Seoul Bike Demand Prediction

Shokoufeh Naseri

Regression

Data

Seoul Bike Sharing Demand

Dimention

11,680 rows
24 columns

Target Variable

Number of bikes rented per hour

Features

weather Data

Temperature
Humidity
Windspeed
Visibility
Dewpoint
SolarRadiation
Snowfall
Rainfall

Date-Time Data

Hour
Day
Month
Year
Weekday
Holiday/Working day

Artificial

feat01
feat02
feat03
feat04
feat05
feat06
feat07
feat08
feat10

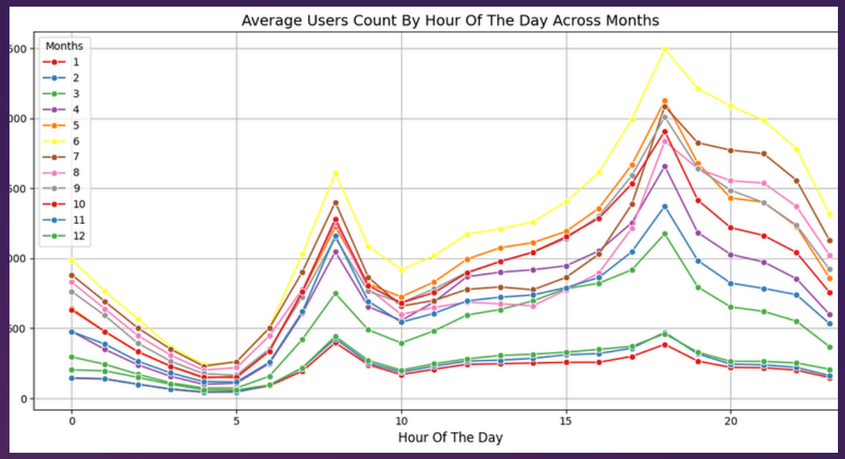
EDA

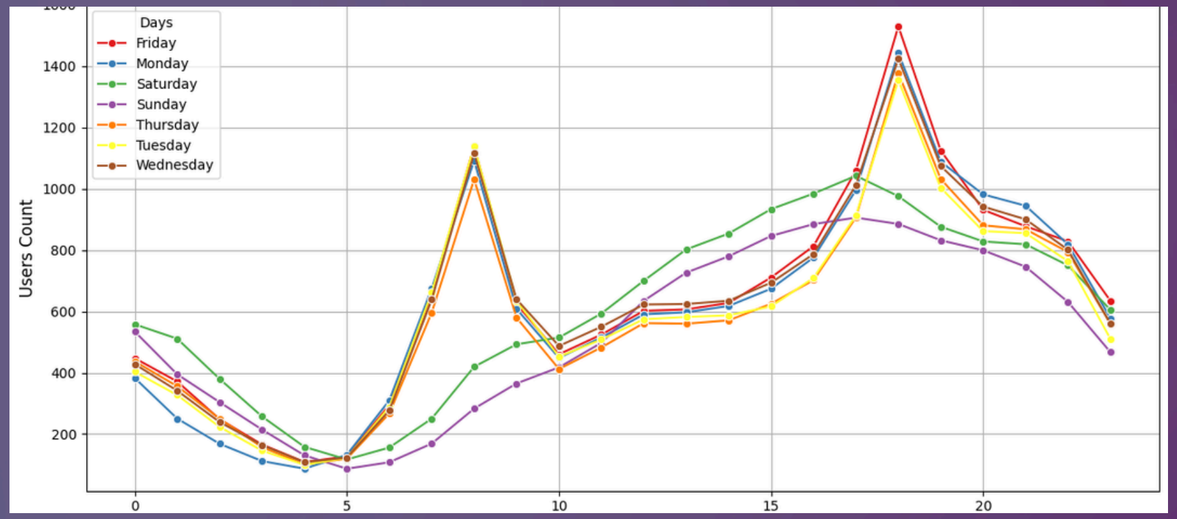
```
Data collected from : 2017-12-01 00:00:00
Data collected untill : 2018-11-30 00:00:00
```

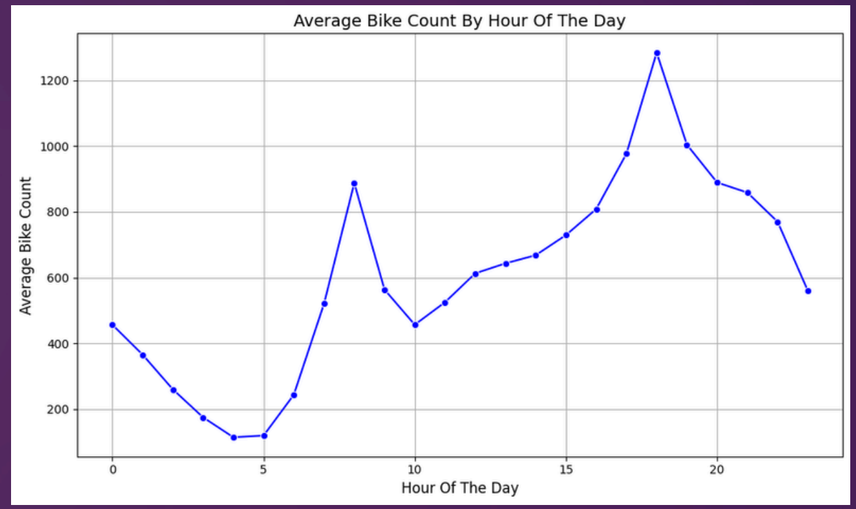
id: removed

Date:

replace with month and Day







Categorical features

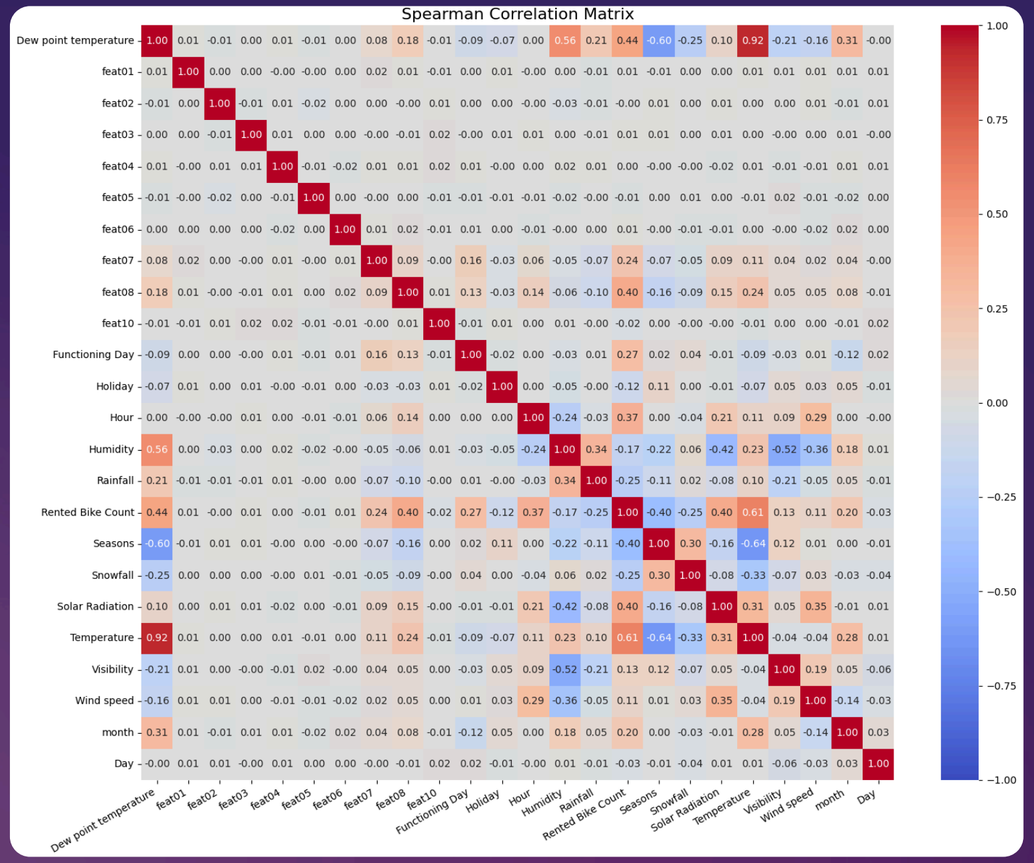
Transformed to numerical manually

Numerical features

Min&MaxScaler for scaling

Boruta algorithm

Boruta is a feature selection algorithm that identifies relevant variables by comparing their importance to shuffled shadow features using Random Forest



'Dew point temperature', 'feat07', 'feat08', 'Functioning Day', 'Holiday', 'Hour', 'Humidity', 'Rainfall', 'Seasons', 'Solar Radiation', 'Temperature', 'Visibility', 'Wind speed', 'month', 'Day'

Evaluation Metrics

RMSE

Root Mean Squared Error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

MAE

Mean Absolute Error

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MAPE

Mean Absolute Percentage Error

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

R²

R-squared

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Decision Tree Regression

```
--- Decision Tree Regression ---
Training Metrics:
  MSE: 97441.24
  RMSE: 312.16
  R²: 0.73
  MAE: 203.42
Testing Metrics:
  MSE: 97549.09
  RMSE: 312.33
  R²: 0.72
  MAE: 204.51
```

Tuned model using GridSearchCV

```
---  Tuned Decision Tree Regression ---
Training Metrics:
  MSE: 66008.71
  RMSE: 256.92
  R²: 0.82
  MAE: 157.81
Testing Metrics:
  MSE: 69987.43
  RMSE: 264.55
  R²: 0.80
  MAE: 164.00
```

Best Parameters: {'max_depth': 7, 'min_samples_leaf': 4, 'min_samples_split': 10}

Best MSE: 74526.47380022038

Bagging Regressor

```
---  Bagging Regressor ---
Training Metrics:
  MSE: 3602.15
  RMSE: 60.02
  R²: 0.99
  MAE: 31.97
Testing Metrics:
  MSE: 20990.66
  RMSE: 144.88
  R²: 0.94
  MAE: 81.21
```

Tuned model using GridSearchCV

```
--- Tuned Bagging Regressor ---
Training Metrics:
  MSE: 48107.55
  RMSE: 219.33
  R²: 0.87
  MAE: 137.08
Testing Metrics:
  MSE: 48958.37
  RMSE: 221.27
  R²: 0.86
  MAE: 141.04
```

RandomizedSearchCV

```
--- Tuned Bagging Regressor ---
Training Metrics:
  MSE: 23297.38
  RMSE: 152.63
  R²: 0.94
  MAE: 89.04
Testing Metrics:
  MSE: 28637.40
  RMSE: 169.23
  R²: 0.92
  MAE: 103.65
```

Best Parameters: {'estimator__max_depth': 7, 'estimator__min_samples_split': 2, 'max_features': 1.0, 'max_samples': 0.5, 'n_estimators': 100} Best Score (MSE): 54029.12414256516

Best Parameters: {'n_estimators': 100, 'max_samples': 1.0, 'max_features': 1.0, 'estimator__min_samples_split': 5, 'estimator__min_samples_leaf': 4, 'estimator__max_depth': 10} Best Score (MSE): 34623.884522274086

Gradient Boosting Regressor

```
--- Gradient Boosting Regressor ---
Training Metrics:
  MSE: 44171.62
  RMSE: 210.17
  R²: 0.88
  MAE: 137.36
Testing Metrics:
  MSE: 48892.45
  RMSE: 221.12
  R²: 0.86
  MAE: 146.13
```

Tuned model using GridSearchCV

```
--  Tuned Gradient Boosting Regressor -
raining Metrics:
 MSE: 23179.36
 RMSE: 152.25
 R²: 0.94
 MAE: 95.04
esting Metrics:
 MSE: 34848.93
 RMSE: 186.68
 R²: 0.90
 MAE: 117.14
```

Best Parameters: {'learning_rate': 0.05, 'max_depth': 7, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 100, 'subsample': 0.7}

Best Score (MSE): 36365.884009811074

Neural Network Model (NNL)

Epochs	Batch Size	Learning Rate	RMSE (Train)	RMSE (Test)	R^2 (Train)	R^2 (Test)	MAE (Train)	MAE (Test)
50	32	0.01	225.18	222.50	0.86	0.86	144.22	145.86
100	32	0.01	224.10	224.97	0.86	0.85	139.68	141.04
150	32	0.01	203.95	202.70	0.89	0.88	132.52	132.96
100	64	0.01	209.58	207.71	0.88	0.88	135.58	133.92
100	16	0.01	308.39	308.93	0.74	0.73	207.87	209.40
100	16	0.1	278.09	276.04	0.79	0.78	179.61	177.35

The model with 150 epochs, batch size 32, learning rate 0.01 performed the best (lowest RMSE, highest R^2).

CONCLUSION

	Model	Training MSE	Training RMSE	Training R^2	Training MAE	Testing MSE	Testing RMSE	Testing R^2	Testing MAE
0	Neural Network (epochs=150, batch_size=32, lr=0.01)	41593.650000	203.950000	0.890000	132.520000	41086.510000	202.700000	0.880000	132.960000
1	Decision Tree Regression	66008.710000	256.920000	0.820000	157.810000	69987.430000	264.550000	0.800000	164.000000
2	Bagging Regressor (Config 1)	48107.550000	219.330000	0.870000	137.080000	48958.370000	221.270000	0.860000	141.040000
3	Bagging Regressor (Config 2)	23297.380000	152.630000	0.940000	89.040000	28637.400000	169.230000	0.920000	103.650000
4	Gradient Boosting Regressor	23179.360000	152.250000	0.940000	95.040000	34848.930000	186.680000	0.900000	117.140000

The Bagging Regressor with tuned parameters achieved the best balance of training and testing metrics, showcasing the lowest RMSE (169.23) and highest R^2 (0.92).

n_estimators: 100

max_samples: 100% (or 1.0) of the data used in each bootstrap sample

max_features: 100% (or 1.0) of the features used to train each base estimator

estimator__min_samples_split: 5 (minimum number of samples required to split an internal node)

estimator__min_samples_leaf: 4 (minimum number of samples required to be at a leaf node)

estimator__max_depth: 10 (maximum depth of each decision tree in the ensemble)

THANK YOU!