Increasing Profits Through Demand Forecasting

Using Seoul Bike Sharing Demand Data

Benjamin Baugh

Introduction of Business Case

GOAL

- Our goal is to predict the hourly demand of bike rentals. With this information, we can make sure the correct number of bikes are available for customers to rent. We want to avoid stock outs to maximize revenue. We can also eliminate waste and the cost associated with having too many bikes.

PROCESS

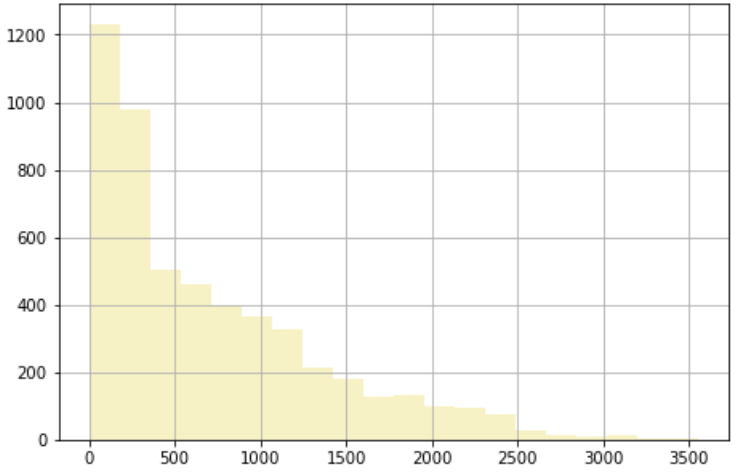
- We will clean and prepare the data and try many regression algorithms. After we find the model best suited to our task, we will fine-tune it. Finally, we will test the model with new data.

RESULTS

- The results indicate that after fine-tuning, the XGBoost model was the most accurate. Looking at an Actual vs Predicted plot, we see the model is more accurate when demand is between 0 and about 1300. When demand goes over 1500 then the model is less accurate. That being said, the hourly demand stays below 1500 most of the time.

Data Description

- There are 13 independent variables. Our dependent variable is Rented Bike Count. There are three columns that are not numeric and will need to be prepared before modeling. There are no missing values in any columns.
- The number of bikes rented has a Poisson distribution. Temperature & Humidity are close to a normal distribution. Wind speed, Visibility (10m), Solar Radiation, Rainfall and Snowfall have either a floor or a ceiling.
- Distribution of Bikes Rented



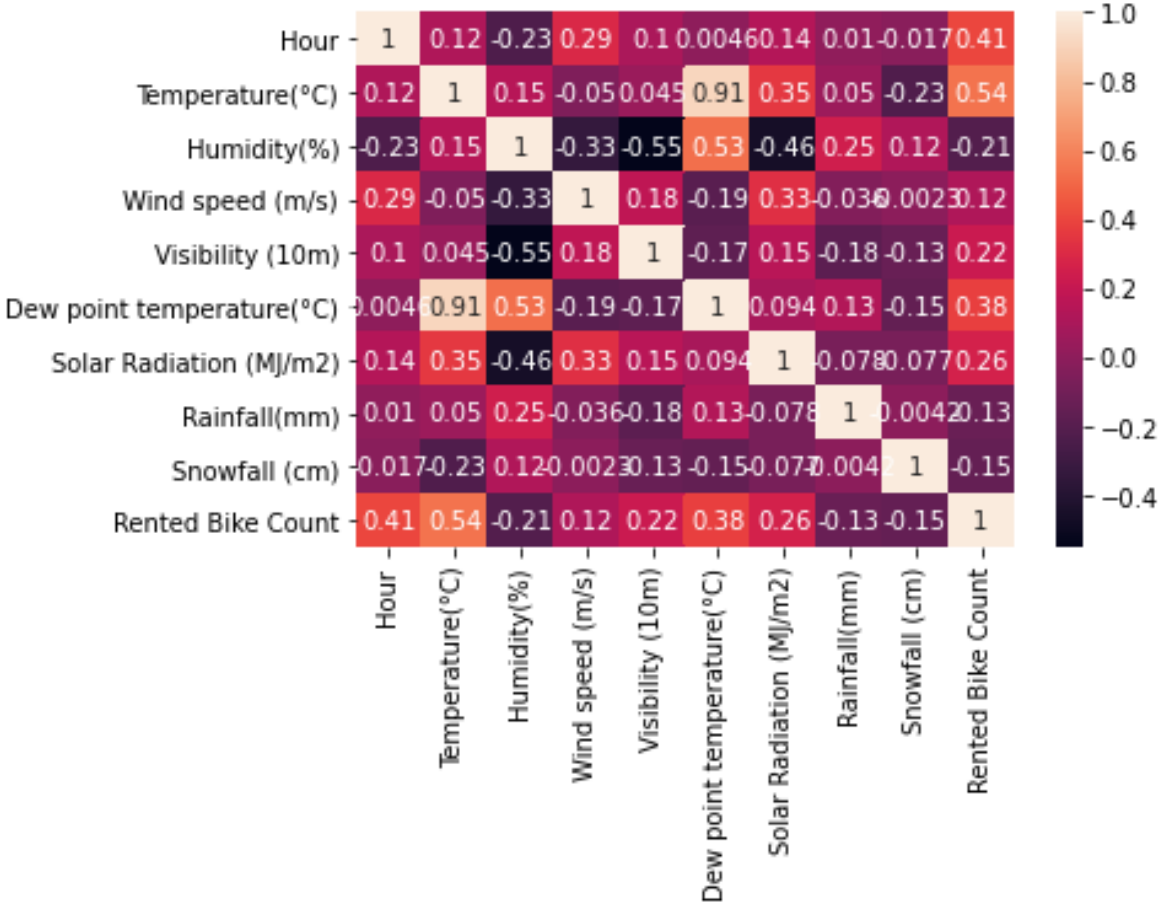
Variables

1. Date: year-month-day
2. Rented Bike count: Count of bikes rented at each hour
3. Hour: Hour of the day
4. Temperature: Temperature in Celsius
5. Humidity - %
6. Windspeed - m/s
7. Visibility - 10m
8. Dew point temperature: Celsius
9. Solar radiation - MJ/m2
10. Rainfall - mm
11. Snowfall - cm
12. Seasons: Winter, Spring, Summer, Autumn
13. Holiday: Holiday/No holiday
14. Functional Day: NoFunc(Non Functional Hours), Fun(Functional hours)

Correlation

- After creating a correlation matrix, we see “Temperature” and “Dew point temperature” are highly correlated.
- We want to avoid multicollinearity so we will drop one of the variables.
- We will keep “Temperature” because it has a higher correlation with Rented Bike Count.

Correlation Matrix



Cleaning and Preparing the Data

- Turned “Seasons”, “Holiday”, and “Functioning Day” into dummy variables.
- Broke out the date column into 5 variations of date:
  1. Year
  2. Month
  3. Day of Month
  4. Week Day
  5. Day of Week.
- Dropped variable “Dew point temperature”
- Scaled the variables with a min max scaler

Shortlist Promising Model

- Test some models to find which algorithm has the most promise with our data. Models tried:
  1. Linear Regression
  2. Decision Tree
  3. Random Forest
  4. Support Vector Machine (SVM)
  5. XGBoost
- The metrics used to compare models were:
  1. Mean Absolute Error (MAE)
  2. Mean Squared Error (MSE)
  3. Root Means Squared Error (RMSE)
  4. R Squared
- Put the most weight behind the RMSE metric

Results of Shortlisting

- Used Cross Validation with 10 folds. We took an average of the metrics for each model and a standard deviation of RMSE.
- Looking at the results the Random Forest and XGBoost models were the most accurate.

	Model	MAE	MSE	RMSE	RMSE StnDev	R2
1	Linear Regression	328	189,853	435	22.42	0.556
2	Decision Tree	147	66,307	257	19.73	0.843
3	Random Forest	107	33,663	183	17.50	0.921
4	SVM	412	356,117	596	29.47	0.169
5	XGBoost	103	27,737	166	13.69	0.935

Fine-Tune the Models

- Tuned both the XGBoost and Random Forest models.
- Used cross-validation and grid search to find the best set of hyperparameters.
- RMSE was used as the metric to re-fit for testing
- After running many versions of both models the XGBoost model was the most accurate
- The best hyperparameters were "max_depth=10,n_estimators=1000".

Model	MAE	MSE	RMSE	RMSE StnDev	R2
XGBoost	95	27,462	165	16.16	0.935

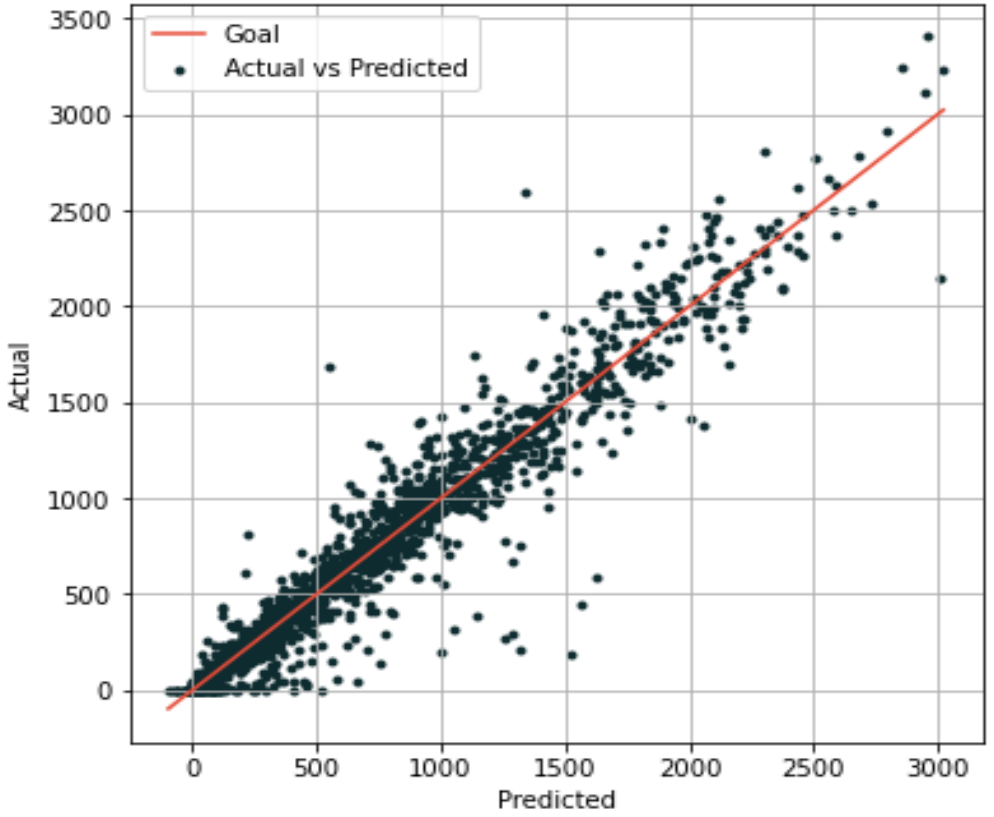
Conclusion

- Our goal at the onset of the projects was to predict the hourly demand of bike rentals.
- With an accurate prediction we could increase profitability.
- After finding the best model and fine-tuning it, we tested the model on a new set of data. At the beginning of our projects we split out a test data set that was 20% of the total data.
- The results were the best yet as measured by our selected metrics.

Model	MAE	MSE	RMSE	R2
XGBoost	96	25,987	161	0.933

Graph of Results

- This Actual vs Predicted graph shows the clustering around the red line which was our goal.
- We see the clustering around the line becoming looser when it passes approximately 1500.



Code

The Code For This Project Can Be Found At This Link