Seoul Bike Sharing Demand Prediction

The dataset contains weather information such as temperature, humidity, windspeed, visibility, dewpoint, solar radiation, snowfall and rainfall. It also contains the number of bikes rented per hour and date information. The goal is to forecast bike rental demand of Bike sharing based on historical usage patterns in relation with weather, time and other data. Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes. People can rent a bike through membership or on demand basis.

We followed the following sequence of steps to solve

this problem statement:

1. Basic inspection was done on the raw data to check the number of columns, understanding distribution of data and checking statistics of the data in each variable. Missing values were checked and cleaned.
2. Feature engineering was performed by creating some new features, dropping unnecessary features and encoding the data into numeric form. The dependent variable was transformed to make it normally distributed.
3. Bi-variate analysis was done to check whether there’s any linear relationship between independent and the dependent variable. Correlation analysis was used to visualize the severity of multicollinearity. Multicollinearity was removed based on VIF factor.
4. The data was scaled and different algorithms were experimented with. Simple models like Linear Regressor and Decision Tree were tried first, then more complex algorithms like Tree ensemble were used to enhance accuracy.
5. Since there was not much linear relation between the independent and dependent variables, the linear regressor model did not perform well, so Tree based algorithms were used instead. Performance was drastically improved by using some boosting and ensemble algorithms and tuning the hyperparameters. The best performance was given by XGBoost model.
6. That’s great! It’s important to evaluate the models to overcome underfitting or overfitting and also have a rough i

We observed following results after completing the

task:

• Functioning day is the most influencing feature and temperature is at the second place for Linear Regressor.

• Temperature is the most important feature for DecisionTree, RandomForest and Gradient Boosting Regressor.

• Functioning day is the most important feature and Winter is the second most for XGBoost Regressor.

• RMSE Comparisons:

o LinearRegressor RMSE: 370.46

o DecissionTreeRegressor RMSE : 302.53

o RandomForestRegressor RMSE: 290.02

o XGBoostRegressor RMSE : 242.72

o GradientBoostingRegressor RMSE : 248.18

• The feature temperature is on the top list for all the regressors except XGBoost.

• XGBoost is acting different from all the regressors as it is considering whether it is winter or not. And is it a working day or not. Though winter is also a function of temperature only but it seems this trick of XGBoost is giving better results.

• XGBoostRegressor has the Least Root Mean Squared Error. So, it can be

considered as the best model for given problem.

XGBoost is giving the best accuracy of around 82%