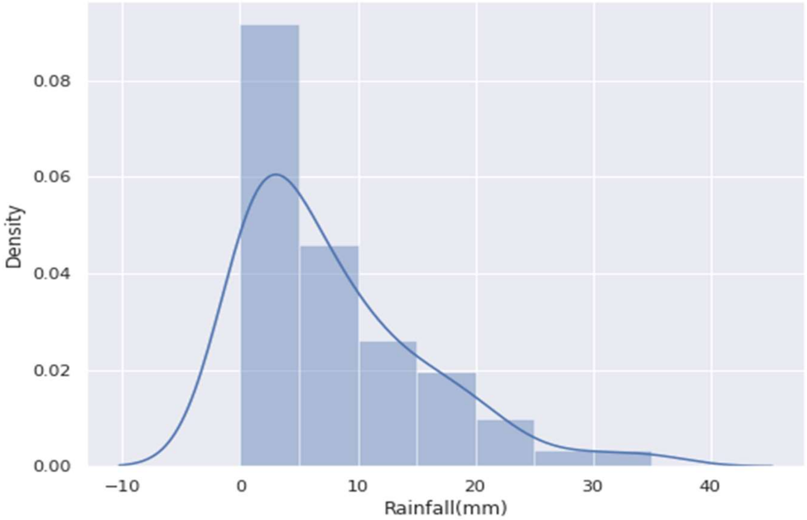


5) Most of the bikes have been rented on working days.

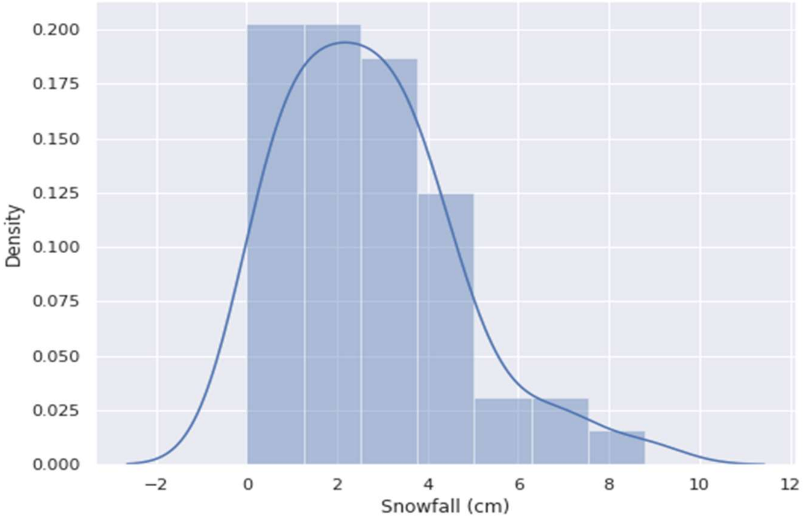
6) Very few bikes have been rented in December which is winter season.

7) Most bikes have been rented in December in the year 2017 as we don't have data before that.

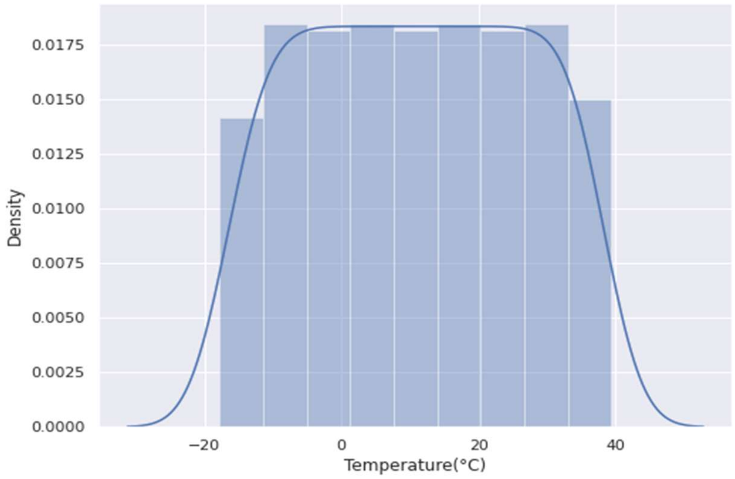
8) People tend to rent bikes when there is no or less rainfall.



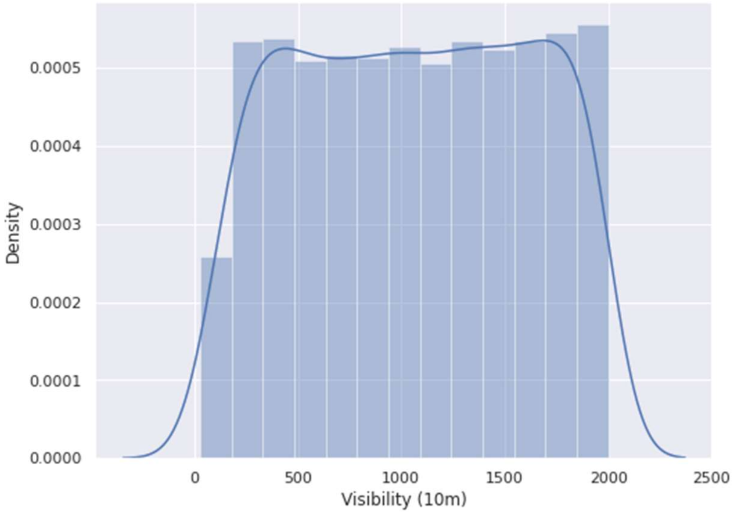
9) People tend to rent bikes when there is no or less snowfall



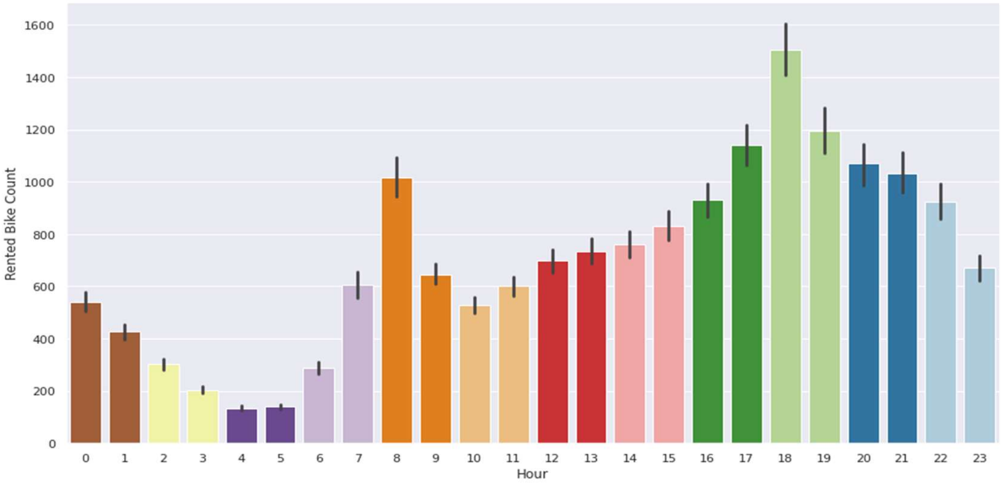
10) People tend to rent bikes when the temperature is between -5 to 25 degrees.



11) People tend to rent bikes when the visibility is between 300 to 1700.



12) The rentals were more in the morning and evening times. This is because people not having personal vehicles, commuting to offices and schools tend to rent bikes.



III. Training The Model

1) Assigning the dependent and independent variables
2) Splitting the model into train and test sets.
3) Transforming data using minmaxscaler.
4) Fitting linear regression on train set.
5) Getting the predicted dependent variable values from the model.

IV. Evaluating metrics of our model

A. Getting MSE, RMSE, R2-SCORE, ADJUSTED-R2 SCORE for different models used.

   1) MSE - the mean squared error or mean squared deviation of an estimator measures the average of the squares of the errors.
   2) RMSE - Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are
   3) R2-SCORE - R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.
   4) ADJUSTED-R2 SCORE - Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

B. Comparing the r2 score of all models used, to get the desired prediction.

Models used

▪ Linear regression:

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis

▪ Lasso regression model

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. The acronym “LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso solutions are quadratic programming problems, which are best solved with software (like MATLAB). The goal of the algorithm is to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- **Ridge regression model**

  Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

  The cost function for ridge regression:

  Min (||Y – X(theta)||^2 + λ||theta||^2)

  Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. Higher the values of alpha, bigger is the penalty and therefore the magnitude of coefficients is reduced.

  - It shrinks the parameters. Therefore, it is used to prevent multicollinearity
  - It reduces the model complexity by coefficient shrinkage

- **Decision tree regression model**

  Linear model trees combine linear models and decision trees to create a hybrid model that produces better predictions and leads to better insights than either model alone. A linear model tree is simply a decision tree with linear models at its nodes. This can be seen as a piecewise linear model with knots learned via a decision tree algorithm. LMTs can be used for regression problems (e.g., with linear regression models instead of population means) or classification problems (e.g., with logistic regression instead of population modes).

- **Random forest regression model**

  Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.[1][2] Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

- **Gradient boosting regression model**

The term gradient boosting consists of two sub-terms, gradient and boosting. We already know that gradient boosting is a boosting technique. Let us see how the term ‘gradient’ is related here.

Gradient boosting re-defines boosting as a numerical optimisation problem where the objective is to minimise the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimisation algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimising a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification, etc

Gradient Boosting Machine (GBM) builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

How does it work?

- Let's take a dataset {(x1, y1), (x2, y2), (x3, y3) ...., (xn, yn)}
- Choose a loss function, let's say MSE.
- Fit a naive model on the dataset, a simple tree or just take y¯, call this model F0(x)

First iteration

- Get residuals of all predictions, ri1(x)=yi−F0(xi)
- Fit a model (can be regression tree) on residuals {(x1,r11),(x2,r21),(x3,r31),....,(xn1,rn1)}, call this model h1(x)
- New predictor is F1(x)=F0(x)+γ1h1(x). Find γ1 which minimizes MSE.

Second iteration

- Get residuals of all predictions, $ri2(x)=yi-F1(x)$
- Fit a model (can be regression tree) on residuals $\{(x1,r12),(x2,r22),(x3,r32),....,(xn2,rn2)\}$, call this model $h2(x)$
- New predictor is $F2(x)=F1(x)+\gamma 2h2(x)$. Find $\gamma 2$ which minimizes MSE.

And so on…

- Get residuals of all predictions, $rim(x)=yi-Fm-1(x)$
- Fit a model (can be regression tree) on residuals $\{(x1,r1m),(x2,r2m),(x3,r3m),....,(xnm,rnm)\}$, call this model $hm(x)$
- **Final predictor** is $Fm(x)=Fm-1(x)+\gamma mhm(x)$. Find γm which minimizes MSE.

Challenges faced

- Pre-processing the data was one of the challenges we faced which includes removing highly correlated variables from the data so as to not hinder the performance of our regression model.

- Exploring all the columns and calculating VIF for multicollinearity was challenging because it might decrease the model's performance.

- Selecting the appropriate models to maximize the accuracy of our predictions was one of the challenges faced.

Conclusion

We are finally at the conclusion of our project!

During the time of our analysis, we initially did EDA on all the features of our dataset. We first analysed our dependent variable, 'Rented Bike Count' and also transformed it. Next, we analysed categorical variable and dropped the variable who had majority of one class, we also analysed numerical variable, found out the correlation, distribution and their relationship with the dependent variable. We

also removed some numerical features who had mostly 0 values and hot encoded the categorical variables.

Next, we implemented 6 machine learning algorithms Linear Regression, lasso, ridge, decision tree, Random Forest and Gradient boosting. We did hyperparameter tuning to improve our model performance. The results of our evaluation are:

Model name	**R2 Score**
1. Linear regression	0.595799
2. Lasso regression	0.597134
3. Ridge regression	0.536137
4. Decision tree regressor	0.790813
5. Decision tree GridsearchCV	0.821301
6. Random forest regressor	0.830182
7. Gradient boosting regressor	0.805943
8. Gradient boosting regressor CV	0.882374

- **No overfitting is seen.**

- **From above its clear that Gradient Boosting regressor (CV) model is the best model for this dataset.**

- **Random forest Regressor and Gradient Boosting gridsearchCV gives the highest R2 score of 83% and 88% respectively for Test set.**

- **Pre-processing the data was one of the difficult challenges we faced.**

- **We were able to get relevant information from the dataset using exploratory data analysis.**

- **We can deploy Random Forest Regressor and Gradient Boosting model.**

- **The most important features who had a major impact on the model predictions were; hour, temperature, Humidity, solar-radiation, and Winter.**

- **Demand for bikes got higher when the temperature and hour values were more.**

- **Demand was high for low values of Humidity and solar radiation.**

- **Demand was high during springs and summer and autumn and very low during winters.**

- **However, this is not the ultimate end. As this data is time dependent, the values for variables like temperature, windspeed, solar radiation etc., will not always be consistent. Therefore, there will be scenarios where the model might not perform well. As Machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time. Therefore, having a quality knowledge and keeping pace with the ever-evolving ML field would surely help one to stay a step ahead in future.**