

Capstone Project Bike Sharing Demand Prediction

By Nayana Pradeep



PROBLEM STATEMENT

- Given is a rich dataset of the Seoul Bike Demand Analysis which includes the weather condition details with number of bikes rent every hour for an entire year.
- Explore and analyze the dataset to have an understanding on how the bike rental demand change with the weather and time.
- Predict the count of bikes being rented every hour.



OVERVIEW

Bike riding has become quite popular in last few years owing to the ecofriendly and economical reasons.

South Korea is actually one of the best countries in the world to explore by bicycle, and almost no one knows, not even cyclists!

So it is no surprise that the bike rental services keep in check with the supply and demand of bikes.





DATA SUMMARY

- Date : year-month-day
- Rented Bike count Count of bikes rented at each hour
- Hour Hour of the 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 and No holiday
- ☐ Functional Day Non Functional Days and Functional Days

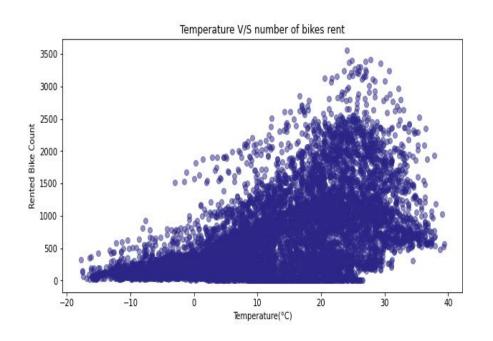


EXPLORATORY DATA ANALYSIS

- ♦ 8760 rows and 14 columns
- No null values or duplicates
- Date, Hour, Seasons, Functioning Day and Holiday are the categorical features
- People rent most number of bikes during the summer season and least during the winter season.
- ❖ Bike rental demand is less on holidays. This indicates that people prefer to use these bikes as mode of transportation to work.



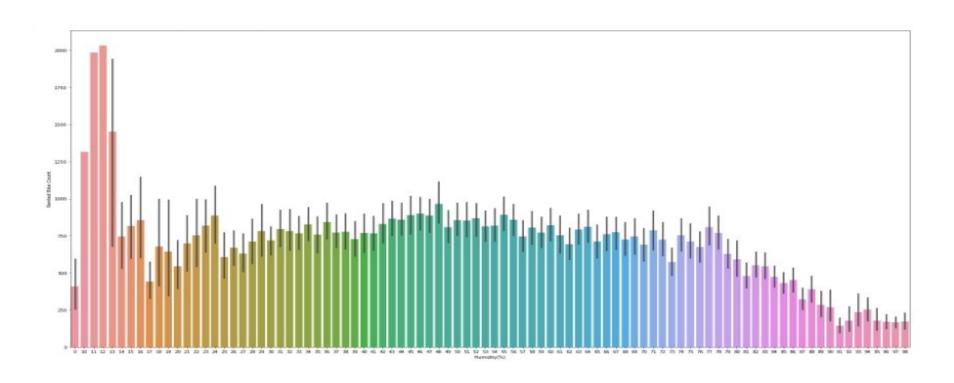
TEMPERATURE AND BIKE COUNT



- Rent Bike Count has an almost linear relationship with the temperature.
- ☐ The conclusion of increasing demand during the summer is clear.

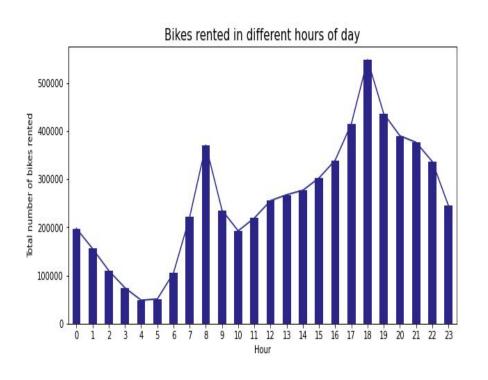


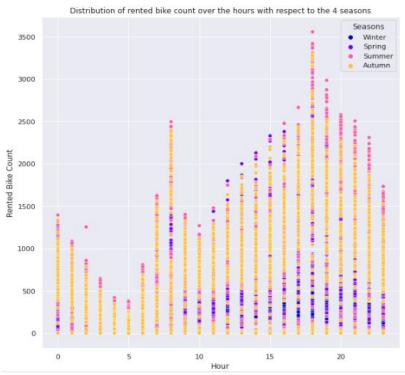
HUMIDITY AND BIKE COUNT





SEASONS AND HOURS WITH BIKE COUNT

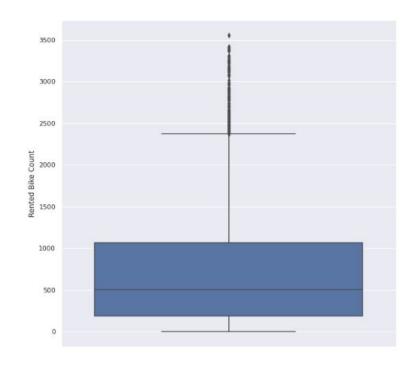






FEATURE EXTRACTION AND OUTLIERS

- Day, month , year and weekday extracted from date column and made new columns out of it
- Enabled better analysis
- Median of the Rent Bike Count is 500
- ☐ Less number of outliers





LABEL ENCODING

- ☐ Holiday
- Functioning Day
- Seasons
- ☐ Day of the week





CORRELATION HEATMAP

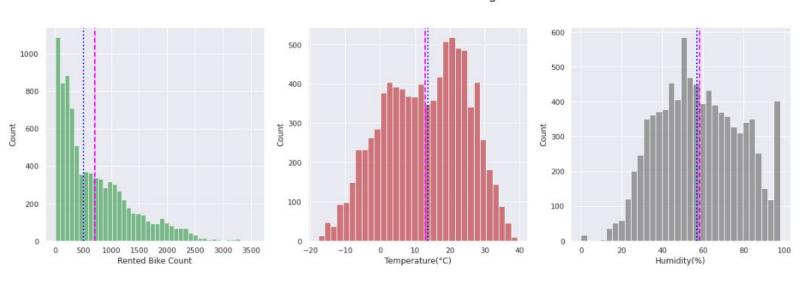
| Rented Bike Count | 1 | 0.41 | 0.54 | -0.2 | 0.12 | | 0.38 | 0.26 | -0.12 | -0.14 | -0.25 | 0.072 | 0.2 | | 0.071 | 0.047 | -0.014 |
|---------------------------|-------------------|---------|-----------------|-------------|------------------|------------------|--------------------|-------------------|--------------|---------------|---------|---------|-----------------|---------|---------|---------|-------------|
| Hour | 0.41 | 1 | 0.12 | -0.24 | | 0.099 | 0.0031 | 0.15 | 0.0087 | -0.022 | 9.5e-18 | | 0.0054 | | 2.7e-18 | 1 6e-19 | 4.1e-19 |
| Temperature(°C) | 0.54 | 0.12 | 1 | 0.16 | -0.036 | 0.035 | 0.91 | 0.35 | 0.05 | -0.22 | -0.34 | | -0.05 | 0.38 | | 0.081 | -0.02 |
| Humidity(%) | -0.2 | -0.24 | 0.16 | 1 | -0.34 | -0.54 | 0.54 | -0.46 | | 0.11 | -0.12 | | -0 021 | 0.036 | 0.048 | 0.007 | -0.0071 |
| Wind speed (m/s) | 0.12 | | -0.036 | -0.34 | 1 | | 0.18 | 0.33 | -0.02 | 0.0036 | 0.11 | -0.023 | | -0.0038 | -0.082 | -0.061 | 0.049 |
| Visibility (10m) | 0.2 | | | -0.54 | 0.17 | 1 | -0.18 | 0.15 | -0.17 | -0.12 | -0.015 | -0.032 | -0.026 | 0.052 | 0.078 | -0.11 | 0.038 |
| Dew point temperature(°C) | 0.38 | 0.0031 | 0.91 | 0.54 | -0.18 | -0.18 | 1 | 0.094 | 0.13 | -0.15 | -0.33 | | -0.053 | | | | -0.021 |
| Solar Radiation (MJ/m2) | 0.26 | | | -0.46 | | | 0.094 | 1 | -0.074 | -0.072 | -0.076 | 0.0051 | -0.0077 | | -0.03 | -0.0034 | -0.011 |
| Rainfall(mm) | -0.12 | 0.0087 | | | -0.02 | -0.17 | 0.13 | -0.074 | 1 | 0.0085 | -0.019 | 0.014 | 0.0021 | 0.028 | -0.023 | 0.012 | 0.035 |
| Snowfall (cm) | -0.14 | -0.022 | -0.22 | 0.11 | -0.0036 | -0.12 | -0.15 | -0.072 | 0.0085 | 1 | 0.15 | 0.013 | | -0.21 | | | 0.037 |
| Seasons | -0.25 | 9.5e-18 | -0.34 | -0.12 | 0.11 | -0.015 | -0.33 | -0.076 | -0.019 | 0.15 | 1 | -0.046 | | -0.41 | -0.22 | -0.066 | 0.0025 |
| Holiday | 0.072 | | | | -0.023 | -0.032 | | 0.0051 | 0.014 | 0.013 | -0.046 | 1 | 0.028 | 0.11 | 0.0091 | -0.016 | 0.035 |
| Functioning Day | 0.2 | 0.0054 | -0.05 | -0.021 | | -0.026 | -0.053 | -0.0077 | 0.0021 | 0.032 | | 0.028 | 1 | -0.057 | -0.051 | 0.04 | 0.018 |
| Year | 0.22 | | 0.38 | 0.036 | -0.0038 | 0.052 | | 0.13 | 0.028 | -0.21 | -0.41 | 0.11 | -0.057 | 1 | -0.3 | -0.083 | -0.025 |
| Month | 0.071 | 2.7e-18 | | 0.048 | -0.082 | 0.078 | 0.065 | -0.03 | -0.023 | | -0.22 | 0.0091 | -0.051 | -0.3 | 1 | 0.012 | -0.0081 |
| Day | 0.047 | 1.6e-19 | 0.081 | 0.007 | -0.061 | -0.11 | | -0.0034 | 0.012 | 0.1 | -0.066 | -0.016 | | -0.083 | 0.012 | 1 | 0.006 |
| Day_of_Week | -0.014 | 4.1e-19 | -0.02 | -0.0071 | | 0.038 | -0.021 | -0.011 | 0.035 | 0.037 | 0.0025 | | | -0.025 | -0.0081 | 0.006 | 1 |
| | Rented Bike Count | Hour | Temperature(°C) | Humidity(%) | Wind speed (m/s) | Visibility (10m) | nt temperature(°C) | Radiation (MJ/m2) | Rainfall(mm) | Snowfall (cm) | Seasons | Holiday | Functioning Day | Year | Month | Day | Day_of_Week |

-1.00- 0.75 - 0.50 - 0.25 - 0.00 - -0.25 --0.50 --0.75



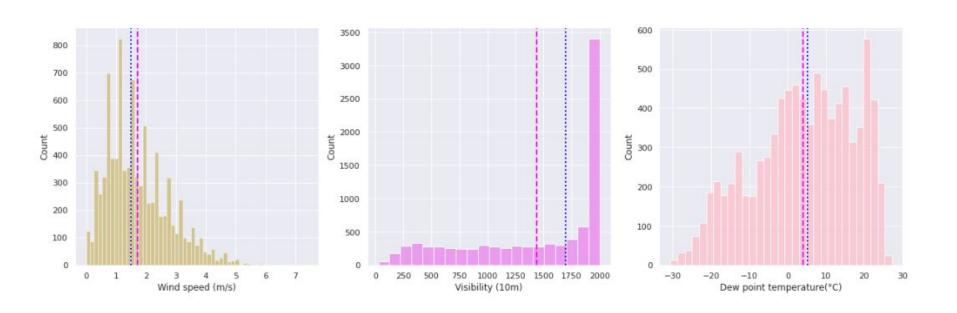
FEATURE DISTRIBUTION

Feature Distribution and Target



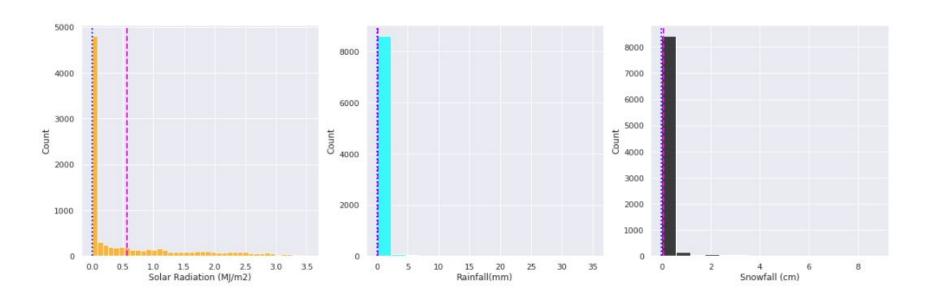


FEATURE DISTRIBUTION





FEATURE DISTRIBUTION





CORRELATION WITH BIKE RENT COUNT

| Correlation | with | Rented | Bike | Count |
|-------------|------|--------|------|-------|
| | | | | |

| | Correlation with Kented Dike Count |
|---------------------------|------------------------------------|
| Rented Bike Count | 1.000000 |
| Temperature(°C) | 0.538558 |
| Hour | 0.410257 |
| Dew point temperature(°C) | 0.379788 |
| Solar Radiation (MJ/m2) | 0.261837 |
| Year | 0.215162 |
| Functioning Day | 0.203943 |
| Visibility (10m) | 0.199280 |
| Wind speed (m/s) | 0.121108 |
| Holiday | 0.072338 |
| Month | 0.070861 |
| Day | 0.046849 |
| Day_of_Week | -0.014004 |
| Rainfall(mm) | -0.123074 |
| Snowfall (cm) | -0.141804 |
| Humidity(%) | -0.199780 |
| Seasons | -0.253058 |
| | |



LINEAR REGRESSION MODEL

□ MAE : 5.660417893430779
 □ MSE : 54.05488568978613
 □ RMSE : 7.352202778064961
 □ R2 : 0.6578077075559197

LASSO REGRESSION MODEL

□ MAE : 5.660956563638153
 □ MSE : 54.0561990479736
 □ RMSE : 7.352292094848626
 □ R2 : 0.6577993933944293



RIDGE REGRESSION MODEL

□ MAE : 5.703254454109305
 □ MSE : 54.60897217620785
 □ RMSE : 7.389788371544062
 □ R2 : 0.6543000852090879

GRADIENT BOOSTING

□ MAE : 3.1410591115406703□ MSE : 19.487435690184608□ RMSE : 4.4144575759865

□ R2 : 0.8766355675794004



XGBOOST MODEL

□ MAE : 3.1139086468938473
 □ MSE : 19.202553766527622
 □ RMSE : 4.382071857754916
 □ R2 : 0.8784390012059464

RANDOM FOREST MODEL

□ MAE : 2.4127930867482856
 □ MSE : 13.38508831250127
 □ RMSE : 3.658563695290991
 □ R2 : 0.9152662336480208



HYPERPARAMETER TUNING

| XGBoost – GridSearchCV | max_depth | 8 |
|--------------------------------------|-------------------|------|
| | min_samples_leaf | 30 |
| | min_samples_split | 10 |
| | n_estimators | 100 |
| Random Forest - GridSearchCV | max_depth | 12 |
| | min_samples_leaf | 30 |
| | min_samples_split | 10 |
| | n_estimators | 100 |
| Random Forest – RandomizedSearchCV | bootstrap | TRUE |
| Nandom i orest – Nandomized Search V | | |
| | max_depth | None |
| | max_features | Auto |
| | n estimators | 11 |



XGBOOST GRIDSEARCHCV

□ MAE : 2.0239312402965672
 □ MSE : 10.00143590368713
 □ RMSE : 3.1625046883265058
 □ R2 : 0.9366863099247676

RANDOM FOREST GRIDSEARCHCV

□ MAE : 3.240885412547648
 □ MSE : 21.01221206348057
 □ RMSE : 4.58390794666304
 □ R2 : 0.8669830317172941



RANDOM FOREST RANDOMIZEDSEARCHCV

MAE
 MSE
 RMSE
 3.930577820354604
 R2
 0.9021979251618604



MODEL COMPARISON

| | MODEL NAME | MAE | MSE | RMSE | R2 |
|---|-----------------------------------|----------|-----------|----------|----------|
| 0 | Linear Regression | 5.660418 | 54.054886 | 7.352203 | 0.657808 |
| 1 | Lasso Regression | 5.660957 | 54.056199 | 7.352292 | 0.657799 |
| 2 | Ridge Regression | 5.703254 | 54.608972 | 7.389788 | 0.654300 |
| 3 | Gradient Boosting | 3.140610 | 19.483598 | 4.414023 | 0.876660 |
| 4 | XGBoost | 3.113909 | 19.202554 | 4.382072 | 0.878439 |
| 5 | Random Forest | 2.404020 | 13.401257 | 3.660773 | 0.915164 |
| 6 | XGBoost- GridSearchCV | 2.023931 | 10.001436 | 3.162505 | 0.936686 |
| 7 | Random Forest- GridSearchCV | 3.252191 | 21.248756 | 4.609637 | 0.865486 |
| 8 | Random Forest- RandomizedSearchCV | 2.554616 | 15.153617 | 3.892765 | 0.904071 |
| | | | | | |



CONCLUSION

- ★ XGBoost model with hyperparameter tuning promises the best accuracy among the experimented models with R2 score of 0.9366 and the least error rate.
- This is followed by Random Forest with R2 score of 0.9151.
- Least accurate model is linear regression model with an accuracy of 65.7%.



CHALLENGES

- Large dataset
- Lot of analyzing and visualization required to unfold the relationship between the features



THANK YOU