**Bike Sharing Demand Prediction Project**

**Mohd Aun Farooqui**

**INDEX**

**SL.NO Topics**

1. Abstract
2. Introduction
3. Project Goal
4. Attributes
5. Exploratory Data Analysis
6. Data Visualization
7. Algorithms
8. Conclusion
9. Challenges
10. **Abstract**

The Bike Sharing Demand Prediction Project aims to develop a machine learning model that accurately predicts the number of bikes rented at a given time based on various parameters like weather conditions, time of day, etc. The dataset used in this project consists of historical bike rental data from a bike-sharing service provider. The project involves data cleaning, exploration, and feature engineering to extract relevant information from the available data.

Different machine learning algorithms like Random Forest, XGBoost, and CatBoost are trained and evaluated using various performance metrics like Root Mean Squared Error (RMSE) and R-squared (R²) score. The CatBoost algorithm outperformed other algorithms with an RMSE of 44.48 and an R² score of 0.901.

The project highlights the importance of feature engineering and the impact of weather conditions on bike rental demand. Additionally, the project provides insights into the usage patterns of bike-sharing services, which can be useful for service providers to improve their services.

Overall, the Bike Sharing Demand Prediction Project is an essential step towards building efficient and reliable bike-sharing services by accurately predicting the demand and improving customer satisfaction.

1. **Introduction**

Bike-sharing services have become increasingly popular in recent years due to their cost-effectiveness, convenience, and environmental sustainability. With the growing demand for such services, it becomes essential for service providers to accurately predict the demand for bikes to meet the needs of their customers. This is where machine learning comes into play, as it provides an efficient way to predict bike rental demand by analyzing various factors like weather conditions, time of day, day of the week, etc.

In this project, we aim to develop a machine learning model that accurately predicts the number of bikes rented at a given time based on the available data. The dataset used in this project is obtained from a bike-sharing service provider, and it contains various parameters like date, time, temperature, humidity, windspeed, etc., which can be used to predict bike rental demand.

1. **Project Goal**

The main goal of this project is to build a machine learning model that can accurately predict the demand for bike sharing services based on various factors such as weather conditions, time of day, and other external factors. This model can be used by bike-sharing companies to better understand and predict demand, allowing them to optimize their resources, reduce costs, and improve customer satisfaction. The project will involve data cleaning, exploratory data analysis, feature engineering, and model selection and evaluation using various machine learning algorithms. The final model will be evaluated based on its accuracy and ability to predict bike rental demand.

1. **Attributes**

The dataset used in this project contains various attributes that can be used to predict the demand for bike sharing services. These attributes include:

* 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)

1. **Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a critical first step in any data analysis project. It helps to understand the data and identify patterns, relationships, and outliers in the dataset. For the bike sharing demand prediction project, EDA was performed on the dataset to gain insights into the data and understand its structure.

1. **Data Cleaning:**
2. **Handling Null Values**

There are no Null values present in our dataset.

1. **Handling outliers**

An outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighbouring co-existing values in a data graph or dataset we work with.

For the column 'Rainfall(mm)', the maximum value observed was 8.2, and there were some values greater than 4, which is considered as an outlier. So, we have replaced those values with 4, which is the maximum value observed in the dataset. Similarly, we have applied the same technique for other columns like 'Solar Radiation (MJ/m2)', 'Snowfall (cm)', and 'Wind speed (m/s)'.

1. **Data Manipulation:**

Data manipulation is the process of transforming and cleaning the data to make it more structured and organized for further analysis. In this project, data manipulation was performed using various techniques such as merging, grouping, filtering, and sorting to prepare the data for modeling.

1. **Data Visualization**

Data visualization is the graphical representation of data and information using visual elements like charts, graphs, and maps. It helps to communicate complex data and information in an easily understandable and accessible format.

* Bar hart
* Line Plot
* Box Plot
* Histogram
* Pie Chart

1. **Algorithms**

* **Linear Regression:**

Linear regression is a statistical technique that is used to model the relationship between a dependent variable and one or more independent variables. It is a linear approach to modeling the relationship between the variables, which means that it assumes a linear relationship between the variables. The aim of linear regression is to find the line that best fits the data points and can be used to predict the values of the dependent variable for a given set of independent variables.

In multiple linear regression, there are multiple independent variables, and the relationship between the dependent variable and independent variables is represented by a plane or a hyperplane, depending on the number of independent variables.

* **Ridge Regression:**

Ridge regression is a regularization technique used to prevent overfitting in linear regression models. It adds a penalty term to the loss function in linear regression that forces the model coefficients to shrink towards zero. This penalty term is controlled by a hyperparameter called the regularization parameter or alpha.

* **Lasso Regression:**

Lasso regression, also known as L1 regularization, is a type of linear regression that uses a penalty term to shrink the coefficients towards zero. It works by adding a penalty term to the sum of squared errors in the regression equation, where the penalty term is proportional to the absolute value of the coefficients.

* **Decision Tree:**

A decision tree is a machine learning algorithm used for both regression and classification tasks. It is a tree-like model where the internal nodes represent features, the branches represent the decision rules, and the leaf nodes represent the outcomes or predictions.

* **Random Forest:**

Random Forest is a supervised machine learning algorithm that is commonly used for classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to create a more accurate and robust model.

* **Gradient Boosting**:

Gradient Boosting is a machine learning technique that is used for regression and classification problems. It is an ensemble learning method that combines multiple decision trees to make a more accurate prediction. The basic idea behind Gradient Boosting is to iteratively improve a weak model by adding new models that complement the already existing models.

1. **Conclusion**

* The temperature, hours, and solar radiation features were found to be more relevant for the bike count required at each hour for the stable supply of rental bikes.
* As we have analyzed the various features, we have seen that people prefer to take bikes on rent when temperature is near about 25 degrees Celsius.
* Other factors such as rainfall and snowfall also have an impact on the requirement of bikes for rent. Because in heavy rainfall and snowfall bike riding sometime becomes dangerous.
* As we have analyzed that the rental bike demands are high in the evening and in the morning. So bikes should be available at that time to fulfill the bike demands.
* The Bike demand increases with an increase in visibility and decreases with an increase with humidity.
* We tried adding possible columns to make the model a bit more complex but for Linear Regression model it is still too general.
* We have to make our model more complex for better discretion or move to tree and ensemble algorithm for better results.
* Random forest gives predictions better than a decision tree model. Predictions made by Gradient Boosting are better than all the models that we have used. The value of the Adjusted R-squared for the Gradient Boosting method is 0.875, which is very good.
* Adjusted R-squared for both Gradient Boosting and CatBoost are almost same. The adjusted R\_squared score for CatBoost is 0.883. Because of the additional benefits, I will choose CatBoost over Gradient Boosting.

1. **Challenges**
2. **The amount of data collected:** There are some unnecessary raw data collected which do not contribute much to the study. Thus, to identify them and eliminate them will be a challenge.
3. **Handling Null Values and outliers:** - Identifying Null values and outliers and handling them. Handling them is different in different cases. So, we need to analyse it and to process.
4. **Processing raw data into meaningful data:** Sometimes the columns can’t be used as to understand how apply functions on these columns to get more relevant results.
5. **Visual representation:** One size doesn’t fit all. Thus, finding exact graphs to represent the data is challenging.
6. **References**
7. <https://www.almabetter.com/notes>
8. <https://www..python.org.com>
9. <https://www.stackoverflow.com>