Bike Demand Prediction

2021

Background

Situation

- Great interest in bike sharing systems due to their important role in traffic, environmental and health issues
- 500 bike sharing platforms around the world employing over 500 thousand bicycles

Question

- Bike sharing demand is highly dependent on environmental and seasonal settings and hence it's difficult to match supply vs demand
- People look at it as a cheaper alternative hence maintaining affordability is key which can achieved by increasing efficiency of use

Analysis

- Scope for research in terms of predicting number of people who might opt to use bike sharing platform on any given day
- Prediction can help the company better manage resources thereby increasing profitability and sustainability of the business model

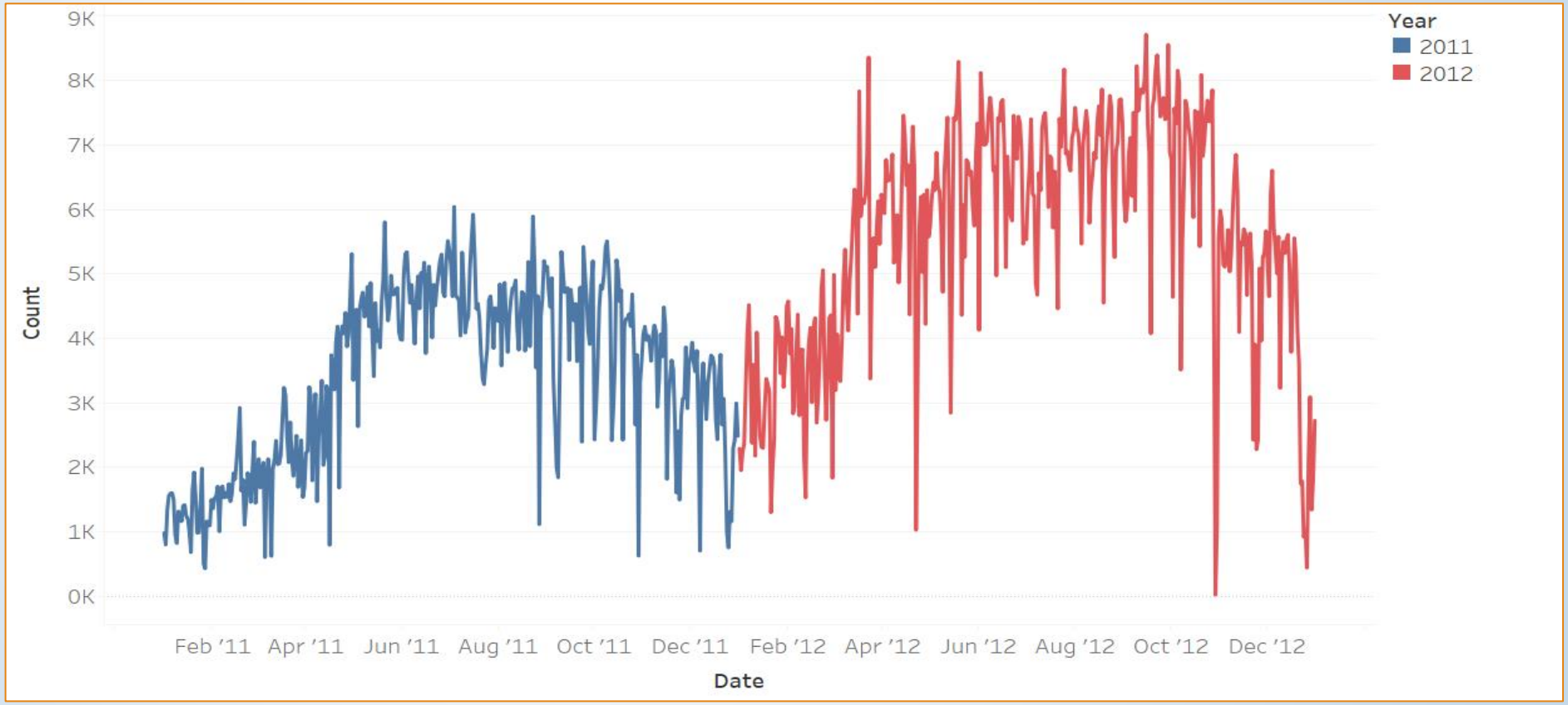
Data Exploration (1/3)

- The core data is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare System, Washington D.C.

Column	Description
Dteday	Date
season	Season (1:spring, 2:summer, 3:fall, 4:winter)
yr	Year
mnth	Month
holiday	Weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
weekday	Day of the week
workingday	If day is neither weekend nor holiday is 1, otherwise is 0
weathersit	1: Clear + Partly Cloudy, 2: Mist + Cloudy, 3: Light Snow/Rain, 4: Heavy Rain + Snow
temp	Normalized temperature in Celsius. The values are divided to 41 (max)
atemp	Normalized feeling temperature in Celsius. The values are divided to 50 (max)
hum	Normalized humidity. The values are divided to 100 (max)
windspeed	Normalized wind speed. The values are divided to 67 (max)
casual	Count of casual users
registered	Count of registered users
cnt	Count of total rental bikes including both casual and registered

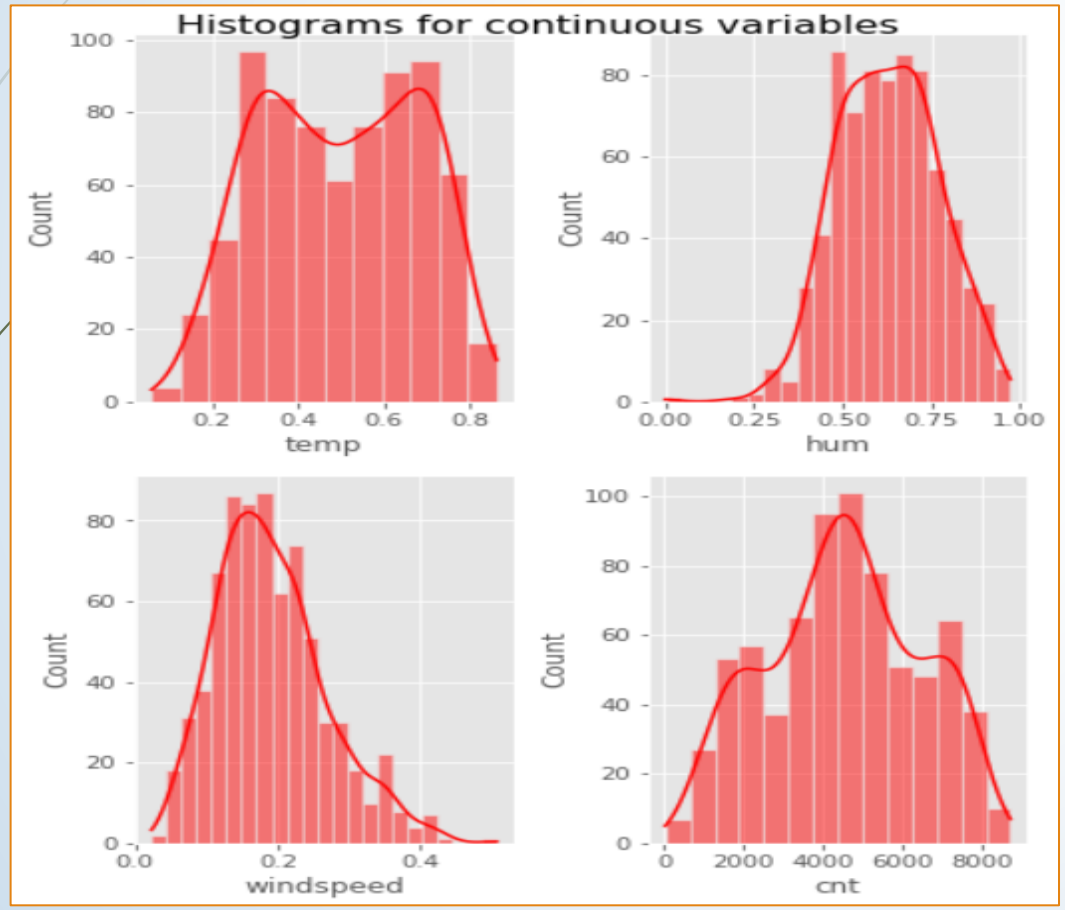
Data Exploration (2/3)

Bike-sharing Daily Count Time-series plot



Data Exploration (3/3)

Continuous Variables



Categorical Variables

season		holiday		workingday		wearthersit	
Category	Count	Category	Count	Category	Count	Category	Count
1	181	0	710	0	231	1	463
2	184	1	21	1	500	2	247
3	188					3	21
4	178						

Approach

PREDICTION MODELS (Supervised Learning Algorithms)

Linear Regression	CART	Random Forest	Neural networks	Forecasting models
• Subset variable selection • Lasso	• Regression Decision Tree	• Prediction with random forest technique	• Timeseries forecasting with RNN	• Prophet package (additive model)
PERFORMANCE RMSE : In-sample RMSE : Out-of-sample	vs	PERFORMANCE RMSE : In-sample RMSE : Out-of-sample	vs	PERFORMANCE RMSE : In-sample RMSE : Out-of-sample	vs	PERFORMANCE RMSE : In-sample Correctness	vs	PERFORMANCE RMSE : In-sample

Linear Regression

Variable subset selection

- Regressed multiple combinations of predictors against the response
- Result: Model with all variables had the highest R-squared and lowest RSS

Stepwise selection using AIC

- Stepwise variables selection technique using AIC suggested model with all variables as the best model with lowest AIC of 8386.5

LASSO variable selection

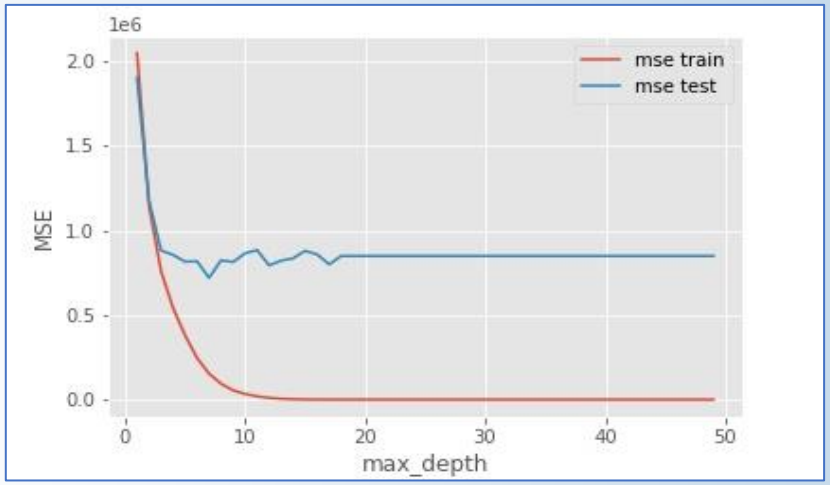
- LASSO variable selection technique suggested use of four variables – instant, season, mnth and weathersit

Linear Regression

- Multiple linear regression with all variables
- All predictors except holiday and weekday were statistically significant
- R-squared: 84.5%
- In-sample RMSE: 759.9
- Out-of-sample RMSE: 752.6

CART (Regression Decision Tree)

- Decision trees are adept at handling non-linear relationships between predictors and response variable and continuous nature of response variable warranted the use of regression decision tree
- 'Depth' of the tree was determined using the MSE vs depth plot which suggested a depth of 6 levels
- On comparison with Linear Regression the Regression Decision Tree provided significant improvement in in-sample performance only



Regression Decision Tree	
In sample RMSE	495.8
Out of sample RMSE	898.5

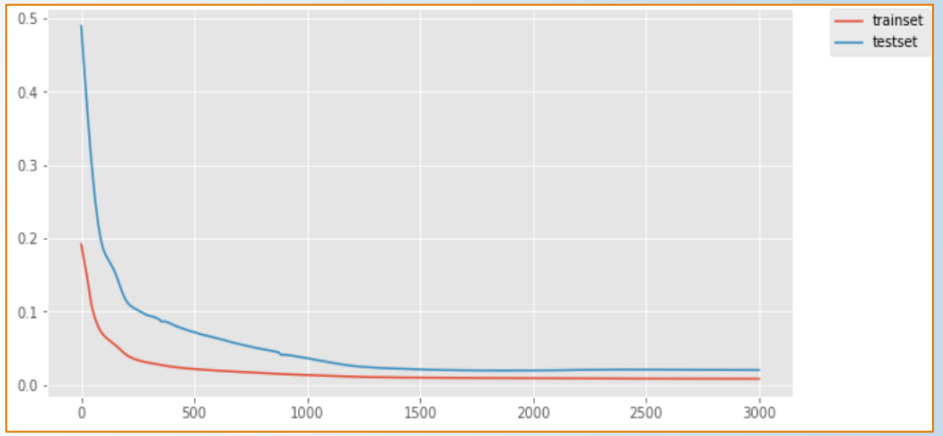
Random Forest

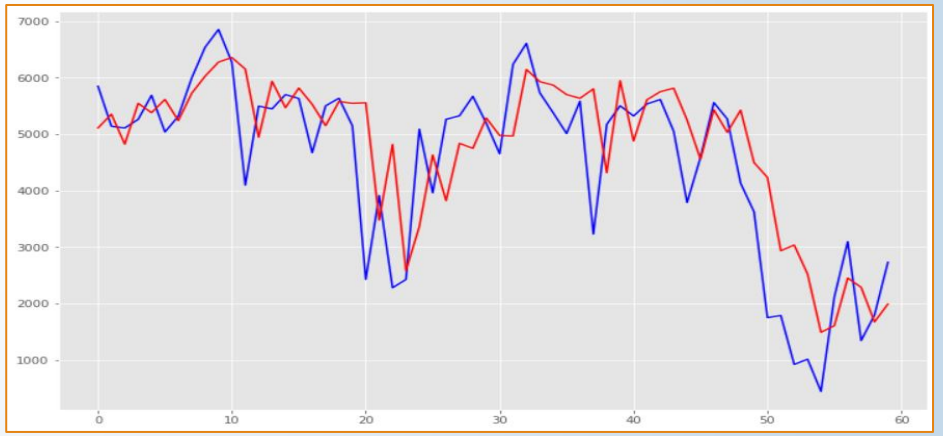
- The idea of random forests is to randomly select m out of p predictors as candidate variables for each split in each tree. By default, m=p/3 for regression tree, and m=√p for classification problem. The reason of doing this is that it can decorrelates the trees such that it reduces variance when we aggregate the trees
- Random Forest algorithm provided no improvement at all compared to Linear Regression or Decision Tree models and hence will be discarded

Random Forest	
In sample RMSE	2670
Out of sample RMSE	2671

Neural Networks (RNN)

- Neural networks are a class of machine learning algorithms used to model complex patterns in datasets using multiple hidden layers and non-linear activation functions
- For modeling, we will be leveraging the Simple RNN (Recurrent Neural Network) algorithm from Tensorflow – Keras package
- Since the data at hand has a time-series element to it, RNN better handles the temporal dynamic behavior
- Plotted in-sample and out-of-sample performance of the model over 3000 epochs (iterations)
- Model confuses noise with signal hence non of the machine learning algorithms are able to predict time series data efficiently

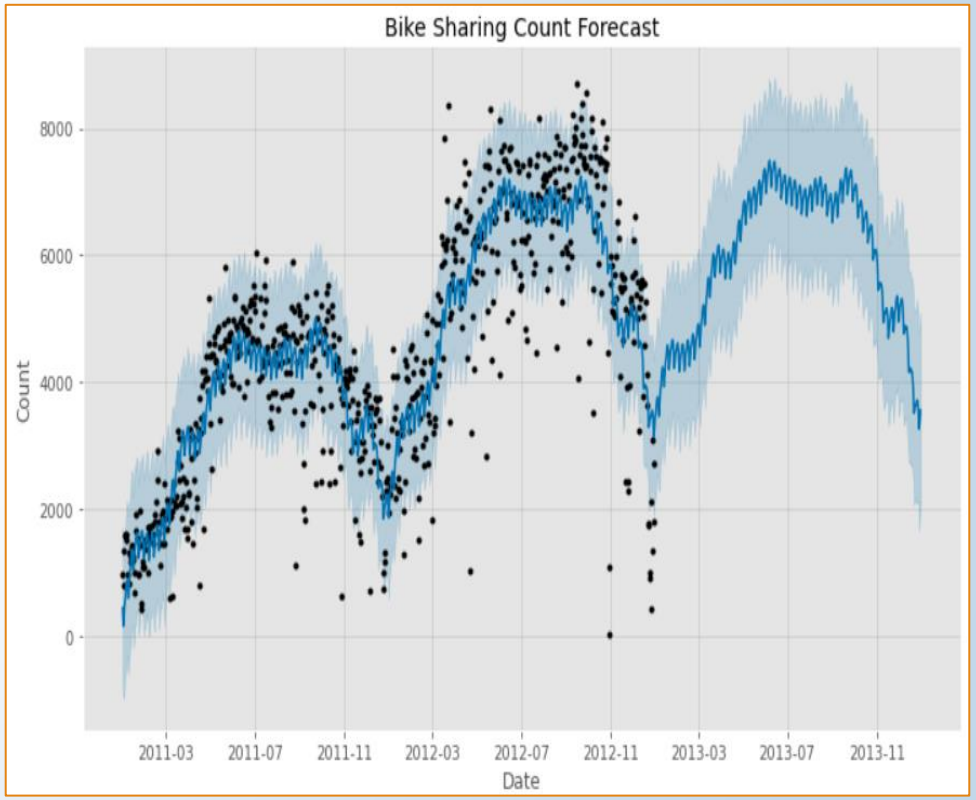




Neural Network - RNN	
Mean	4506.51
MAE	746.25
MAE/Mean Ratio	16.56%
Correctness	83.44%
RMSE	534.29

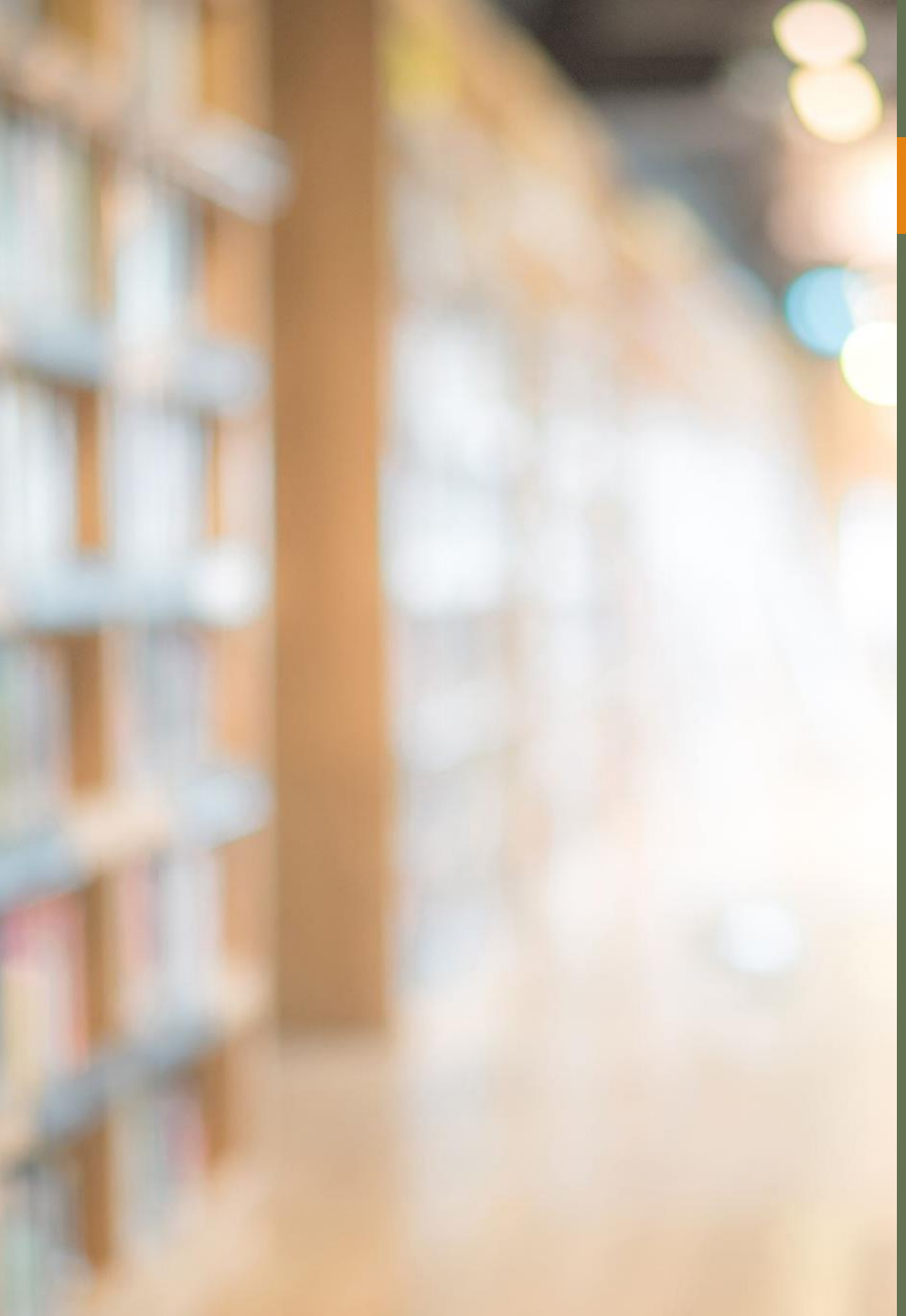
Time series forecasting using Prophet package

- Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects
- In the model we set the change point parameter to 0.15. This hyperparameter is used to control how sensitive the trend is to changes thereby taking care of bias-variance tradeoff
- The RMSE obtained for initial 2 years data is 956.41. The RMSE is high because the actual values have very high variance resulting in increased RMSE. However, we can see the forecast for next year maintains the similar trend and shape as seen for historical data points



Conclusion

- Most of the machine learning algorithms confuse signal with noise on multiple occasions for a time series data and hence fail to efficiently predict the outcome
- For prediction based on the features RNN can be used to predict the count of bikes
- Forecasting techniques use detrending and transforming data to make it stationary thereby making it possible to effectively forecast the data
- For prediction of count of bikes on a given day based on historical counts, time series forecasting using Prophet algorithm from Facebook provides the best results



References

- Data Source - http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#
- Prophet Model - https://towardsdatascience.com/time-series-analysis-in-python-an-introduction-70d5a5b1d52a