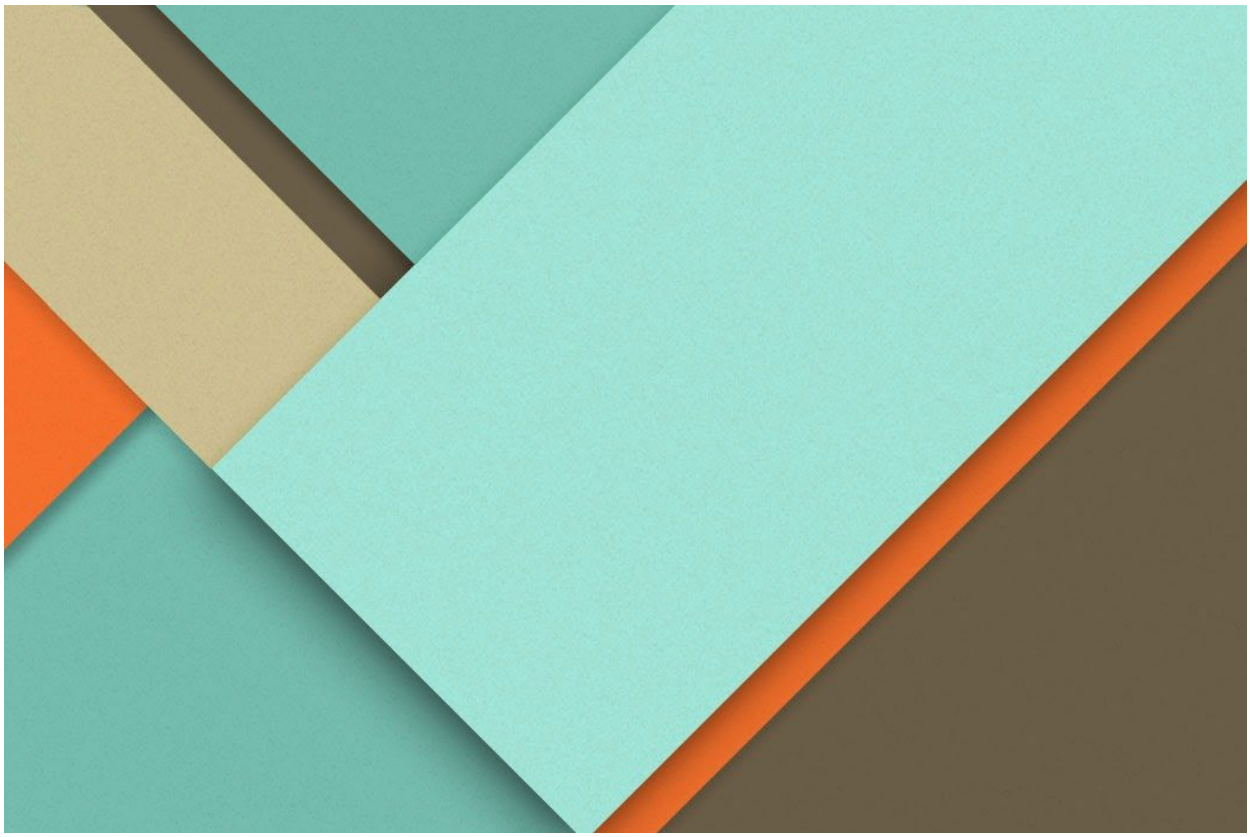
Udacity Machine Learning Engineer Nanodegree



Capstone Project Proposal:

Kaggle competition: M5 Forecasting - Accuracy

03.23.2020

Olga Moreva

Overview

For my capstone project I would like to take part in Kaggle competition M5 Forecasting - Accuracy. The goal of this competition is to predict 'the unit sales of various products sold in the USA by Walmart'[1]

Domain Background

'Walmart Inc. is an American multinational retail corporation that operates a chain of hypermarkets, discount department stores, and grocery stores'[2]. A good forecasting model helps retailers to plan marketing companies, manage cash flow, manage their inventory and optimize their supply chain leading to increase of profit, whereas 'inaccurate business forecasts could result in actual or opportunity losses.' Forecasting is an active area of research, in this project I concentrate on quantitative methods, such as time series analysis and regression methods.

Problem statement

The goal of this competition is to 'to forecast daily sales for the next 28 days' based on the data from 'stores in three US States (California, Texas, and Wisconsin)' . The data 'includes item level, department, product categories, and store details. In addition, it has explanatory variables such as price, promotions, day of the week, and special events. Together, this robust dataset can be used to improve forecasting accuracy.'

Datasets and Inputs

The data provided by the Kaggle platform contain five files.

1. calendar.csv - Contains information about the dates on which the products are sold.
2. sales_train_validation.csv - Contains the historical daily unit sales data per product and store [d_1 - d_1913]
3. sample_submission.csv - The correct format for submissions.
4. sell_prices.csv - Contains information about the price of the products sold per store and date.

[1] All citations in this document if not stated overwise are from the M5 Forecasting - Accuracy challenge https://www.kaggle.com/c/m5-forecasting-accuracy/overview/description
[2] https://en.wikipedia.org/wiki/Walmart

5. *sales_train_evaluation.csv - Available once month before competition deadline. Will include sales [d_1 - d_1941]*

Solution

I am going to explore time series forecasting models (like ARIMA) and tree based algorithms like RandomForest and XGboost to make the predictions.

Benchmark

For a selected item sold on a particular day the benchmark forecast is the average number of items sold at the same day of the week.

Evaluation Metric

The accuracy of the point forecasts will be evaluated using the Root Mean Squared Scaled Error (RMSSE), which is a variant of the well-known Mean Absolute Scaled Error (MASE) proposed by Hyndman and Koehler (2006)[3]. The measure is calculated for each series as follows:

$$RMSSE = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}(Y_t - \hat{Y}_t)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(Y_t - Y_{t-1})^2}}$$

where Y_t is the actual future value of the examined time series at point t, $\hat{Y}_t$ the generated forecast, n the length of the training sample (number of historical observations), and h the forecasting horizon.[4]

Project design

I start with exploratory data analysis and search for missing values and outliers. Then I will try to identify trends and seasonality components. I will also look at difference in sales for special events and SNAP dates. After that I will try out standard time series models (e.g. ARIMA). Depending on the result of previous steps I may come up with additional features. Then I plan to apply some tree-based algorithms like RandomForest or XGBoost.

---

[3] R. J. Hyndman & A. B. Koehler (2006). Another look at measures of forecast accuracy. International Journal of Forecasting 22(4), 679-688.

[4] THE M5 COMPETITION Competitors' Guide, https://mofc.unic.ac.cy/m5-competition/