Thomas Tran
DATA-3421 Data Mining
Professor Roustami

DATA-3421 Final Project Store Sales Time-Series Forecasting

The data can be accessed on the Kaggle website at this link:
https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data

Introduction:

This dataset originates from stores located in Ecuador and uploaded to the Kaggle website. The goal of the challenge is to use time-series forecasting to predict future stores sales on data from "Corporación Favorita".

The data contains information from 54 different store locations, 33 different types of products, which product being promoted on a given day and includes various different possible explanatory variables such as oil price, holiday, etc.

The data was compiled ranges over the course of nearly five years from January 1, 2013 to August 15, 2017.

Data Description:

train.csv - The training data, comprising time series of features **store_nbr**, **family**, and **onpromotion** as well as the target **sales**.
stores.csv - Stores metadata including city, state, type and cluster
oil.csv - Contains daily price of oil. This may be important since Ecuador's economy is largely based around the price of oil.
holidays_events.csv - Contains dates of regional, and local holidays

Preprocessing/Cleaning:

Taking an initial look at the data, the data was already very clean and ready to go as there was almost no missing data. Most of the preprocessing involved just merging some of the datasets together as well as extrapolating some of the missing data in the oil price dataset.

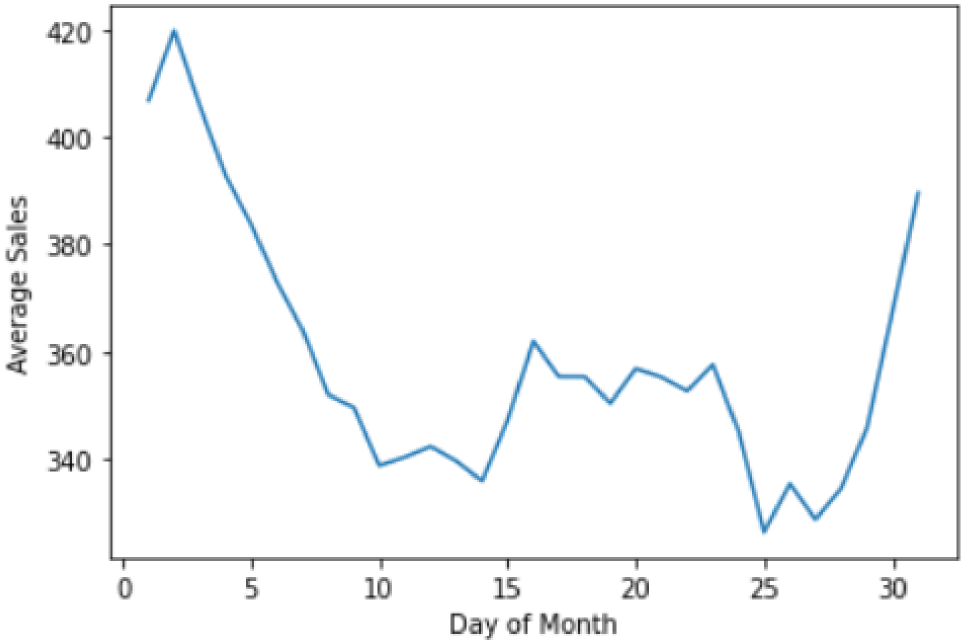
Feature Engineering:

The original training set only contained 4 columns, one being the target variable. Therefore, I thought it would be valuable to add some features to train the model. Some of the features that were created include: the sales of the product over the past 4 days, the rolling average of the past 4 days, the year, day of the month, month of the year and price of oil.

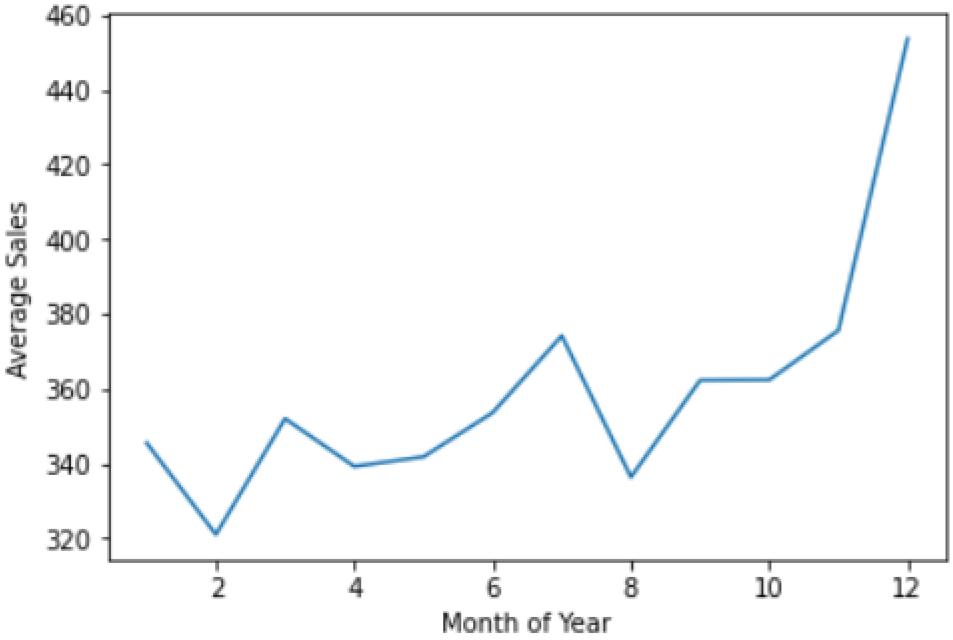
Exploratory Data Analysis:

Here are just some of the plots done after preprocessing:

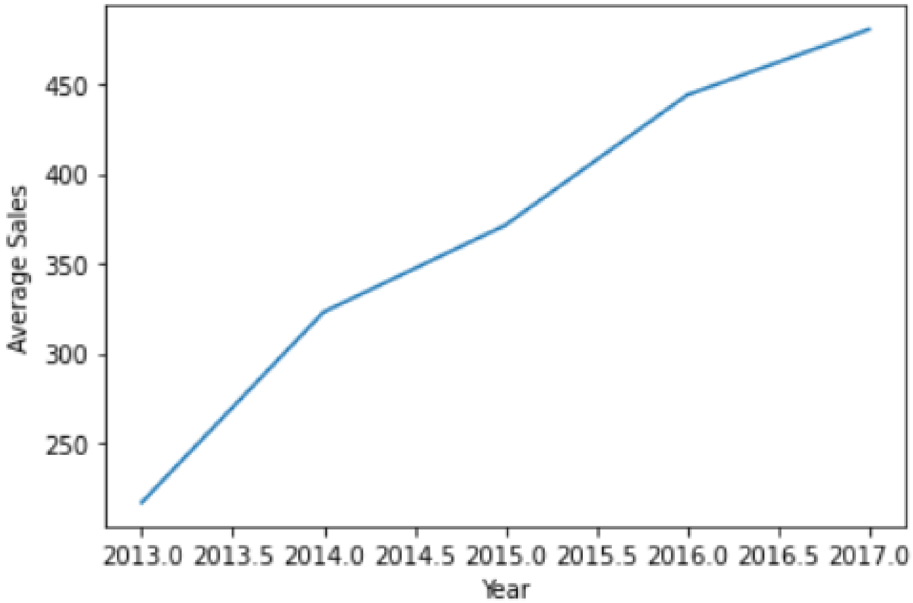
Average Sales by day of month



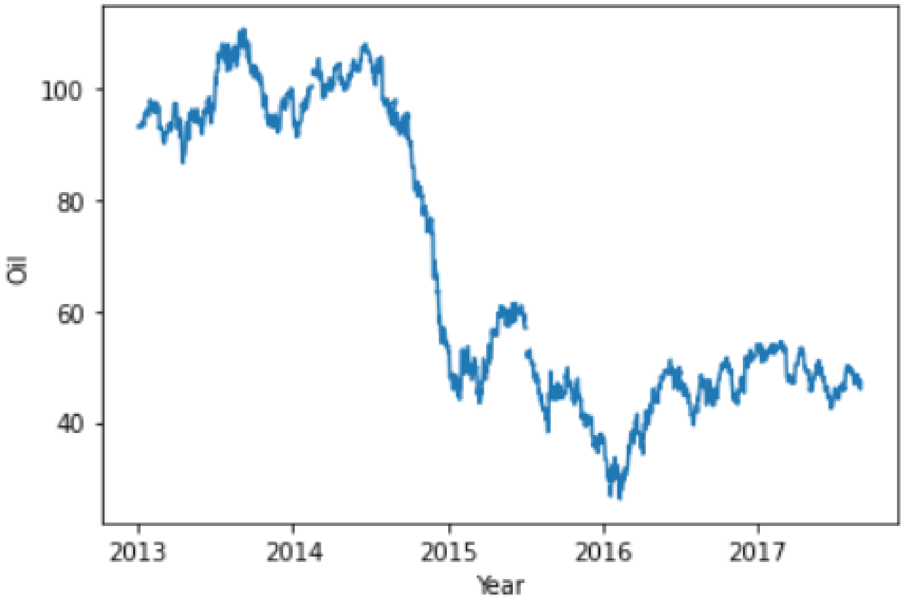
Average Sales by Month in Year



Average Sales over Years



Oil Price Over the Four Years



Some more data visualization can be found in the jupyter notebook under the EDA section.

(There were also some visualizations that I performed and eventually removed because they were difficult to interpret and didn't provide much to understanding the data.)

Training:

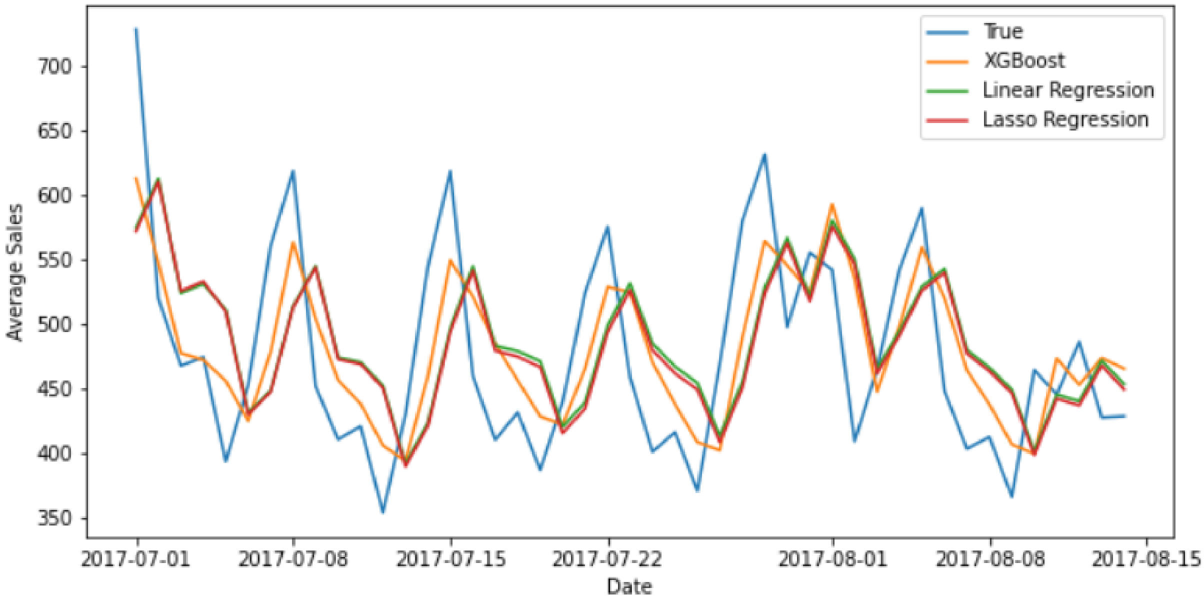
I used 3 different models for this Time-Series forecasting task:

1. XGBoost
2. Linear Regression
3. Lasso Regression

Table containing results of different models on the validation set

Model	r-squared	RMSE	Time Taken
XGBoost	0.93922788	321.21	1 minute
Linear Regression	0.9137132757	382.74	992 ms
Lasso Regression	0.91369741	382.78	3 minutes 5 seconds

Plot of Average Sales over validation time period using the different models



Conclusion:

Linear Regression is the fastest model with a training time of only 992 ms and a Root Mean Squared Error of 382.74.

Lasso Regression appears to be the worst with a training time of 3 minutes and 5 seconds and a Root Mean Squared Error of 382.78

XGBoosting appeared to result in the best results with a Root Mean Squared Error of 321.21 but with a runtime of about 1 minute.

Therefore, in terms of accuracy XGBoosting appears to be the optimal choice. However, if you would rather sacrifice a little bit of accuracy for run time the better choice is a linear regression model.

Citations

https://www.kaggle.com/competitions/store-sales-time-series-forecasting/overview

https://www.turing.com/kb/an-introduction-to-naive-bayes-algorithm-for-beginners#advantages-of-a-naive-bayes-classifier

https://www.youtube.com/watch?v=RRd2wzMRpOc&t=2265s

https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article#:~:text=cost%20to%20accuracy!-,What%20is%20XGBoost%20Algorithm%3F,optimize%20their%20machine%2Dlearning%20models.