BOOMBIKES FORCASTING

By: Tushar Sood

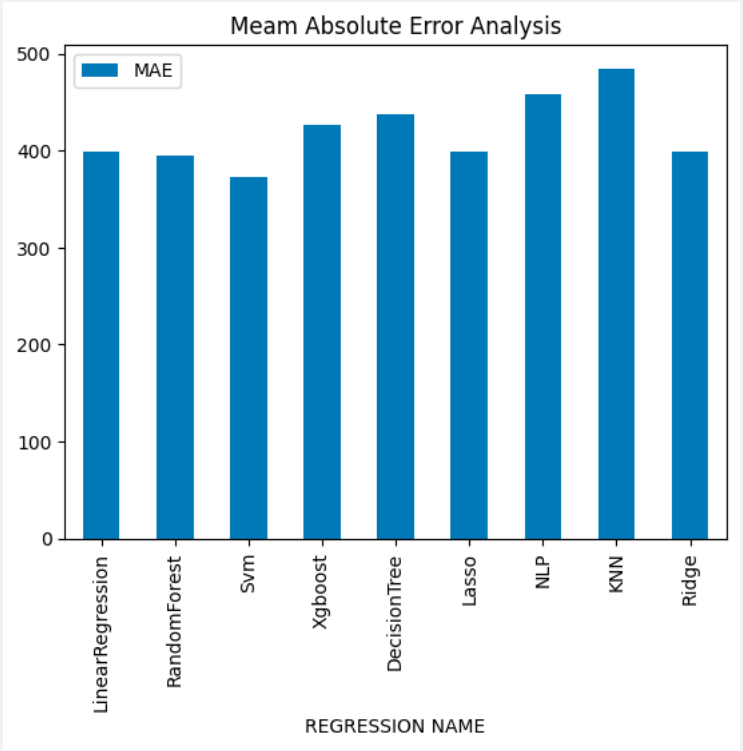
PROBLEM DEFINITION

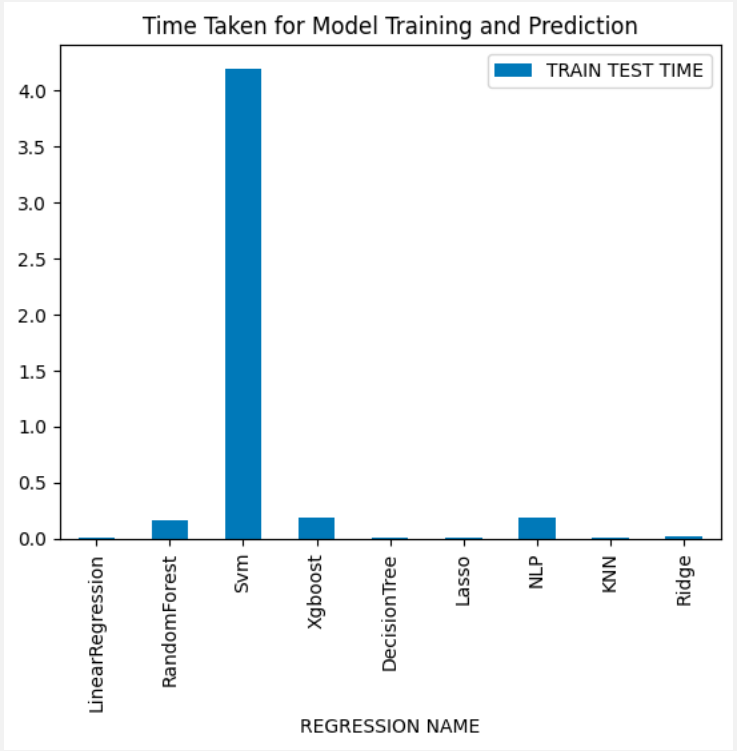
- Boombikes provide commercial commute bikes, however, there is a need to forecast demand to better plan resources. Static forecasting leads to losses and is not good for companies' financial health

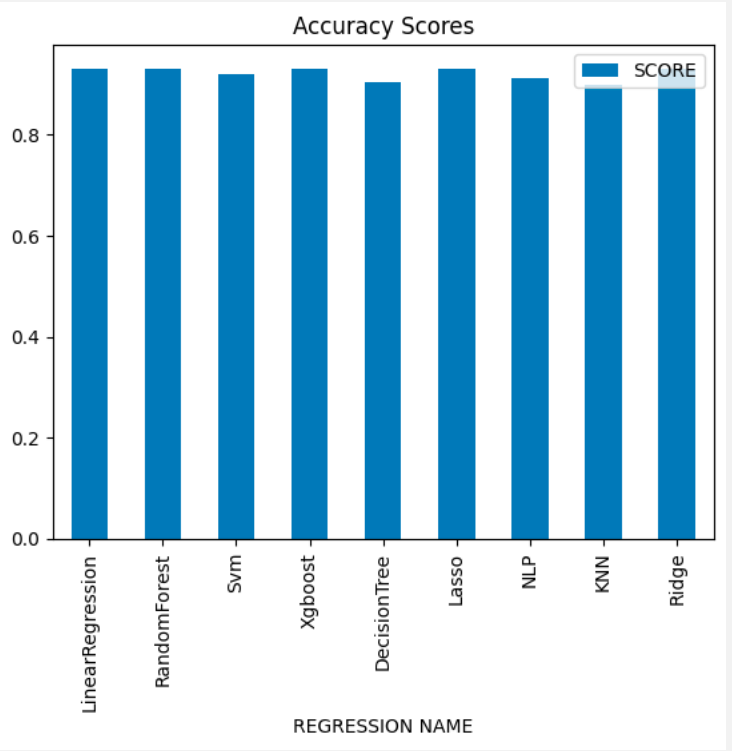
HISTORIC DATA AVAILABLE

- The given data is for 2 years 2018 and 2019
- Data does not contain any missing values
- The date is irrelevant and would like to drop this column as demand based on date is just coincident and does not provide any meaningful correlation with the demand forecast
- The variation between the working day and the non-working day is minimum hence can be ignored
- Dropping columns 'dteday', 'yr', 'holiday', 'weekday', 'workingday', 'casual' as they do not impact accuracy of prediction under any algorithm very well

COMPARISON OF REGRESSION ALGORITHMS #1







COMPARISON OF REGRESSION ALGORITHMS #2

REGRESSION NAME	MAE	TRAIN TEST TIME	SCORE
LinearRegression	398.268449	0.007097	0.931416
RandomForest	394.505785	0.161691	0.931290
Svm	373.274713	4.193126	0.920203

- The top Algorithms with the highest accuracy are Linear Regression, Random Forest and SVM. However, the best pick among these would be Linear Regression for this case.

FEW POINTS

Q: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable

A: From the experiments, I saw accuracy being reduced. The accuracy on average for 9 algorithms experimented dropped below 55%. Dropped 'yr', 'holiday', 'weekday', 'workingday' columns and observed accuracy above 90%

FEW POINTS

Q: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable

A: From the pairplot looks like temp and feeling temp have highest correlation

FEW POINTS

Q: How did you validate the assumptions of Linear Regression after building the model on the training set?

A: I calculated the accuracy scores and concluded Linear Regression is best suited. It had accuracy of 93%

FEW POINTS

Q: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A: temp, atemp and humidity

FEW POINTS

Q: Explain the linear regression algorithm in detail.

A: Linear Regression algorithm intends to find the least cost best fit 2D line among cluster of data points in 2D plane.

Any point on the given line is the optimal solution to find paired variable value. This means the RMSE score between predicted and actual value is minimum all the time.

FEW POINTS

Q: Explain the Anscombe's quartet in detail.

A:

FEW POINTS

Q: What is Pearson's R?

A:

FEW POINTS

Q: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
A:

FEW POINTS

Q: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A:

FEW POINTS

Q: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A:

THANK YOU!