Store Sales - Time Series Forecasting

Thomas Tran
DATA-3421 Final Project

CORPORACIÓN
FAVORITA

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

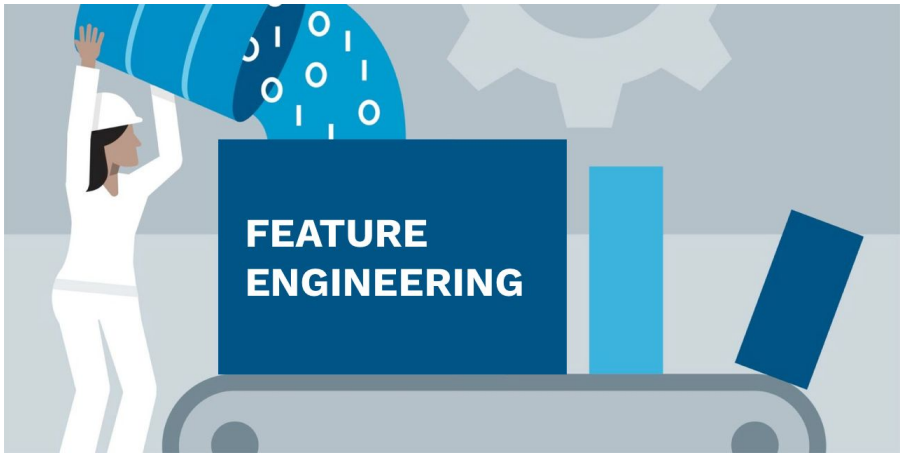
oil.csv - Contains daily price of oil. This may be important since Ecuador's economy is largely based around the price of oil.

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



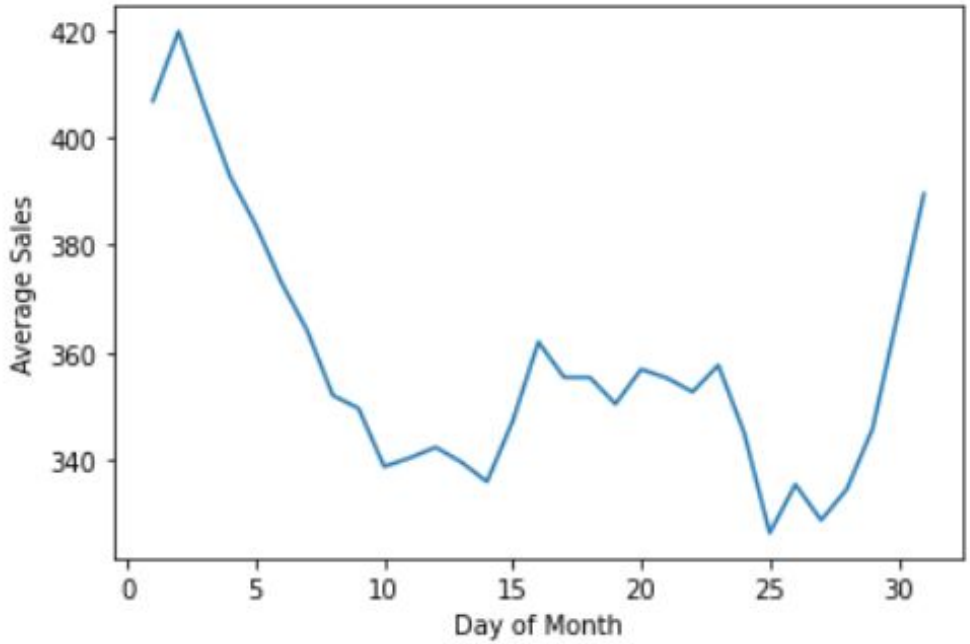
The original training set contained only 4 columns therefore, I felt it was necessary to do some feature engineering. Some features I decided to include was the sales of the past four days, their average as well as some other features.

This increased the number of features from just 3 before to about 10 after.

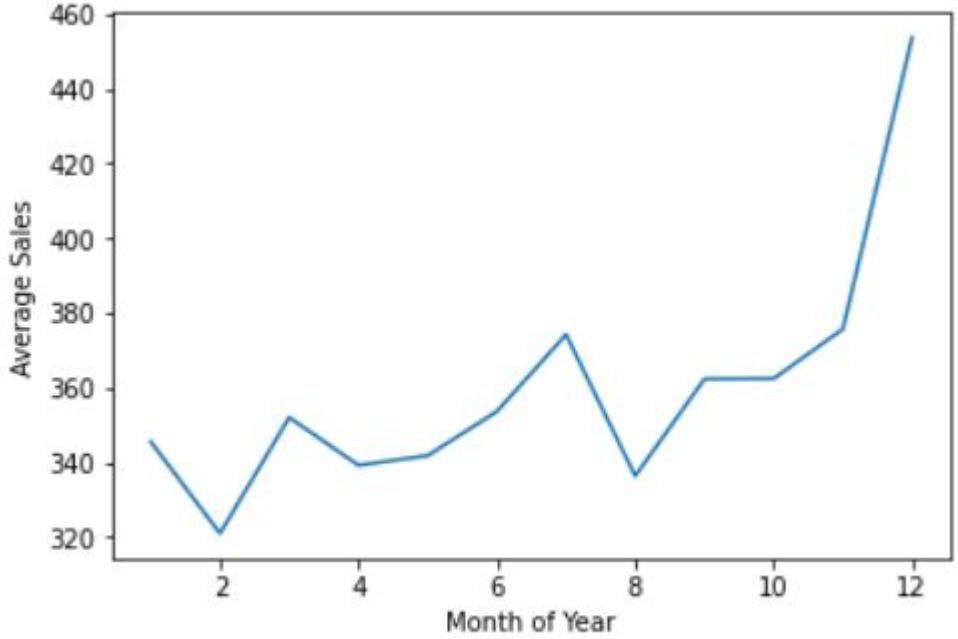
Exploratory Data Analysis

After the preprocessing/feature engineering step I wanted to take an initial look at the data and how some of the different variables interacted with one another.

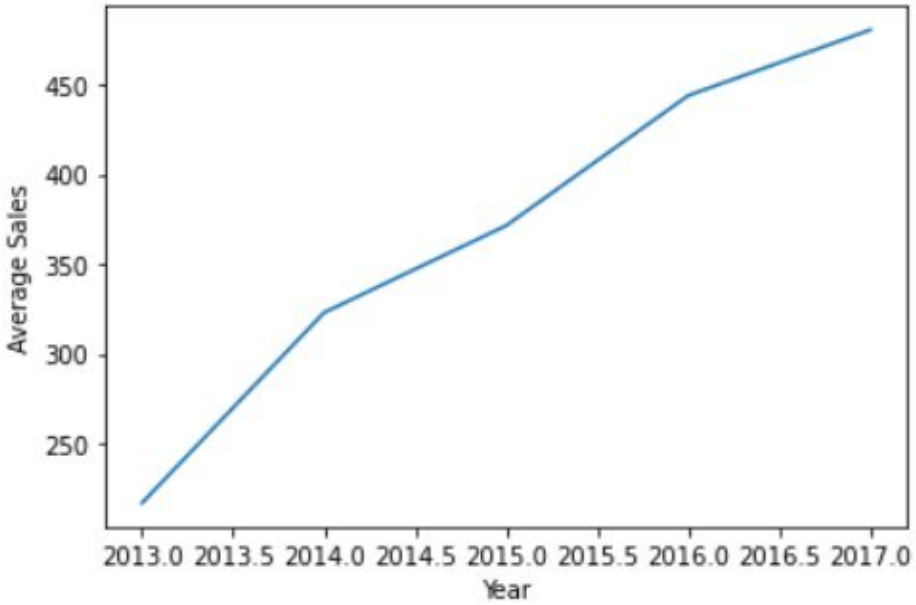
Average Sales by Day of Month



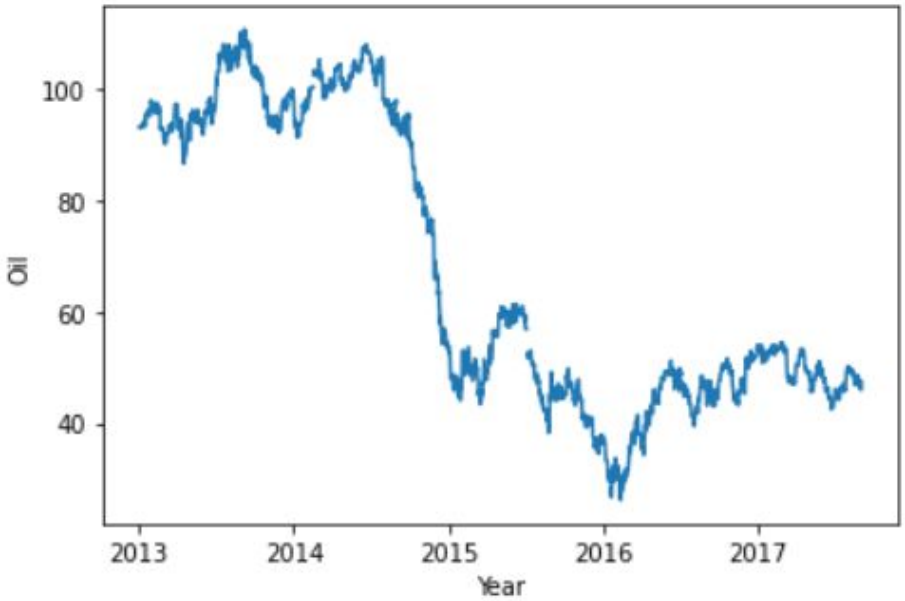
Average Sales by Month in Year



Average Sales by Year



Oil Price Over Course of Four Years



Training

The three models I decided to choose for this regression task are:

1. XGBoost
2. Linear Regression
3. Lasso Regression

Performance Evaluation

XGBoost:

r-squared: 0.93922788

RMSE: 321.21

1 min

Linear Regression:

r-squared: 0.9137132757

RMSE: 382.74

992 ms

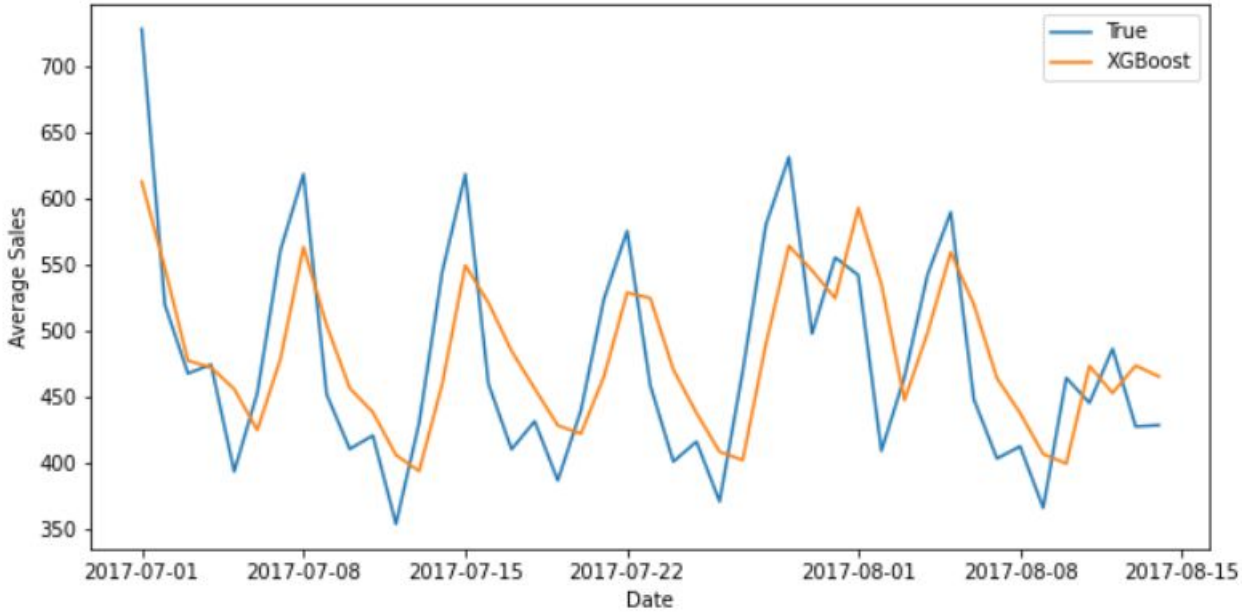
Lasso Regression:

r-squared: 0.91369741

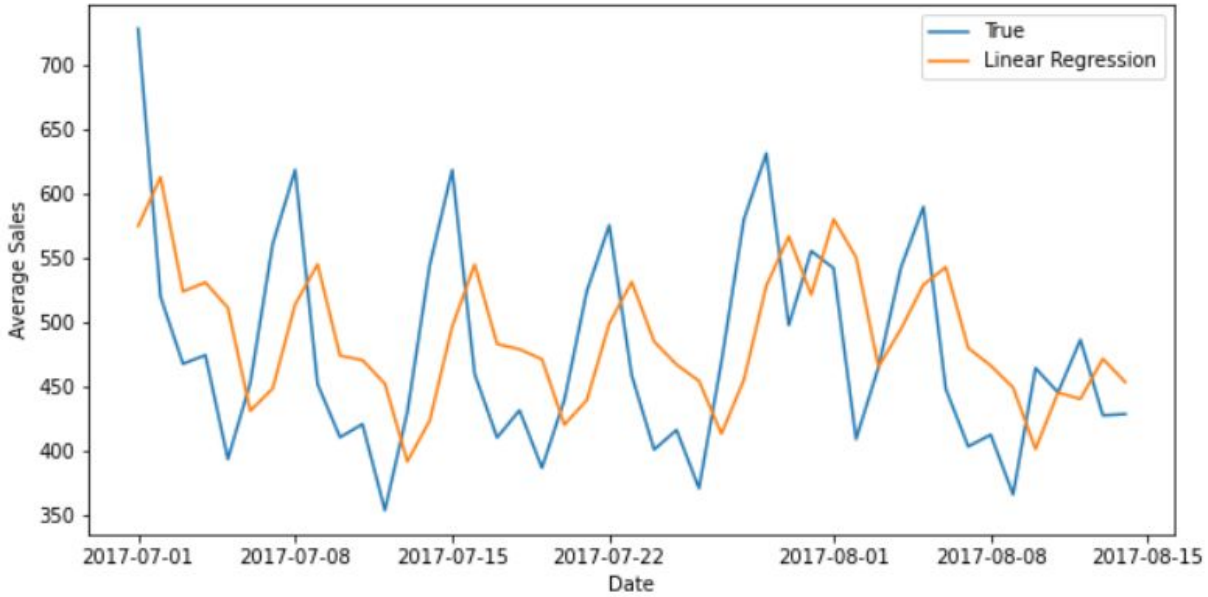
RMSE: 382.78

3 min 5 sec

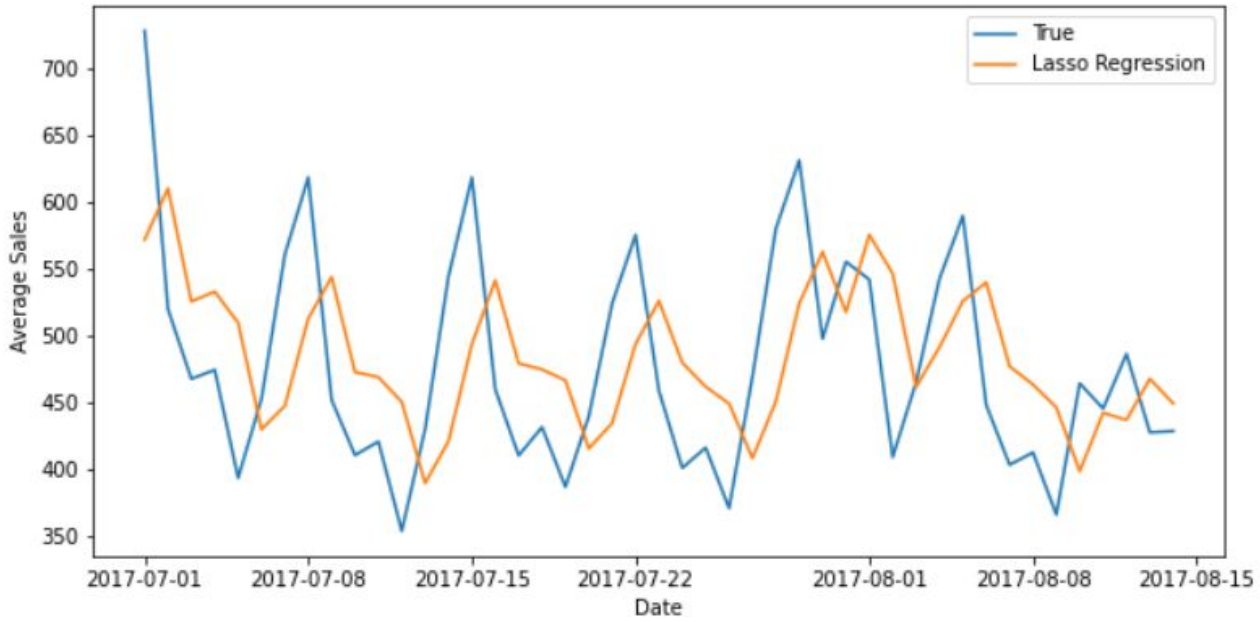
XGBoost Average Sales Predicted vs True



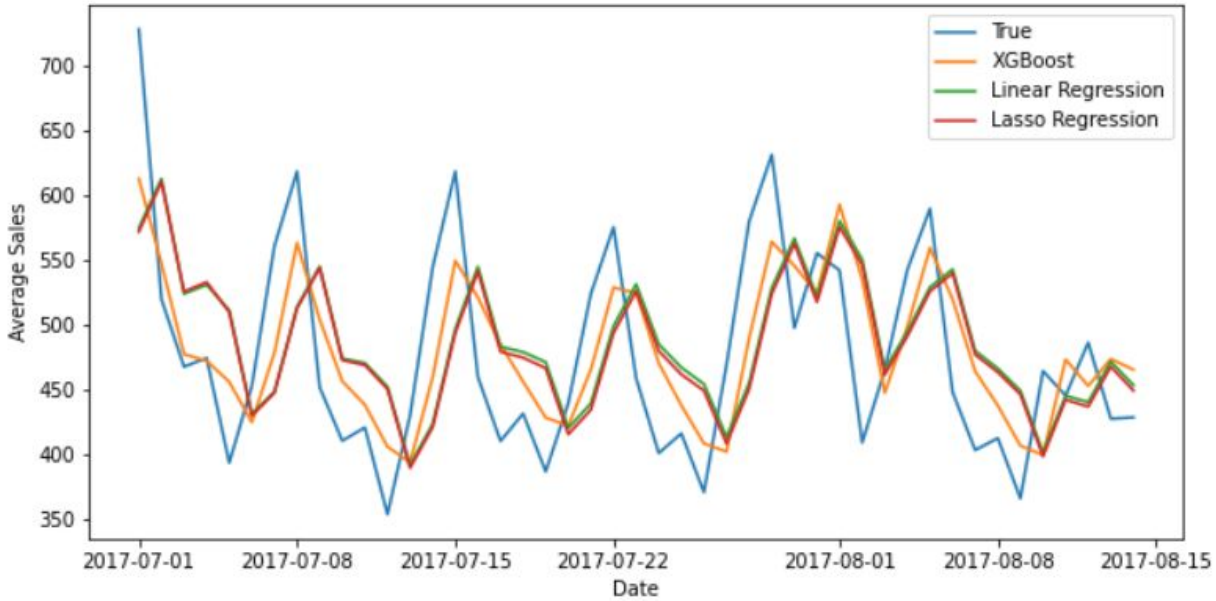
Linear Regression Average Sales Predicted vs True



Lasso Regression Average Sales Predicted vs True



Combined Plots



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.

Naive Bayes Algorithm

$$P(Y|X) = \frac{P(X,Y)}{P(X)}$$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{\sum_Y P(X,Y)}$$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X|Y)P(Y) + P(X|{\sim}Y)P({\sim}Y)}$$

Uses of Naive Bayes Algorithm

1. Sentiment Analysis
2. Text Classification
3. Face Recognition
4. Weather Prediction
5. Recommendation System

Pros and Cons of Naive Bayes Algorithm

Pros

1. Simple to implement and easy to understand
2. Scalable with number of predictors and data points
3. Works well with high-dimensional data
4. Not sensitive to irrelevant features

Cons

1. Number of features must equal number of attributes
2. Assumes that predictors are Independent

References

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.