Capstone Project

Seoul Bike Sharing Demand Prediction

by

Suraj Singh

Content

- Problem Statement
- Data Description
- Data Preparation
- EDA
- Model Deployment
- Observations
- Conclusions
- References

Problem Statement

In this project we'll be using the Seoul Bike Share Demand dataset to:

- Understand bike share use trends.
- Apply machine learning techniques to predict the number of bikes rented at any given hour using the city weather and date information.
- Provide reasonable explanations of the best predicting model to understand factors affecting bike share demands.

Data Description

Feature	Type	Measurement
Date	year-month-day	-
Rented Bike Count	Continuous	0,1,2,...3556
Hour	Continuous	0,1,2,..24
Temperature(°C)	Continuous	Celsius
Humidity(%)	Continuous	%
Wind speed (m/s)	Continuous	m/s
Visibility (10m)	Continuous	m

Feature	Type	Measurement
Solar Radiation (MJ/m2)	Continuous	mJ/m^2
Rainfall(mm)	Continuous	mm
Snowfall (cm)	Continuous	cm
Holiday	Categorical	Holiday/ No Holiday
Seasons	Categorical	Summer, Winter, Spring, Autumn
Functioning Day	Categorical	Yes/No
Month	Count	1,2,3..12

Feature	Type	Measurement
Weekend	Categorical	Yes/No
Day Phase	Categorical	Morning, Afternoon, Evening, Night
Sin Month	Continuous	[-1,1]
Cos Month	Continuous	[-1,1]
Sin Day	Continuous	[-1,1]
Cos Day	Continuous	[-1,1]
Sin Hour	Continuous	[-1,1]
Cos Hour	Continuous	[-1,1]

Data Preparation

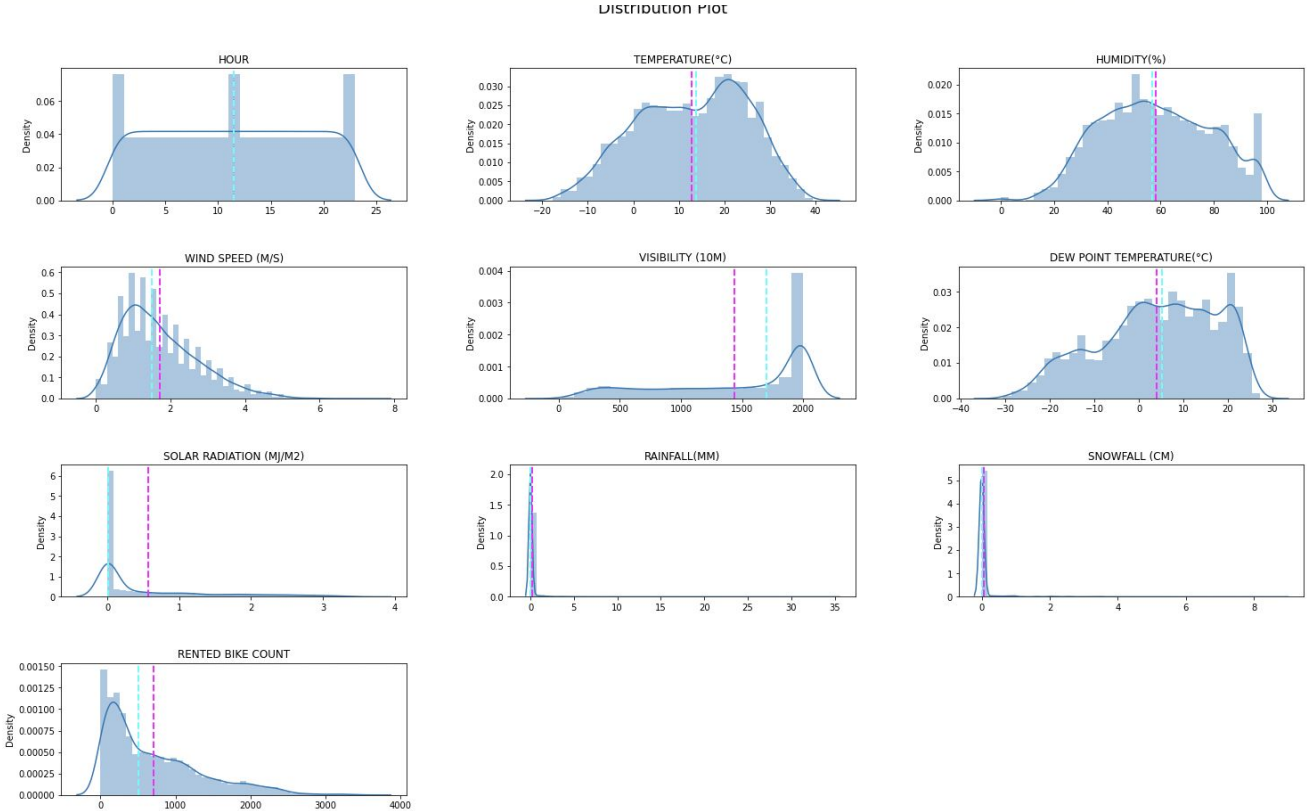
- Date attribute was converted from type string to type datetime
- Features such as day of the week, month and weekend-weekdays were extracted from Date after conversion.
- In order to reflect the cyclic nature of weekdays and time, the features underwent sine and cosine transformations,
- Hour of the day was used to categorise the time of the day into Morning,Afternoon, Evening and Night

Exploratory Data Analysis

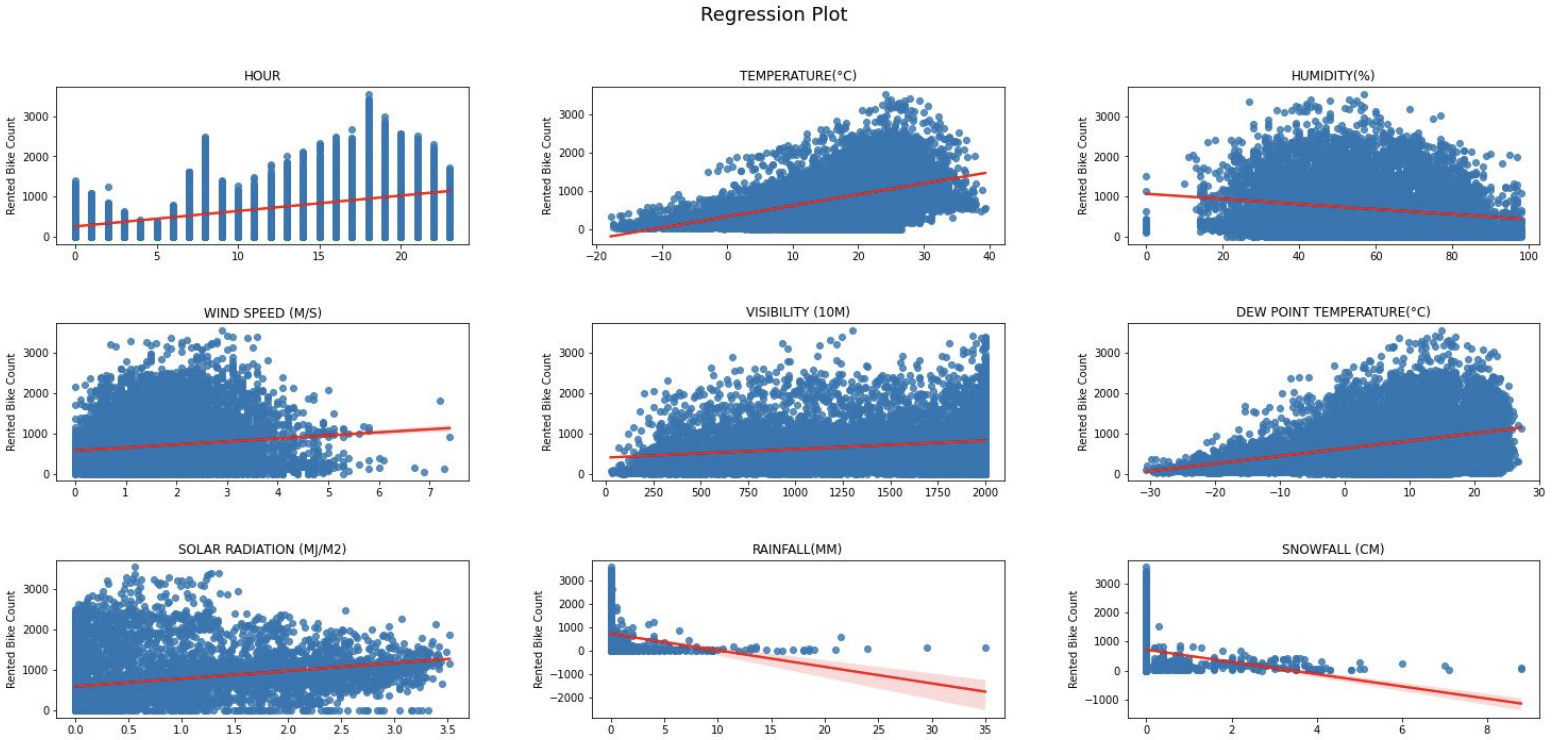
In this part of the project, we inspected and explored:

- Physical Characteristics of the Dataset
- Characteristics of the Numerical and Categorical Attributes
- Relationships between the attributes
- Relationship between Target variable and Independent variables

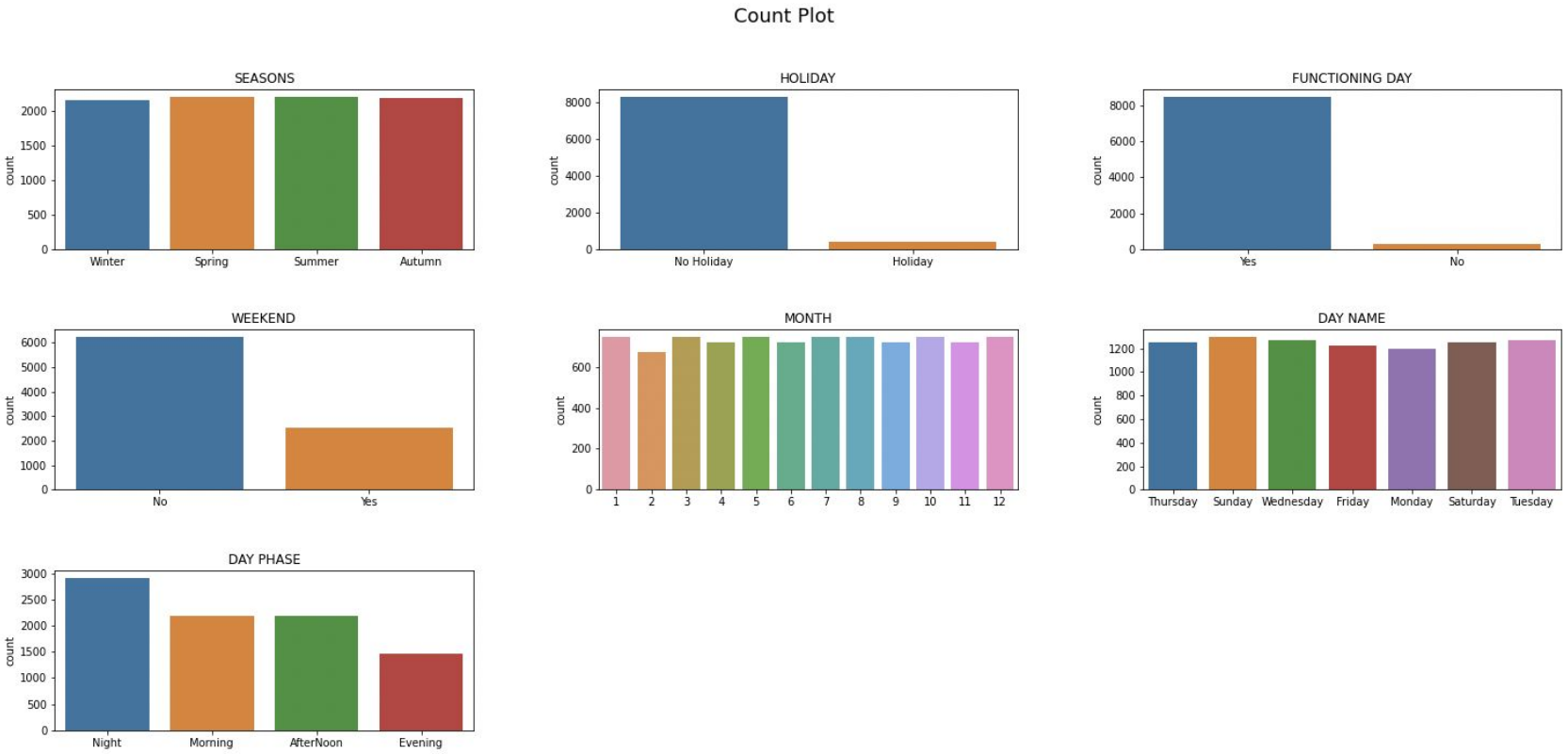
Distribution Plot of Numerical Values



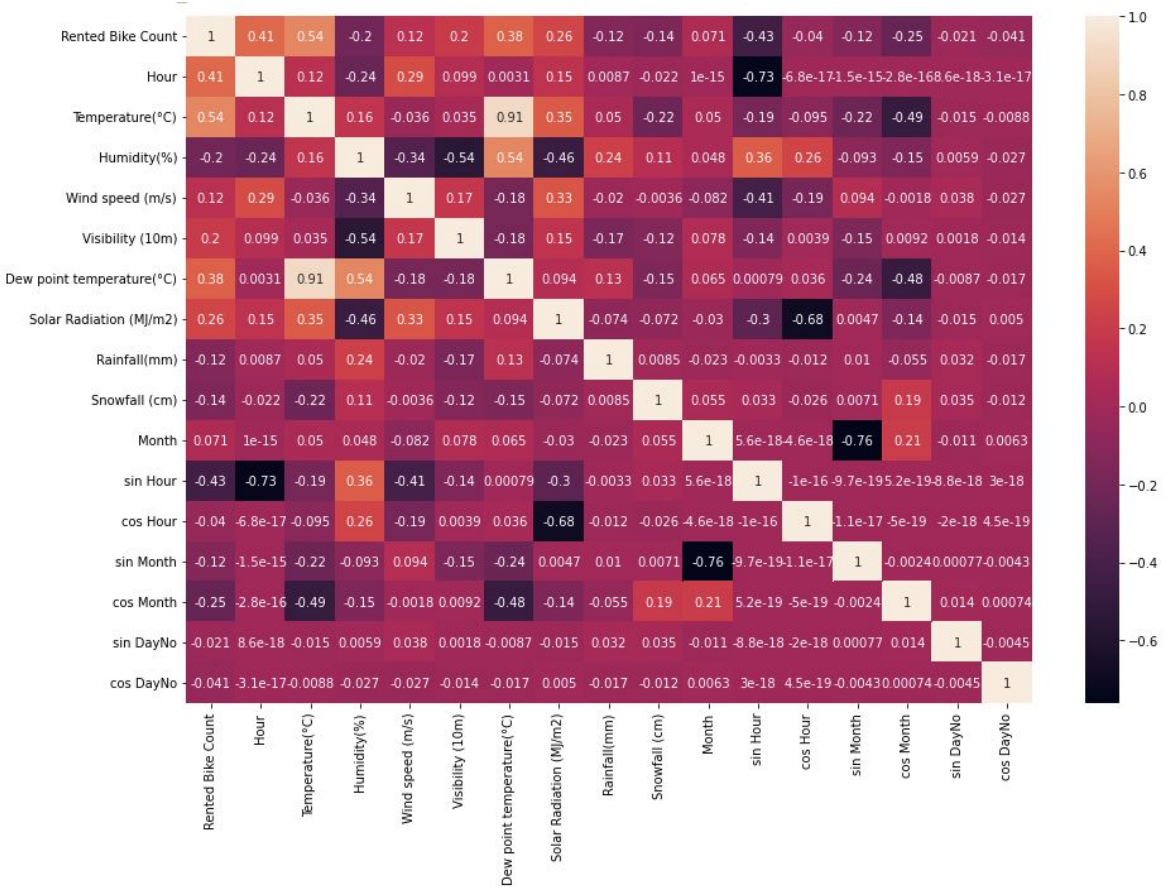
Regression Plot of Numerical Features and Bike Rentals



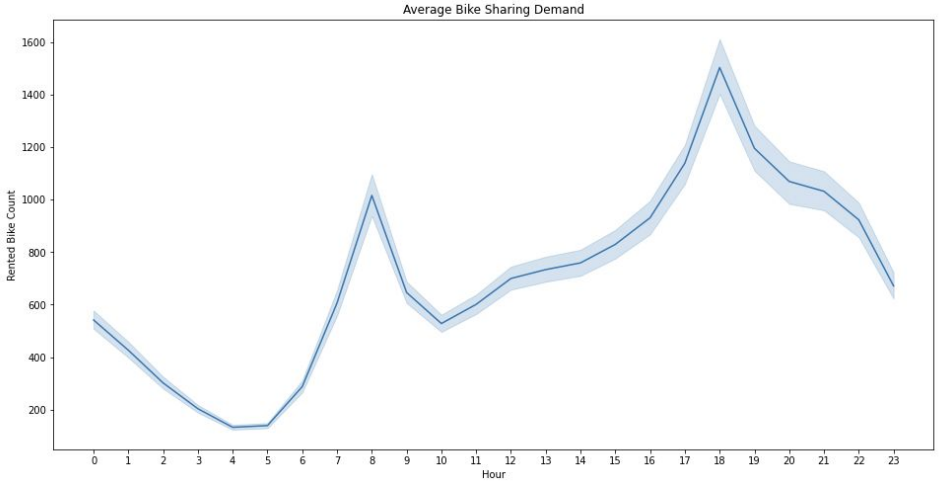
Count Plot of Categorical Variables



Correlation Heatmap

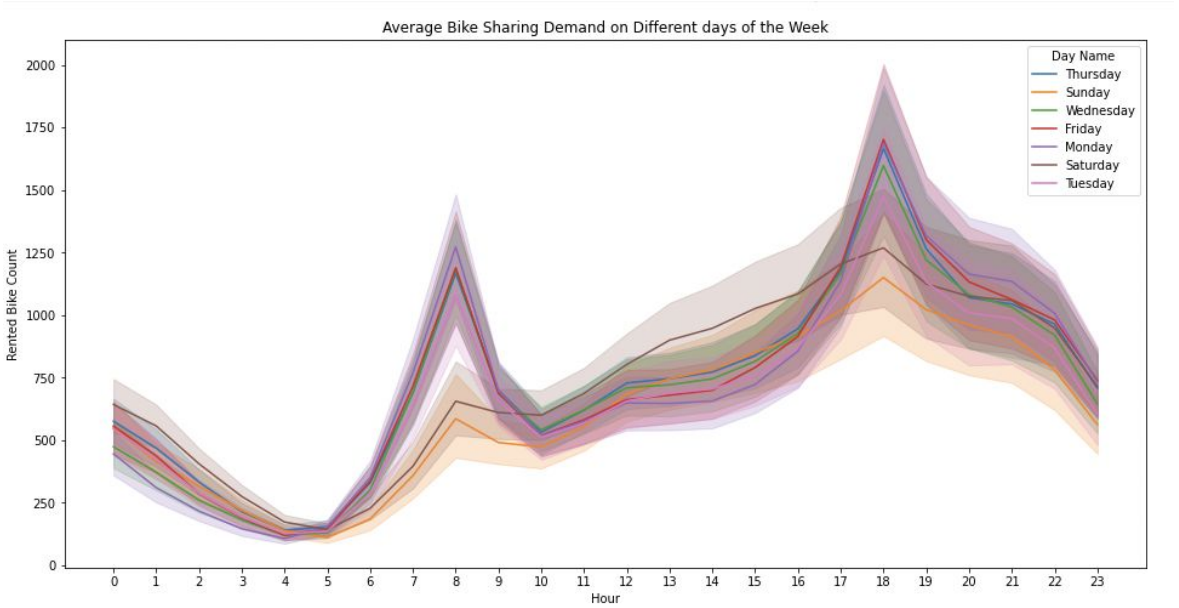


Bike Sharing Trend on an Average Day in Seoul



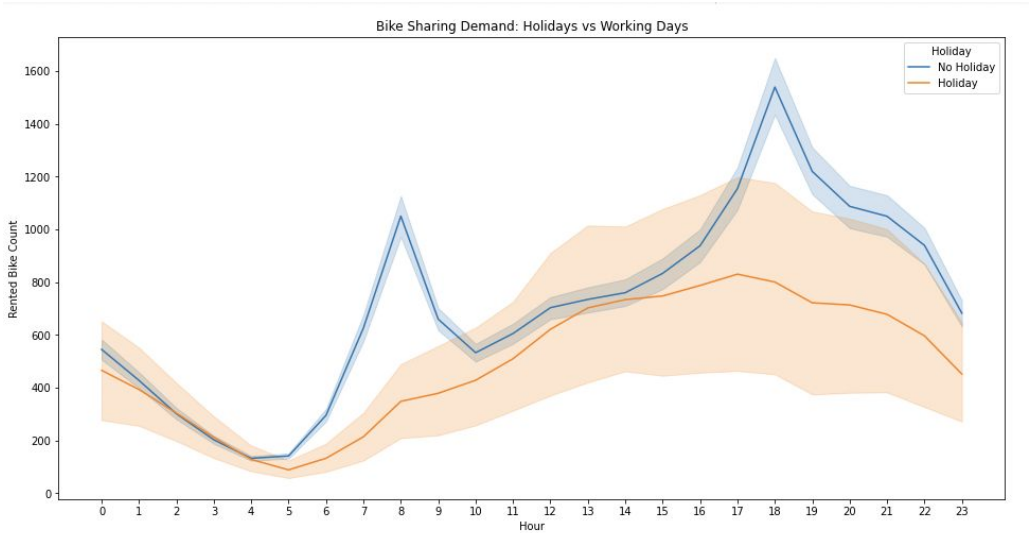
- The busiest times of the day are from 7am to 9am in the morning and 5pm to 7pm in the evening
- The relatively idle hours are witnessed from 11pm to 6am in the morning

Bike Sharing Demand on Different Weekdays



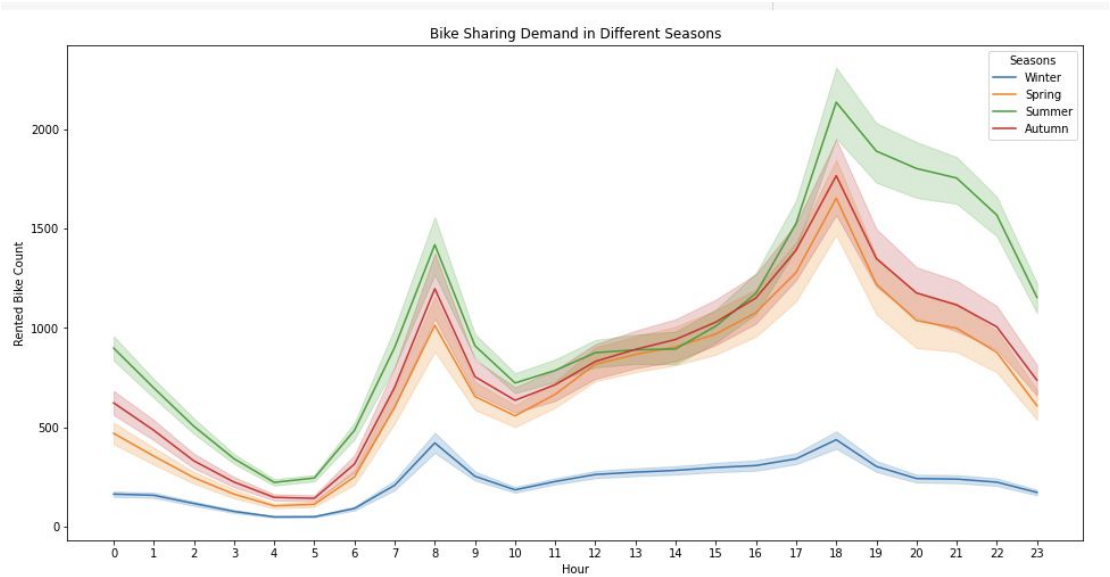
- The busiest days of the week are Fridays and Mondays
- The weekends experience a significantly low demand in Bike Rentals

Effect of Holidays on Bike Sharing Demand



- We can observe that bike demands increase and decrease gradually during holidays
- The mornings and evenings are the peak hours for bike rentals on non-holidays
- This indicates that notable contributions in bike rentals are made from the working class section of the population

Bike Sharing Demand in Different Seasons



- Summer sees the highest demand in bike rentals
- Spring and Autumn have comparable bike rental demand in the mornings but are lower in the evenings
- A significantly low number of Bike Rentals are observed in Winters. This season might be a good season for re-servicing and upgrading docking stations.

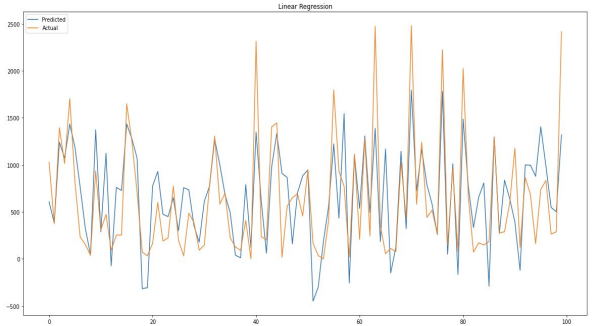
Model Deployment

The performance of ten machine learning algorithms are evaluated and compared to predict bike rental demand in the city of Seoul using the dataset at any given hour.

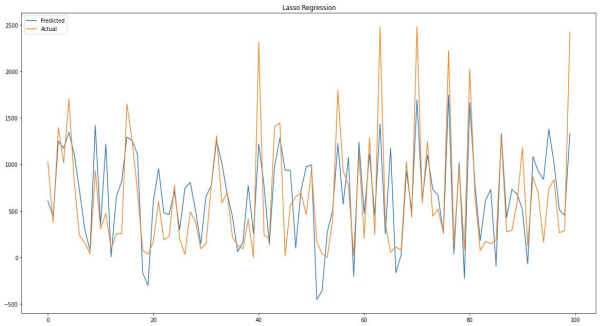
The metrics calculated for evaluating performance are RMSE, MAE, R^2 and adjusted R^2

The top models are picked to tune hyperparameters in order to further optimize their results. These models are also stacked and evaluating using Stacking Ensemble.

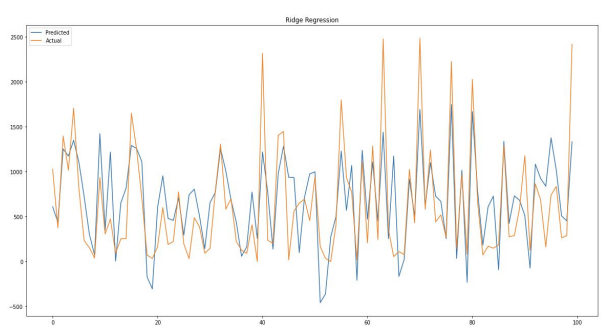
Linear Models



Linear Regression



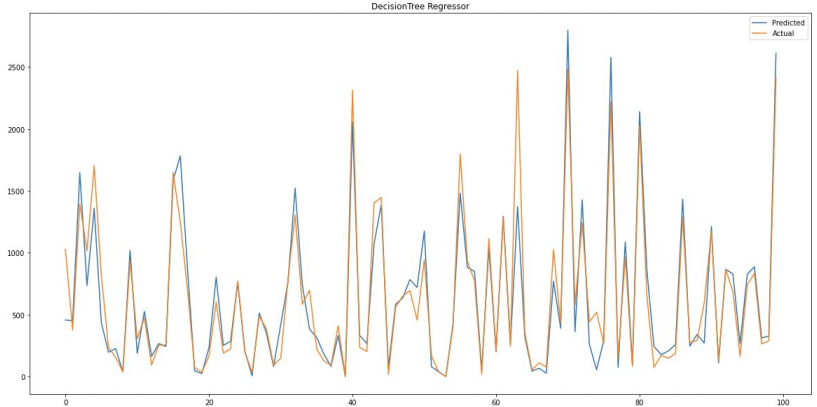
Lasso Regression



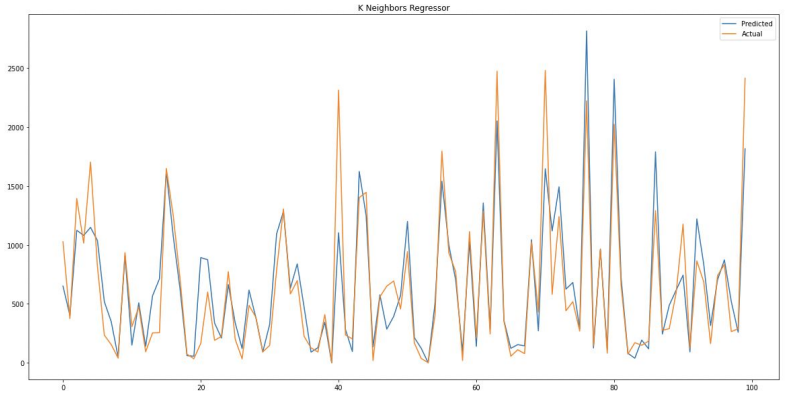
Ridge Regression

Model Name	RMSE	MAE	R^2	Adjusted R^2
Linear Regression	408.7252	312.7789	0.5983	0.5887
Ridge Regression	400.7197	303.3573	0.6139	0.6046
Lasso Regression	400.6544	303.2638	0.6140	0.6047

Non-Linear Models



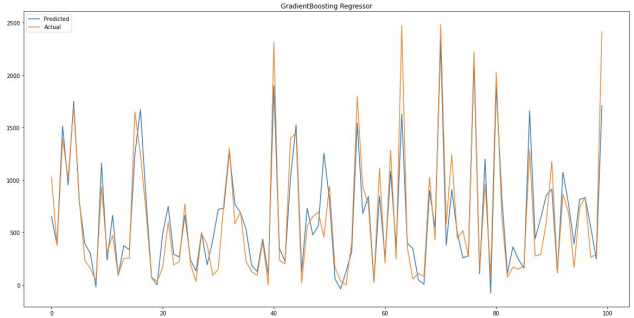
Decision Tree



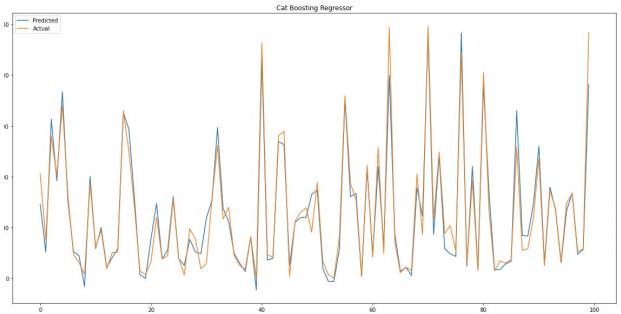
K Nearest Neighbors

Model Name	RMSE	MAE	R^2	Adjusted R^2
Decision Tree Regressor	294.2235	165.7871	0.7918	0.7868
K Neighbors Regressor	301.9540	195.3866	0.7808	0.7755

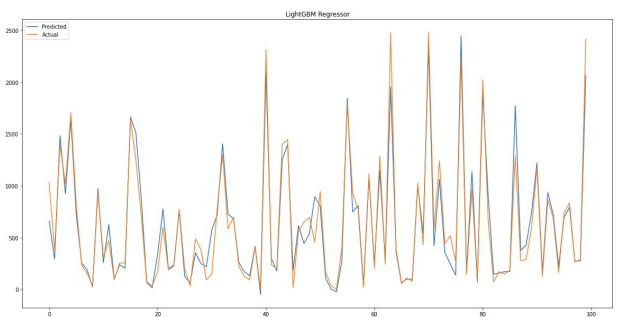
Boosting Ensembles



Gradient Boosting



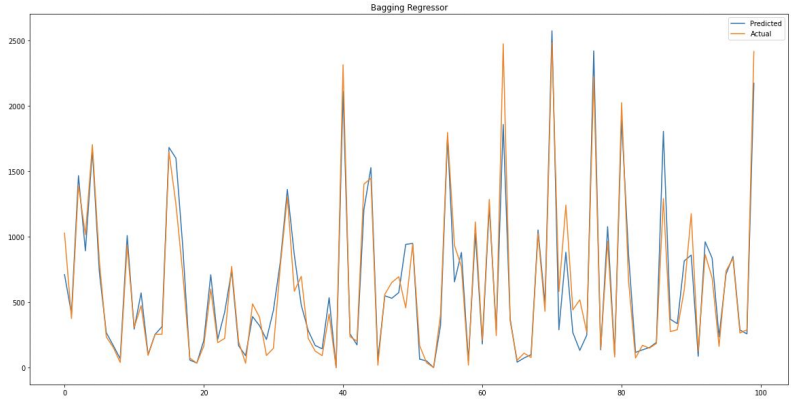
Cat Boosting



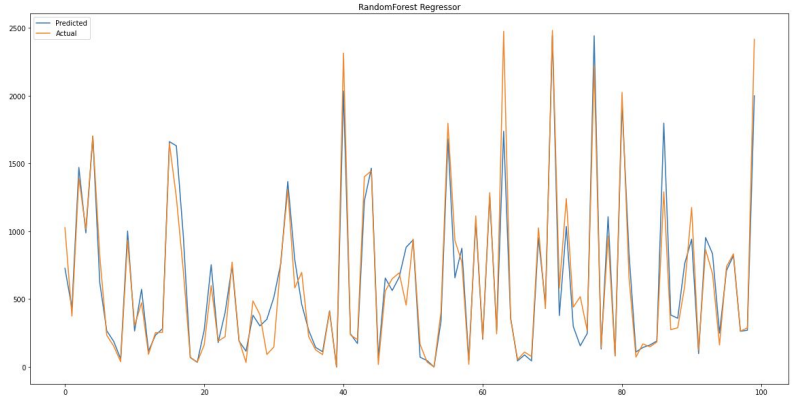
Light Gradient Boosting

Model Name	RMSE	MAE	R2	Adjusted R2
Cat Boosting Regressor	178.3280	108.5684	0.9235	0.9217
LightGBM Regressor	194.2324	119.3688	0.9093	0.9071
Gradient Boosting Regressor	246.4315	163.3944	0.8540	0.8505

Bagging Ensembles



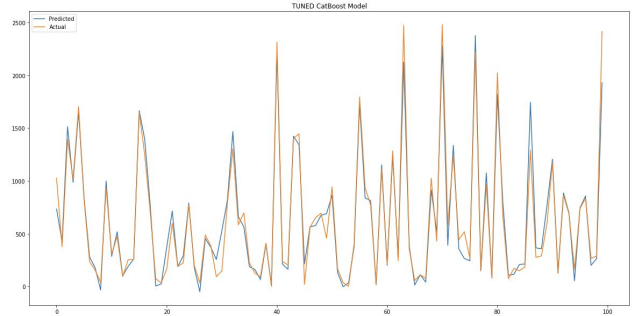
Bagging



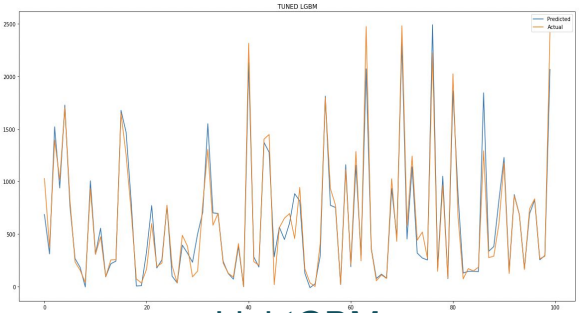
Random Forest

Model Name	RMSE	MAE	R2	Adjusted R2
Random Forest Regressor	206.2215	120.6033	0.8977	0.8953
Bagging Regressor	214.7691	127.2935	0.8891	0.8864

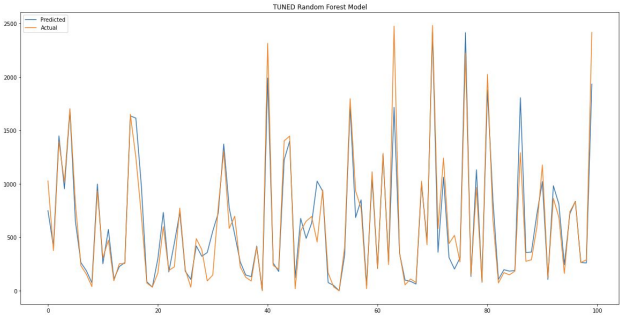
Hyperparameter Tuning



CatBoosting



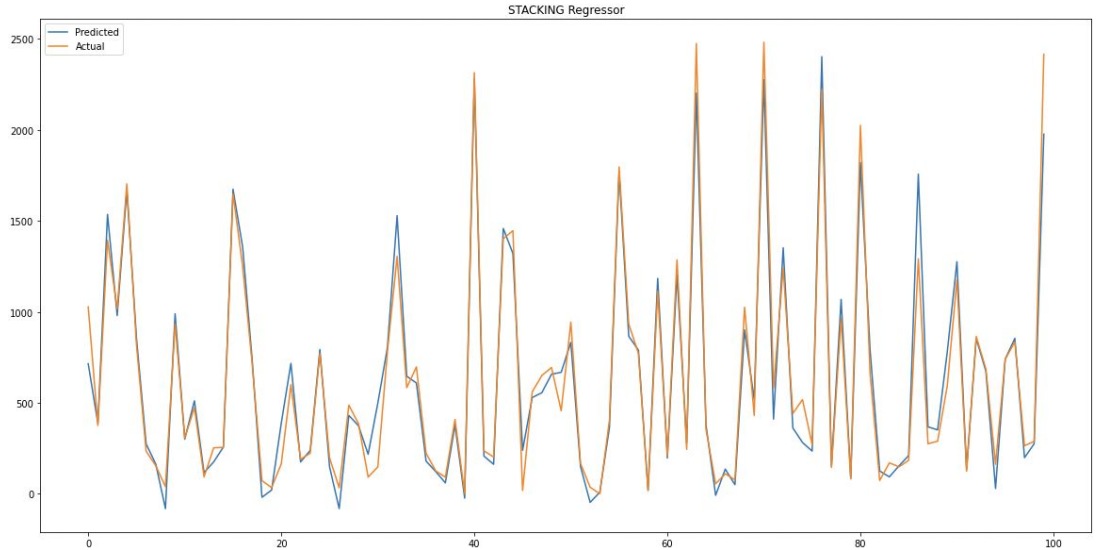
LightGBM



Random Forest

Model Name	RMSE	MAE	R2	Adjusted R2
TUNED CatBoost Model	162.0191	96.6193	0.9369	0.9354
Cat Boosting Model	178.3280	108.5684	0.9235	0.9217
TUNED LightGBM Model	180.5627	107.3737	0.9216	0.9197
LightGBM Model	194.2324	119.3688	0.9093	0.9071
TUNED Random Forest Model	200.6814	120.4977	0.9032	0.9008
Random Forest Model	206.2215	120.6033	0.8977	0.8953

Stacking Ensemble



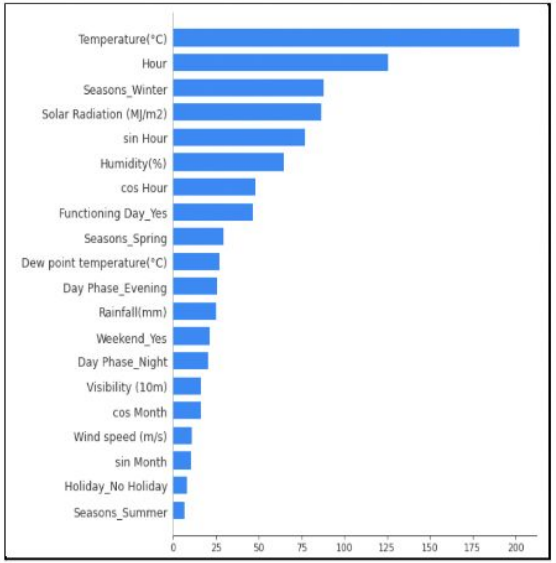
Model Name	RMSE	MAE	R2	Adjusted R2
Stacking Regressor	160.5961	96.9616	0.9380	0.9365

Evaluation Report

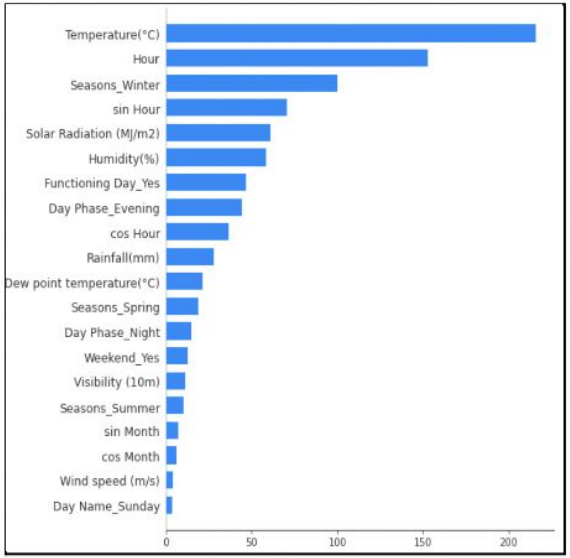
Model Name	RMSE	MAE	R2	Adjusted R2
Stacking Model	160.5961	96.9616	0.9380	0.9365
TUNED CatBoost Model	162.0191	96.6193	0.9369	0.9354
Cat Boosting Model	178.3280	108.5684	0.9235	0.9217
TUNED LightGBM	180.5627	107.3737	0.9216	0.9197
LightGBM	194.2324	119.3688	0.9093	0.9071
TUNED Random Forest Model	200.6814	120.4977	0.9032	0.9008
Random Forest Model	206.2215	120.6033	0.8977	0.8953
Bagging Model	214.7691	127.2935	0.8891	0.8864
Gradient Boosting Model	246.4315	163.3944	0.8540	0.8505
Decision Tree Model	294.2235	165.7871	0.7918	0.7868
K Nearest Neighbors	301.9540	195.3866	0.7808	0.7755
Lasso Regression	400.6544	303.2638	0.6140	0.6047
Ridge Regression	400.7197	303.3573	0.6139	0.6046
Linear Regression	408.7252	312.7789	0.5983	0.5887

- Stacking Ensemble and tuned CatBoost Model produced results with highest R^2 score and lowest errors.

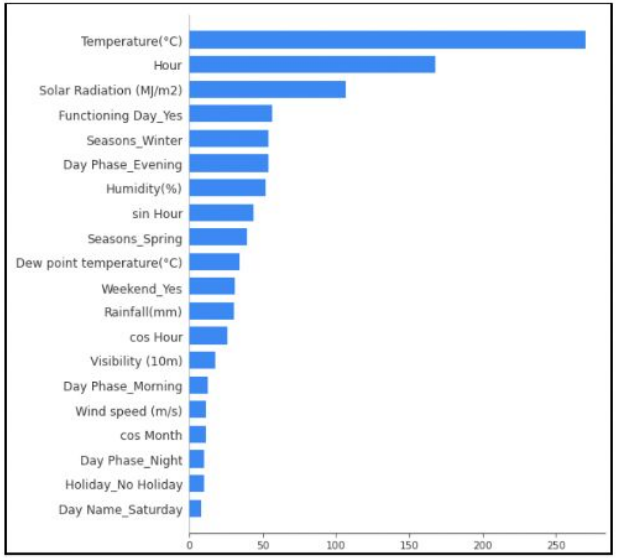
Feature Importance



CatBoosting



Random Forest



LightGBM

- The Catboost models has given high importance to features such as Temperature, Hour of the day, Winter Season and Solar Radiation.
- The Stacking Ensemble uses all three models and hence has high priority to features like Temperature, Hour of the Day, Winter Season and Functioning Day

Conclusions

- Upon Exploratory Data Analysis, we found that the bike rentals follow an hourly trend where it hits the first peak in the morning and the highest peak later in the evening.
- We also found that these trends are prominent only during weekdays and working days, leading us to make a safe assumption that office-goers make a notable contribution in bike sharing demand. In addition, seasons were observed to have a notable effect on bike rentals, seeing high traffic during the summers and a significant low during the winters.
- Upon training and evaluation of the machine learning models, the CatBoost model and the Stacked Ensemble of CatBoost, LightGBM and Random Forest models performed the best when evaluated using the R2 metrics. They produced R2 scores of 0.9369 and 0.9380, with a root mean squared error of 162.01 and 160.59 respectively.
- It was found that the top performing models made predictions based on the weather and time of the day as high weightage was given to seasons, temperature recorded and hour of the day. This confirms the trends observed during the exploratory data analysis stage of the project.

References

1. Christopher M. Bishop, “Pattern Recognition and Machine Learning”, Pg.137-139
2. Zhi-Hua Zhou, “Ensemble Methods Foundations and Algorithms”, Pg. 57-58
3. John T. Hancock and Taghi M. Khoshgoftaar, “CatBoost for big data: An Interdisciplinary Review”
4. Essam Al Daoud, “Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset”.
5. Jason Brownlee,“Machine Learning Mastery With Python, Understand Your Data, Create Accurate Models and Work Projects End-To-End

Thank You