Seoul Bike Sharing Demand prediction

Deepak Kumar Dubey
Data Science Trainee ,
Almabetter,Bangalore

1.Abstract

Bike sharing is the provision of making bicycles available for shared use through a self-service rental scheme in which people can hire bicycles for short-term use.

The bike sharing market was valued at USD 3 billion in 2020, and it is anticipated to reach USD 4 billion by 2026, registering a CAGR of about 6% during the forecast period (2021 – 2026).The growing need for urban transportation periodically increased the number of vehicle usage on-road, which led to heavy traffic jams and high environmental pollution. The bike sharing program has been widely adopted by major regions, like Asia-Pacific, North America, and Europe, to alleviate the above-mentioned issues.The Bike Sharing Market is growing at a CAGR of >6% over the next 5 years.

1.1.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.

- The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

- We have around 8760 hours of observations for the whole year that cover all seasons and weather conditions to predict the demand of rental bikes at any particular hour using supervised machine learning algorithms.

- This is a regression problem that comes under supervised machine learning.

2.Data Summary

We have following columns in our data

- Date: year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Wind Speed - 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 - (Non-Functional Day), Fun (Functional Day)

Further these variables can be divided as following-

- Dependent variable -Rented bike counts per hour

- Independent variable-
  Numerical variables-
  a. Temperature
  b. Humidity
  c. wind speed
  d. Humidity
  e. Rainfall
  f. solar radiation
  g. Wind speed
  h. Visibility
  i. Snowfall
  j. Dew point temperature

  Categorical variables-

  k. Hour
  l. Holiday
  m. Functioning day
  n. Seasons

We have 3760 hours of data for 365 days of the year .Weather conditions (temperature ,humidity,wind,speed,humidity,rainfall etc) and rented bike counts of each hour were available.All seasons were equally represented in data.There were hours which had heavy rainfall and snowfall.All these features had been represented in model and helped model to become robust.

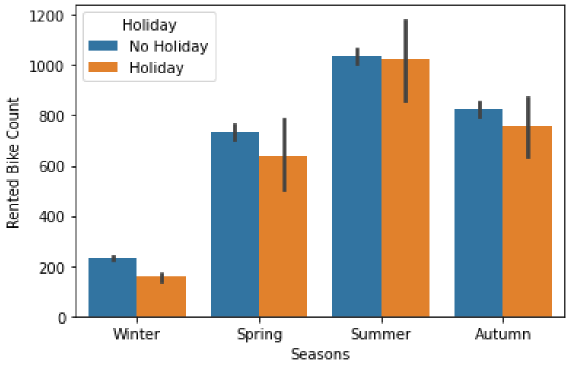
3.Approach

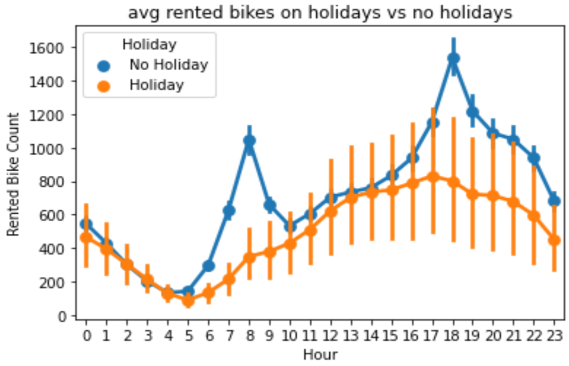
This problem has been solved following these steps-

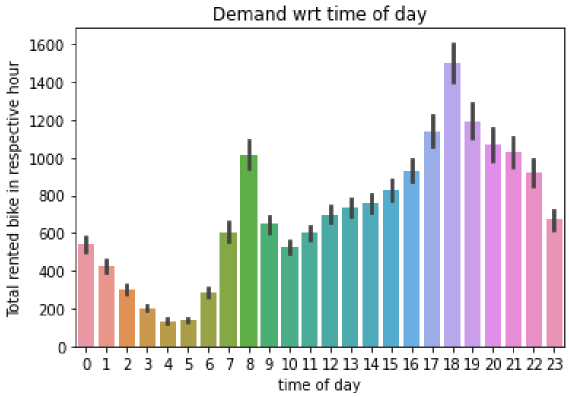
- Data understanding
- Data exploration
- Data preprocessing
- Feature engineering
- Model training
- Model validation
- Error analysis-Mean square error,mean absolute error,r2 score for test and train score,adjusted r2 score were calculated and compared for various models.

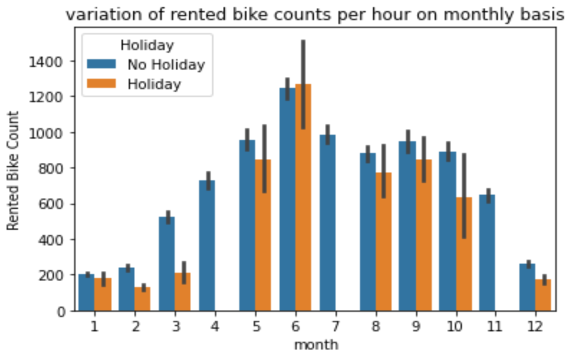
3.1.Data understanding and exploration

What hours of the day had the highest rented bike counts was determined.What month of the year had the highest rented bike counts.Whats is the difference in rented bike counts in those hours which had unusually high rainfall or snowfall.What was the count in hours which had heavy wind blowing. Which seasons had the highest bike counts.Was there any difference in count there was holiday and no holiday or the day was a functioning day or non functioning day.Graphs were drawn for better visualizations.Histograms ,barplots,pointplots and scatterplots were drawn using seaborn and matplotlib.







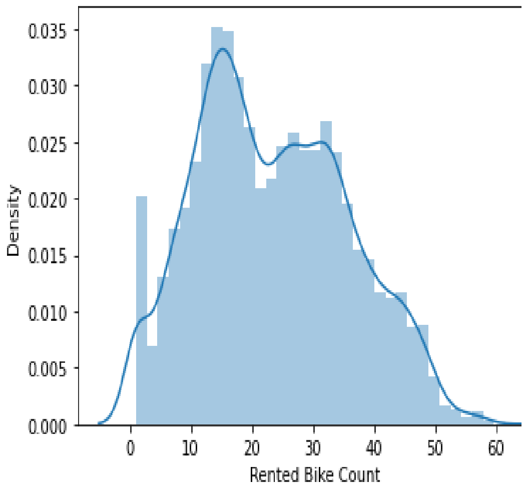


3.2.Data preprocessing

Our data didn’t have any null or missing values.Date was converted from object to datetime object and then months were extracted for analysis.

3.3.Feature engineering

- One hot encoding was used for categorical features.
- Z score pruning method was used to detect outliers.
- Square root transformation is used to convert dependent variable distribution to normal distribution.
- Collinearity was removed by using correlation coefficient.Vif was also calculated.

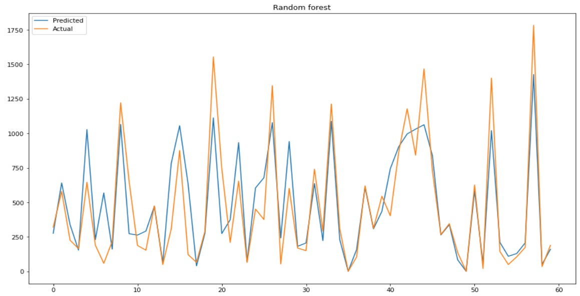


3.4.Model Training and validation

Data was split into train and test data .Cross validation as well as hyperparameter tuning both were used to get better results.The following regression models were used-

- Linear Regression
- Polynomial Regression
- Lasso Regression
- Ridge Regression
- Elastic net regression
- Decision tree regressor
- Random forest regressor
- Gradient boosting method
- XGboost regression model
- adaboost regression model
- Bagging regression model

3.5.Error Analysis



Mean absolute error,mean square error,R2 score adjusted R2 score was calculated to compare different model performances.

model	MSE	RMSE	MAE	Training Score	R2_score	Adjusted_R2
Linear regression	89591.33	299.41	202.52	0.78	0.78	0.78
Polynomial regression	53336.51	230.95	144.84	0.88	0.87	0.869
Lasso regression	89588.76	299.31	202.51	0.78	0.78	0.78
Ridge regression	89608.41	299.34	202.56	0.78	0.78	0.78
Elastic net regression	89614.45	299.35	202.56	0.78	0.78	0.78

Decision Tree regression	106589.86	326.48	210.79	0.77	0.74
Random forest regression	93720.39	306.14	198.11	0.81	0.7
Gradient boost method	52276.59	228.64	137.67	0.95	0.8
Adaboost	197572.58	444.49	311.62	0.6	0.5
Bagging	55946.63	236.53	140.28	0.96	0.8
Xgboost	52096.23	228.24	138.47	0.97	0.8

4.Conclusion-

1)We observed that the bike rental count is higher during no holidays than during holidays.

2)The rental bike count was at its peak at 8 AM in the morning and 6pm in the evening, an increasing trend can be observed from 5am to 8am, the graph touches the peak at 8am in the morning then dips a bit. Later we can see a gradual increase in the demand until 6pm, the demand is highest at 6 pm, and reduces until midnight.

3)People prefered to rent bikes when the temperature was between 20 degrees to 35 degrees celsius and even when it was a little windy.

4)Highest bike rental count was found in Autumn and summer seasons and the lowest is in the winter season.

5)Bike rentals were highest during clear days and lowest on snowy and rainy days.

6)When there was not a functioning day no bikes were booked.

7)Linear models such as linear regression ,lasso,ridge are performing good on both train and test data.R2 score is around 0.8 for both train and test set.

8)Polynomial regression is giving better results than linear models and giving a good r2 score on both test and train set.r2 score is around 0.85 for both train and test sets.

9)Ensemble models are giving quite good results on train data i.e R2 score more than 0.9 but are a little bit overfit i.e R2 score of 0.86-0.88 on test data.

10)Based on the feature importance plot, Temperature,whether day was functioning or non functioning,6:00pm evening time ,seasons were most important parameters in predicting the demand.

5.References

- Analytics vidhya
- Towards Data Science
- Stack overflow
- Geek for geeks