Demand Forecasting for Bike Rentals

Capstone Project

Sonal Kiran Hansra

Table of Contents

Introduction

According to a paper published in the journal - 'Sustainable Cities and Society', which looked at the life cycle of a public self-service bike-sharing system in the city of Edinburgh, UK, modeling the production, operation, and disposal elements of the system, bike-sharing saves between

716 and 4,300 tonnes of CO_2e per year

At the level of the functional unit, a cycle releases **9.6 g** CO_2e/km, which is

50 times less

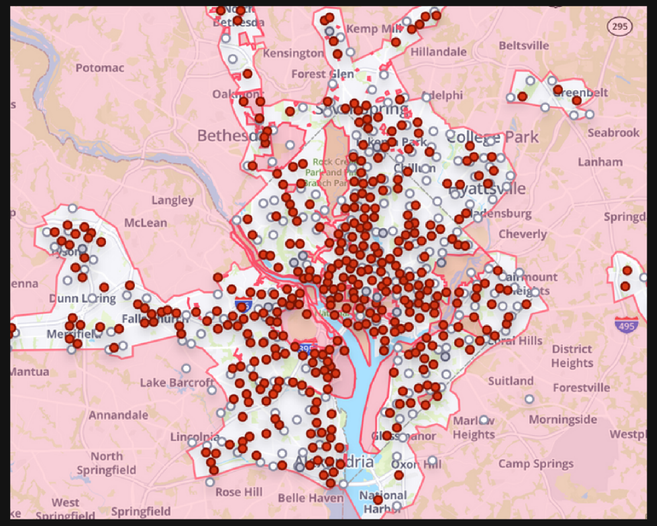
than an average car with petrol which releases 259 g CO_2e/km

The Bike Sharing market is expected to grow at a CAGR of 11% over the next decade and reach

$18.4 Billion

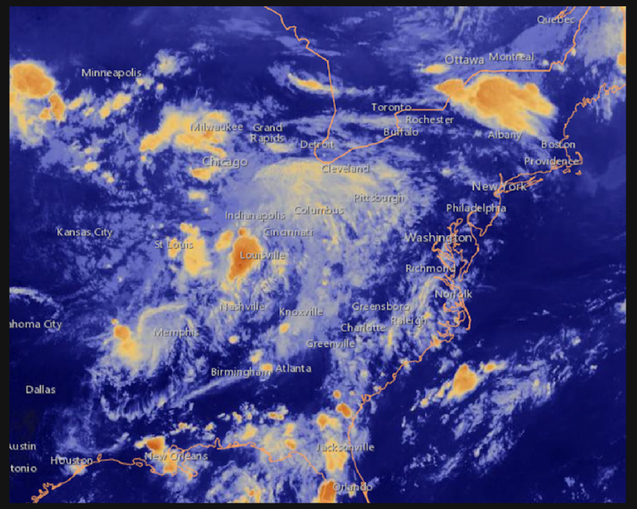
The objective of this project is to contribute to this sector by forecasting bike rental demand.

Meet the Data



Capital Bikeshare

Bike sharing data for Washington, DC



NOAA Weather

Weather information such as temperature, rain, snow, hail

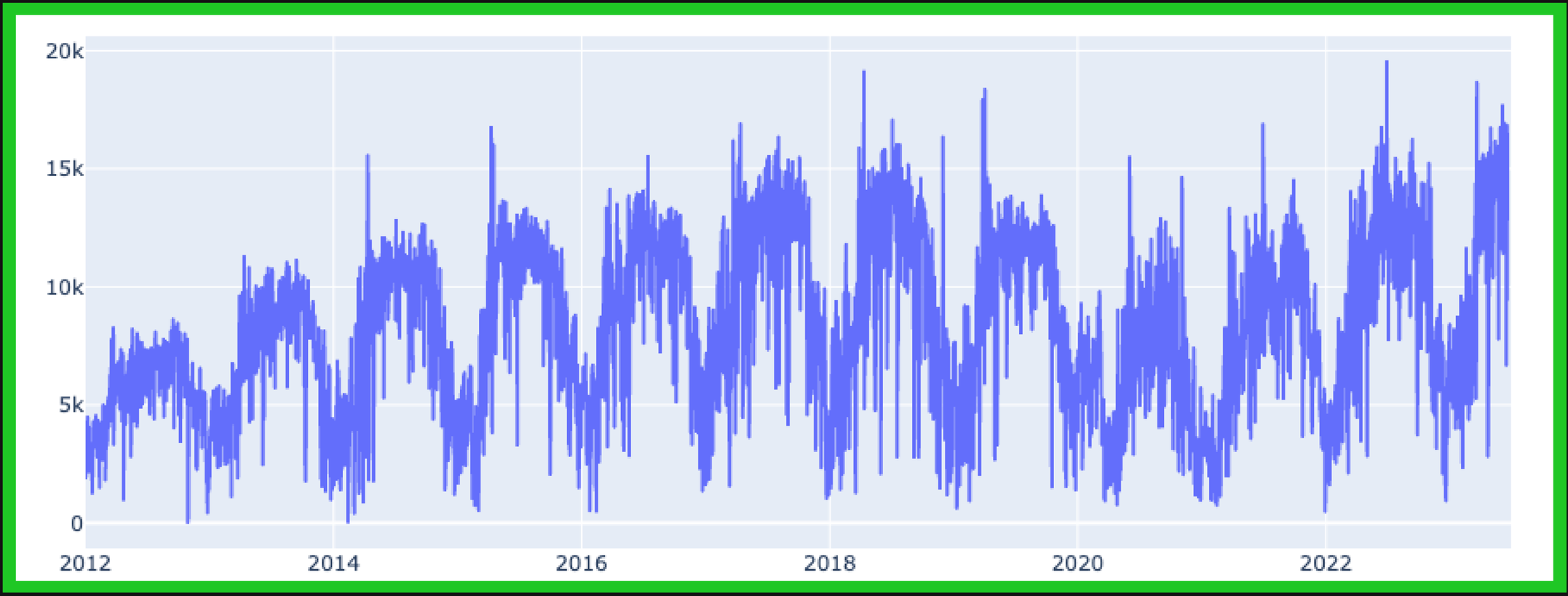
DC Holidays

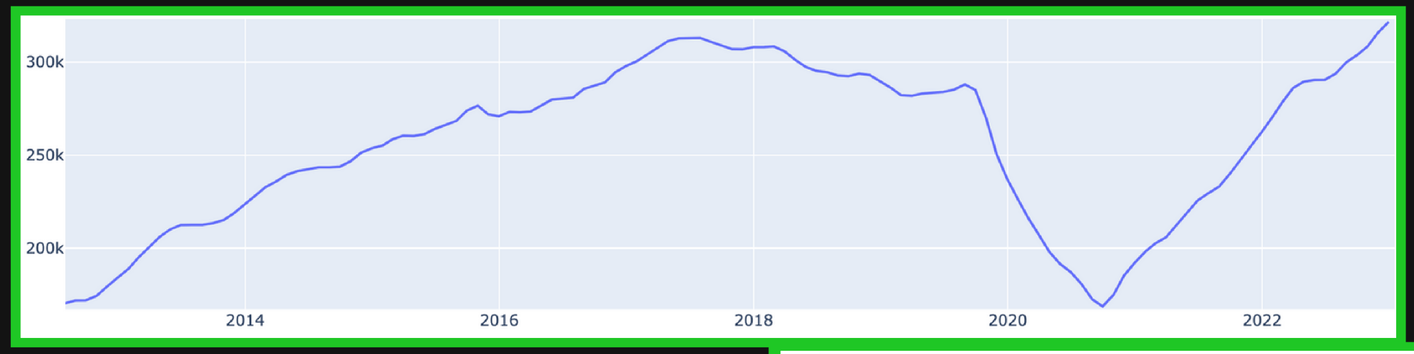
Holidays, weekend/weekday information

EDA

- Addressed missing time series and weather data
- Performed feature engineering - dropped/ merged/ created columns
- Performed time series decomposition to explore trend and seasonality

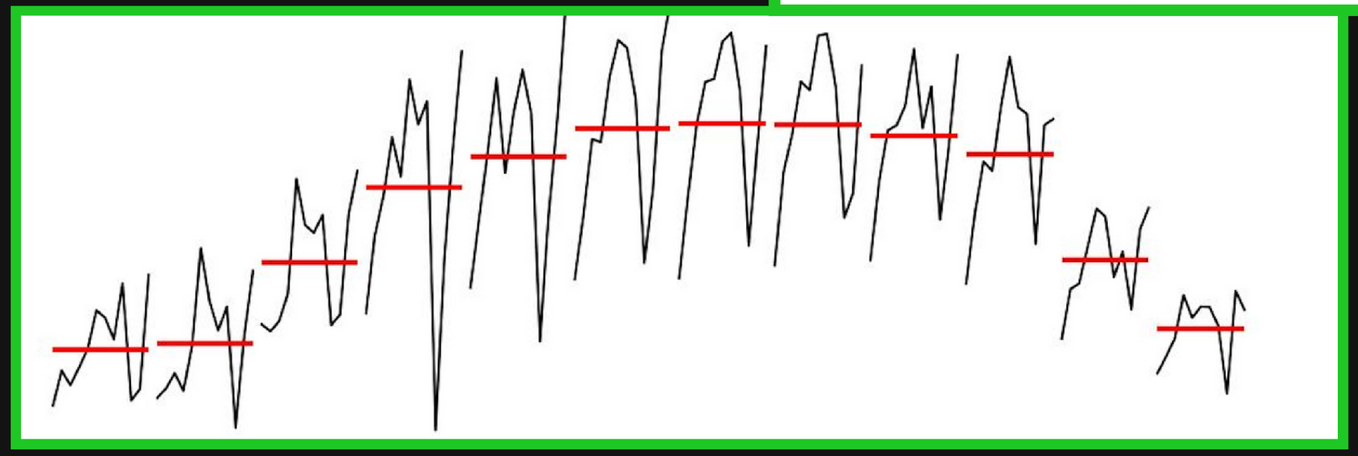
Daily Demand Over Time





Trend

Weekly Seasonality



Monthly Seasonality

COVID!

One of the biggest challenges with the data was the effect of the pandemic on bike rental demand.

After carefully evaluating multiple approaches (on the right), I decided to -

Drop the data for the pandemic years (2020 & 2021)

Use the data as it is

A limitation of ARIMA models is that they are incapable of capturing non-linear patterns hidden within a time series

Forecast & update demand for COVID years

This is time-intensive and wasn't feasible within given deadlines

Add a feature for the pandemic

May work for the tree-based ensemble model, but won't work with (S)ARIMA

Use past data

COVID impacted almost 2 years, which is a considerable amount of time given the span of our data, using past data would've biased the models

Baseline Modeling

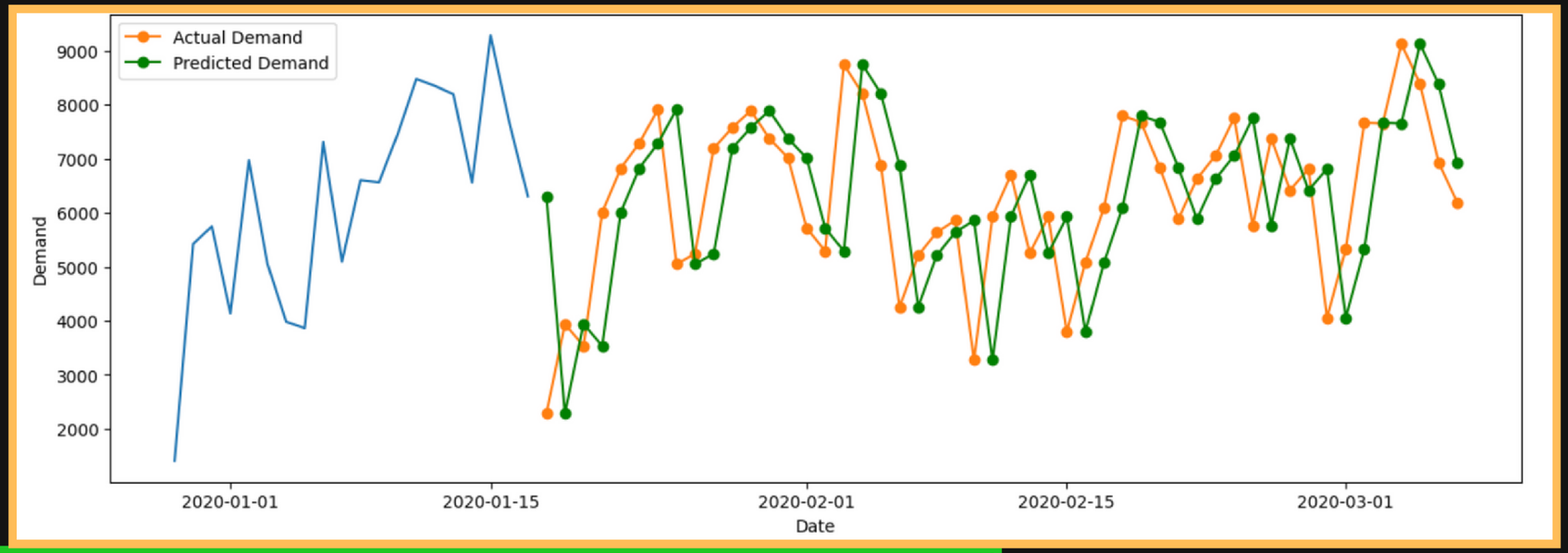
Naïve Forecasting

Since we intend to provide short-term forecasts, our first baseline model assumes that the predicted value at time t is equal to the actual value of demand at time t-1.

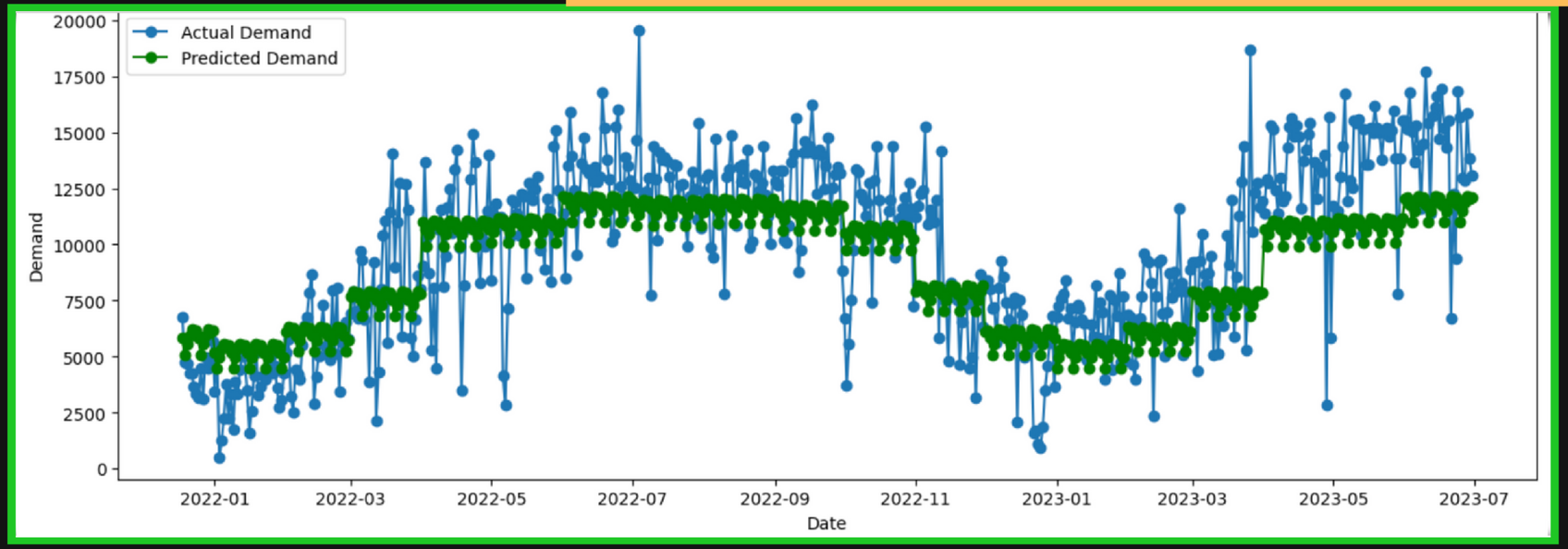
Seasonal Decomposition

In our second baseline model, we extract the trend and seasonalities from our training data and use them to forecast demand.

Demand vs. Forecast



Naïve Forecast



Seasonal Decomposition

Modeling

XBGoost

XGBoost stands for eXtreme Gradient Boosting. It is a tree-based ensemble method that has witnessed a lot of success in recent years. XGBoost uses advanced regularization compared to gradient boosting and trains faster as well

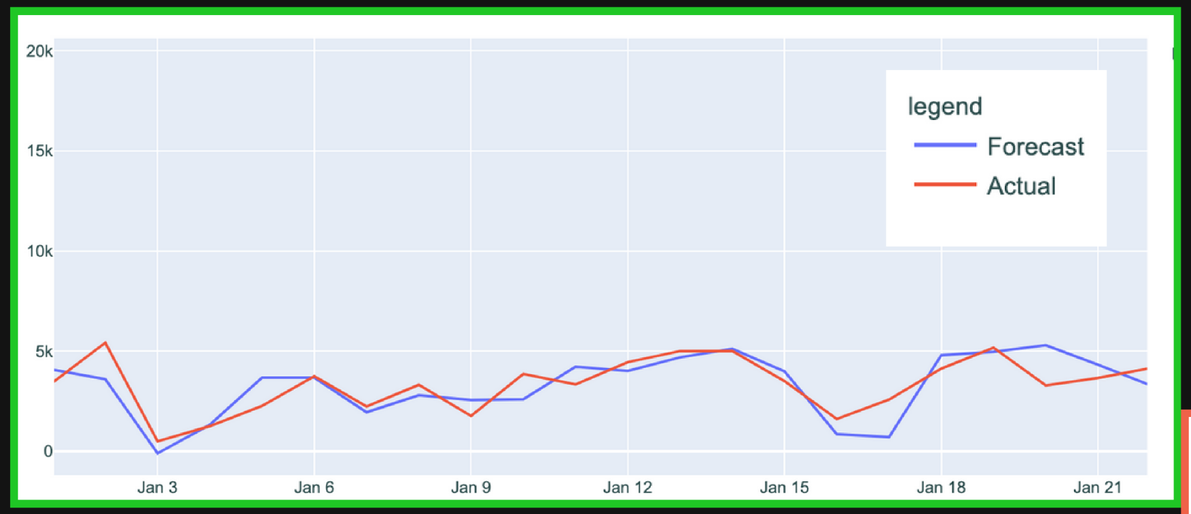
ARIMA

ARIMA stands for Autoregressive Integrated Moving Average. It is a statistical model that forecasts based on only the historical data. The `integrated` part here corresponds to the initial differencing step that needs to be applied to the data to make it stationary.

SARIMA

SARIMA stands for `Seasonal AutoRegressive Integrated Moving Average`. It is an extension of ARIMA that accounts for seasonal patterns in the data. The 'X' in SARIMAX implies that we can provide exogenous variables to this model as well.

XGBoost



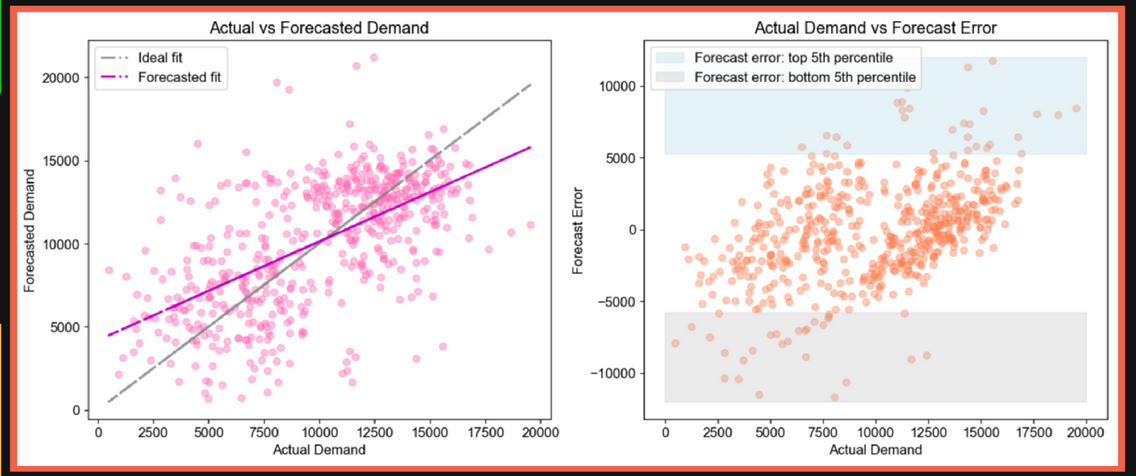
ARIMA

SARIMA

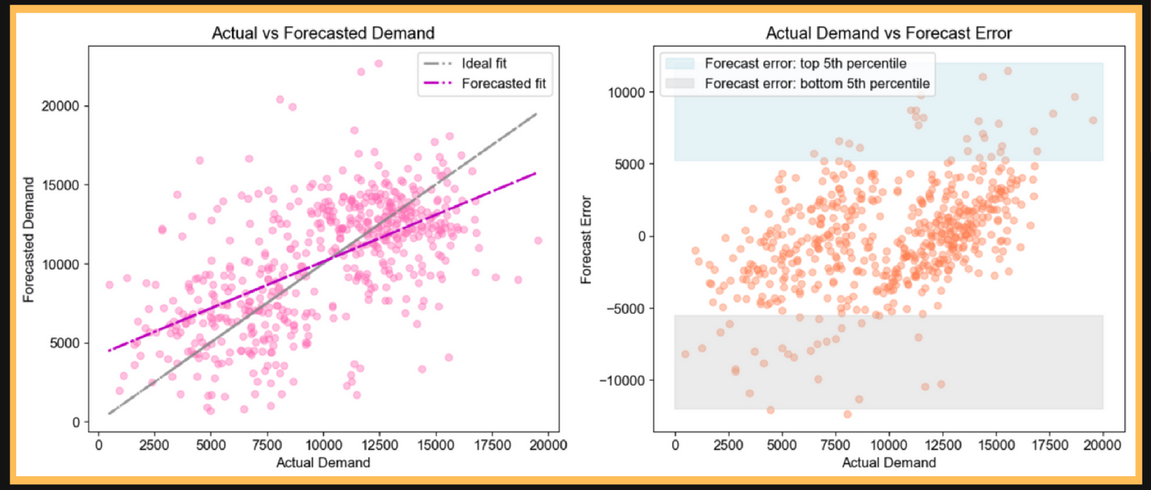
XGBoost

This model does relatively poorly when it comes to predicting demand for the weekends.

The proportion of weekends in cases of extreme (top and bottom 5 percentile) forecast errors was found to be 1.7 times of expected!



ARIMA



SARIMA

Results

The best model amongst the ones tried so far is XGBoost, by a huge margin.

Naïve model

MAPE = 0.278 %
RMSE = 2710

Seasonal Decomposition

MAPE = 0.304 %
RMSE = 2672

XGBoost

MAPE = 0.178 %
RMSE = 1970

ARIMA

MAPE = 0.38 %
RMSE = 3433

SARIMA

MAPE = 0.4 %
RMSE = 3451

Next Steps

There is a lot of repetitive code in this notebook, create classes/functions where possible and clean this notebook up.

Investigate the outliers in bike rental demand data, and check if the spikes in the data correspond to promotional events or any other such critical information that may help inform data preparation.

Investigate if seasonal decomposition can be used effectively to remove multiple seasonalities (consider using statsmodels-MSTL) from the data before employing (S)ARIMA.

Explore RNNs like LSTM to forecast demand and compare with XGBoost.

Thank You!

Questions?