BIKE SHARING DEMAND PREDICTION-
(Supervised ML Regression)

Aditya Saw and Parijat Krishna
Data science trainees,
AlmaBetter, Bangalore

1.Objective

Currently many metropolitans cities have adopted the use of rental bikes to improve mobility comfort. It is crucial to make the rental bikes accessible and available to the general public at the appropriate time since it reduces waiting. Eventually, maintaining a steady supply of rental bikes for the city emerges as a top priority. Predicting the number of bikes needed to maintain a steady supply of rental bikes at each hour's interval is essential..

2.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.

.3. Dataset

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

4.Steps involved-

I. Executing EDA (Exploratory Data Analysis)

- Examining the data's head and tail to gain understanding of the information provided.
- Searching for null values and eliminating them if they have an impact on the model's performance.
- Putting the information into the proper data types to build a regression model.
- Constructing dataframes that facilitate the extraction of insights from the dataset.
- Adding new columns to our dataset that will be useful for the model-building process.
- Coding the data of the string type to improve the fit of our regression model.
- Computing the interquartile range and data filtering.
- Extracting correlation heatmap and calculating VIF to remove correlated and multicollinear variables.

II. Drawing conclusions from the data

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

- The summer months have seen the majority of bike rentals.
- The winter season has the lowest number of rented bikes.
- The number of bikes rented throughout the autumn and spring seasons is nearly equal.
- The majority of the bicycles were hired in 2018.
- The majority of the bicycles were rented during business hours.
- December, which is the winter season, saw a very low number of bike rentals.
- Because we don't have statistics prior to 2017, the majority of bikes were hired in December of 2017.
- When it rains less or not at all, people frequently rent bicycles.
- When there is little or no snowfall, people choose to rent bikes.
- When the temperature ranges from -5 to 25 degrees, people frequently rent bicycles.
- When visibility ranges from 300 to 1700, people frequently rent bicycles.
- The morning and evening hours saw increased rentals.
- This is due to the prevalence of renting bikes among commuters to

offices and schools who lack personal vehicles.

III. Training the Model-

- The dependent and independent variables are assigned in step A.
- Distinguishing the train and test sets from the model.
- Minmaxscaler data transformation.
- Using a train set to fit a linear regression.
- Obtaining the model's projected values for the dependent variables.

IV. Evaluating metrics of our Model-

- Getting **MSE , RMSE, R2-SCORE, ADJUSTED-R2 SCORE** for different models used.
- **MSE -** *the mean squared error or mean squared deviation* of an estimator measures the average of the squares of the errors.
- **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.
- **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.
- **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.
- **Comparing the r2 score of all models used , to get the desired prediction.**

5.Research Methodology:

Univariate Analysis-

Univariate Analysis is a quantitative-statistical method of evaluation. With this approach of analysis, each variable in a data set is examined separately and the results are each summarised separately.

Therefore, unlike bivariate and multivariate analysis, which look at interactions between several variables, univariate data only serves to describe one component of a piece of research. Although different forms can be utilised, a frequency distribution table or bar graph is the simplest way to combine the data for a single variable (e.g. pie chart, histogram etc.). This indicates that one of these selected modes of presentation is used to analyse the number of examples in a given category (variable).

Bivariate analysis-

Bivariate analysis is one of the statistical analysis where two variables are observed. One variable here is dependent while the other is independent. These variables are usually denoted by X and Y. So, here we analyse the changes occured between the two variables and to what extent. Apart from bivariate, there are other two statistical analyses, which are Univariate (for one variable) and Multivariate (for multiple variables).

In statistics, we usually interpret the given set of data and make statements and predictions about it. During the research, an analysis attempts to determine the impact and cause in order to conclude the given variables.

Bivariate analysis is stated to be an analysis of any concurrent relation between two variables or attributes. This study explores the relationship of two variables as well as the depth of this relationship to figure out if there are any discrepancies between two variables and any causes of this difference. Some of the examples are percentage table, scatter plot, etc.

6.Libraries and Tools used in Data Visualisation:

Library used in Data Visualisation are-

1.Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication quality plots.
- Make interactive figures that can zoom, pan, update.
- Customize visual style and layout.
- Export to many file formats.
- Embed in JupyterLab and Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib.

2.Seaborn- Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

Graphical tools used in Data Visualization-

- Bar Plot.
- Histogram.
- Scatter Plot.
- Pie Chart.
- Line Plot.
- Heatmap.
- Box Plot

7. Models Used-

Linear Regression-

Linear regression is a linear approach for modeling 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 modeled 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.

Logistics Regression

Assumptions-

- There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response

Y due to one unit change in X¹ is constant, regardless of the value of X¹. An additive relationship suggests that the effect of X¹ on Y is independent of other variables.

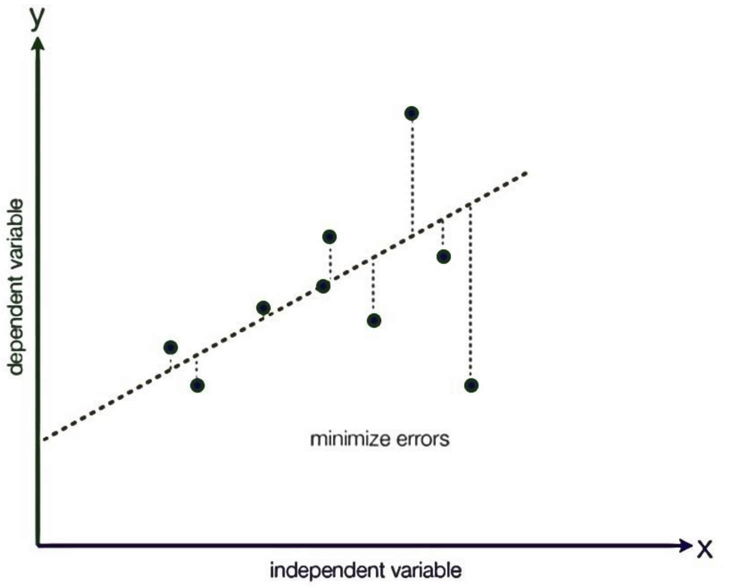
- There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
- The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
- The error terms must have constant variance. This phenomenon is known as homoscedasticity. The presence of non-constant variance is referred to heteroskedasticity.

  The error terms must be normally distributed.

  - We have to train our model considering the above assumptions.

Properties of Logistic Regression-

- The line reduces the sum of squared differences between observed values and predicted values.
- The regression line passes through the mean of X and Y variable values
- The regression constant (b0) is equal to y-intercept the linear regression
- The regression coefficient (b0) is the slope of the regression line which is equal to the average change in the dependent variable (Y) for a unit change in the independent variable (X).



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 multi collinearity 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 goalof 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. Random decision forests correct for decision trees' habit of over fitting 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.

Extra Trees Regression Model:

Extra Trees is an ensemble machine learning algorithm that combines the predictions from many decision trees.

It is related to the widely used random forest algorithm. It can often achieve as-good or better performance than the random forest algorithm, although it uses a simpler algorithm to construct the decision trees used as members of the ensemble.

It is also easy to use given that it has few key hyperparameters and sensible heuristics for configuring these hyperparameters.

Elastic Net Regularization Model:

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

The elastic net method improves lasso’s limitations, i.e., where lasso takes a few samples for high dimensional data. The elastic net procedure provides the inclusion of “n” number of variables until saturation. If the variables are highly correlated groups, lasso tends to choose one variable from such groups and ignore the rest entirely.

To eliminate the limitations found in lasso, the elastic net includes a quadratic

expression (||β||2) in the penalty, which, when used in isolation, becomes ridge regression. The quadratic expression in the penalty elevates the loss function toward being convex. The elastic net draws on the best of both worlds – i.e., lasso and ridge regression.

8. Conclusion:

Coming from the beginning we did EDA on the dataset and also cleaned the data according to our needs.After that we were able to draw relevant conclusions from the given data and then we trained our model on linear regression and other models .

Out of all models used , with extra-trees regression model we were able to get the r2-score of 0.85.The model which performed poorly was elastic net regularization with r2-score of 0.42.

Given the size of data and the amount of irrelevance in the data , the above score is good.

9. Challenges:

1. Huge chunk of the data was to be handled keeping in mind not to miss anything which is even of little reference.

2. Feature selection was quiet challenging as our data set had many futuristic features which had no relevance for initial detection.

3. Computation time