Bike sharing demand prediction

Sahil pardeshi, pravin Bejjo , kirtesh verma
Data science Trainee, almabetter nashik

➢ Introduction:

Bike share can be broadly defined as any setting where bicycles are pooled for multiple users. Models include Public Bike Share (PBS) – self-service on-street docked or dockless stations – workplace pool bikes, railway station hubs, loans, lockers and peer to peer sharing. Free-floating bike sharing (FFBS) is an innovative bike sharing model. FFBS saves on start-up cost, in comparison to station-based bike sharing (SBBS), by avoiding construction of expensive docking stations and kiosk machines.

➢ Problem statement :

1) **Maximize: The availability of bikes to the customer.**
2) **Minimize: Minimize the time of waiting to get a bike on rent.**

The main goal of the project is to :

Finding factors and causes that influence shortage of bikes and time delay of availing bikes on rent. Using the data provided by almabetter and to analyze the data and to determine what variables are correlated with customers. Hours count on renting a bike can also be predicted.

➢ Dataset preparation:

The bike sharing demand prediction on rented bike company has 14 features and 8760 observations in complete years.

Dataset description table :

	Date	8760 non-null	object
1	Rented Bike Count	8760 non-null	int64
2	Hour	8760 non-null	int64
3	Temperature(°C)	8760 non-null	float64
4	Humidity(%)	8760 non-null	int64
5	Wind speed (m/s)	8760 non-null	float64
6	Visibility (10m)	8760 non-null	int64
7	Dew point temperature(°C)	8760 non-null	float64
8	Solar Radiation (MJ/m2)	8760 non-null	float64
9	Rainfall(mm)	8760 non-null	float64
10	Snowfall (cm)	8760 non-null	float64
11	Seasons	8760 non-null	object
12	Holiday	8760 non-null	object
13	Functioning Day	8760 non-null	object

Features description

Breakdown of Our Features:

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

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

Hour: The hour of the day, starting from 0-23 it's in a digital time format, type : int, we need to convert it into category data type.

Temperature(°C): Temperature in Celsius, type : Float

Humidity(%): Humidity in the air in %, type : int

Wind speed (m/s) : Speed of the wind in m/s, type : Float

Visibility (10m): Visibility in m, type : int

Dew point temperature(°C): Temperature at the beginning of the day, type : Float

Solar Radiation (MJ/m2): Sun contribution, type : Float

Rainfall(mm): Amount of raining in mm, type : Float

Snowfall (cm): Amount of snowing in cm, type : Float

➢ **Steps involved**

1)Performing EDA (exploratory data analysis):

Exploratory Data Analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

1)To maximize the analyst's insight into a data set and into the underlying structure of a datasets.

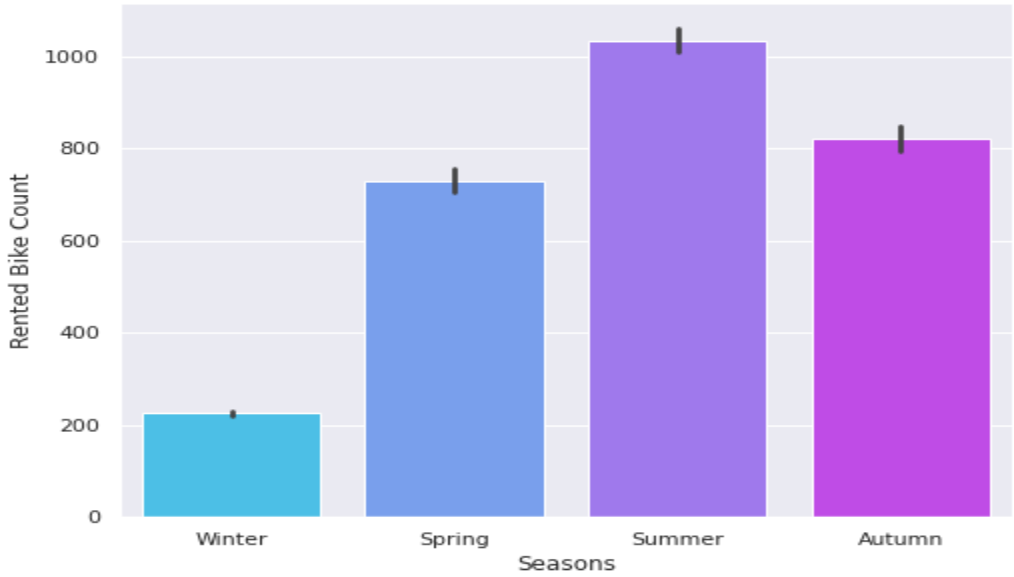
2)Encoding string type data to better fit our regression model.

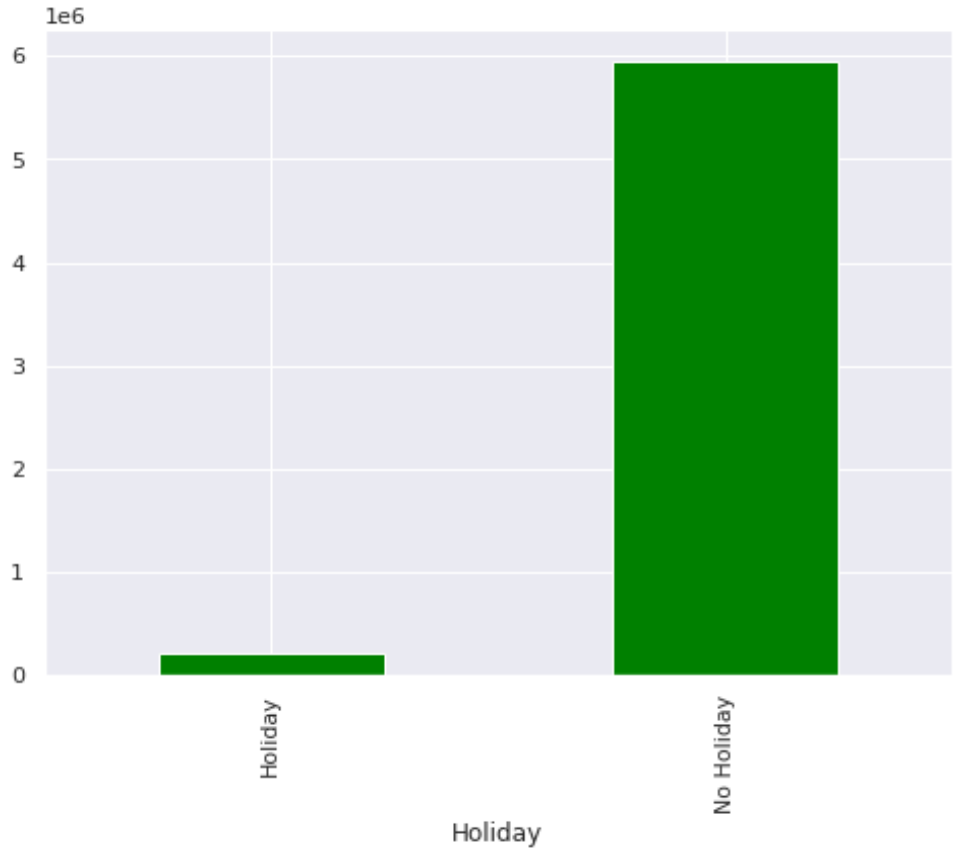
3) Calculating inter-quartile range and filtering our data.

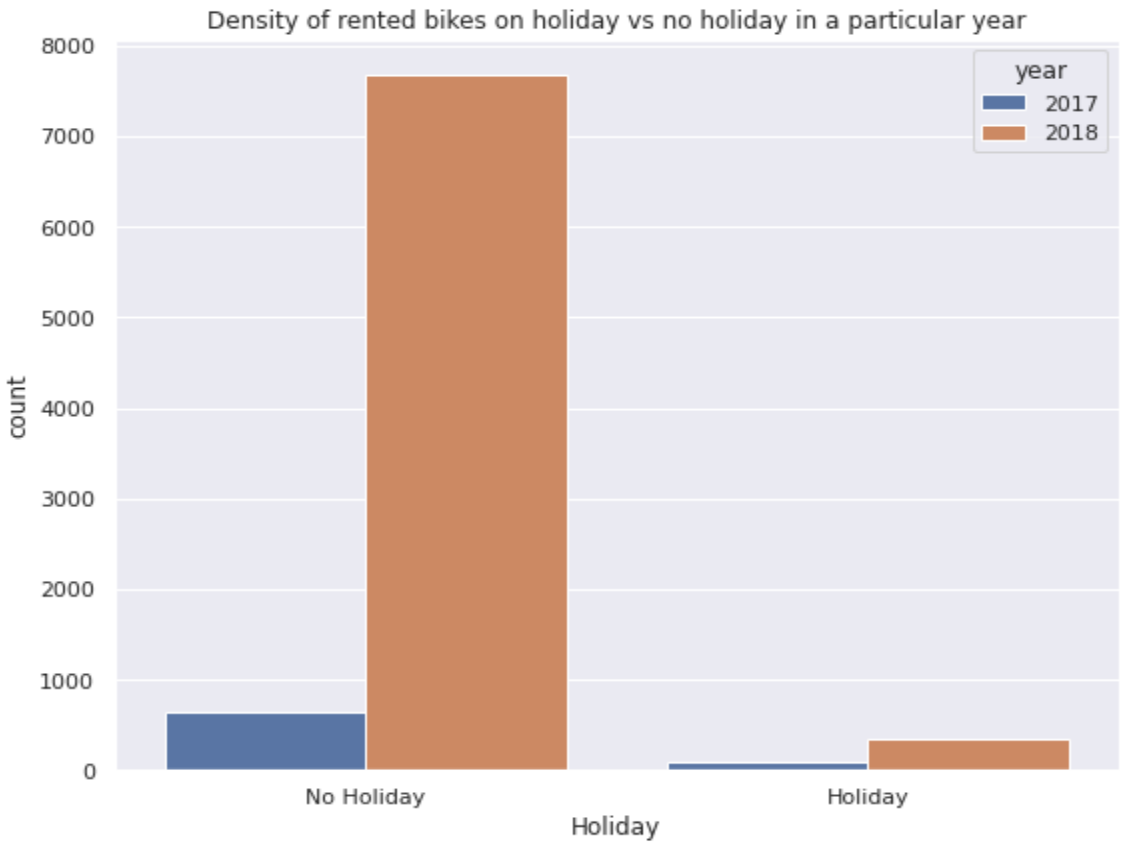
4)Looking for Null values and removing them if it affects the performance of datasets.

5)Extracting correlation heatmap and calculating VIF variance inflation factor to remove correlated and multicollinear variables.

6)visualize the missing values.


Rented Bike Count
1000
800
600
400
200
0
Winter
Spring
Summer
Autumn
Seasons


1e6
6
5
4
3
2
1
0
Holiday
No Holiday
Holiday



2) Drawing conclusion from the data :

1) Most bikes are rented in 2018 less in 2017.

2)Most of the bikes are rented on working days.

3)Bikes were rented high in summer.

4)Most bikes were rented low during winter.

5)Spring and autumn have equal amounts of bikes rented.

6)Most people tend to rent bikes when the temperature is between -5 to 25.

3)Training the model :

1)Assigning the dependent and independent variables.

2)Splitting model into train and test.

3)Fitting linear regression on train sets

- **4)Evaluating metrics of models :**
- **MSE**
- **RMSE**
- **R2-score**
- **ADJUSTED R2 score for different model used**

1)MSE: Mean square error or mean square deviation The mean square error (MSE) provides a statistic that allows for researchers to make such claims. MSE simply refers to the mean of the squared difference between the predicted parameter and the observed parameter.

2)RMSE: Root mean square error is the measure of how well a regression line fits the data points. RMSE can also be construed as Standard Deviation in the residuals.

3)R2-score: The R2 score is a very important metric that is used to evaluate the performance of a regression-based machine learning model. It is pronounced as R squared and is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset.

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.

1. **Linear regression model:** linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.The sign of a regression coefficient tells you whether there is a positive or negative correlation between each independent variable and the dependent variable
2. **Lasso regression:** In statistics and machine learning, lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.lasso regression is a type of linear regression that uses shrinkage. LASSO, short for Least Absolute Shrinkage and Selection Operator, is a statistical formula whose main purpose is the feature selection and regularization of data models.
3. **Ridge regression:** Ridge regression is a model tuning method that is used to analyze 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 being far away from the actual values

4. **Decision Tree:**Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves.

5. **Random forest regression model:** Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.In terms of interpretability, most people place it between conventional machine learning models and deep learning. Many consider it a black-box. Despite widely used, the random forest is commonly interpreted with only feature importance and proximity plots. These visualizations are very useful but not sufficient.
6. **Gradient boosting model:** i) Gradient Boosting Algorithm is generally used when we want to decrease the Bias error. ii) Gradient Boosting Algorithm can be used in regression as well as classification problems. In regression problems, the cost function is MSE whereas, in classification problems, the cost function is Log-Loss. Gradient boosting is based on minimizing a loss function, different type of loss function can be used, resulting in a flexible techniques that can be applied to regression,multi-class classification.

Challenges which we faced they are :

- **Preprocessing the data was one of the challenges we faced which includes removing highly correlated variables from the data.**
- **Calculating vif for multicollinearity was challenging because it might decrease the model performance.**

- ➢ **Selecting the appropriate model to maximize the accuracy of our prediction was one of the challenges faced.**

Conclusion:1) No overfitting is seen.

2)Preprocessing data was one of the difficult challenges.

3) Random forest and gradient boosting model deployed.

4)Number of bikes rented in 2018 is the highest.

5)Demand for bikes is high during no holiday and functioning days.

6) We have did hyperparameter tunning to improve model performance.