Time-Series Sales Forecasting using Store, Promotion, and Competitor Data

Team members:

Ashwathy Mohan Menon (am5683), Gefei Zhang (gz2315), Shangzhi Liu (sl4973), Yash Jain (ycj2103), Yuzhao Liu (yl4897)

Background and context to the problem statement

Every business can improve its financial performance by accurately estimating customer demand and future sales of products and services. Demand for a product or service changes continuously. In sales forecasting, we refer to the process of forecasting demand for or sales of a particular product over an extended period of time.Through using machine learning techniques, we can forecast the future sales while abandoning the traditional and inconsistent empiricism methods. Upon this ideology, we will explore and implement various machine learning methods on real sales data records to predict the future sales over a certain time period.

Identification and description of the data set

Dataset: https://www.kaggle.com/c/rossmann-store-sales/data

We will use the historical sales data for 1,115 Rossmann stores to forecast future sales by considering store information, promotions and competitors. We got four files from the link above: (1) train.csv, which has historical data including sales; (2) test.csv, which has historical data excluding sales; (3) sample_submission.csv, where we need to fill in our forecast numbers; (4) store.csv, where supplemental information about the stores are provided. The dataset includes a good number of details about the stores, such as their sales number, customers number, whether they're open or closed, whether it was a state holiday or school holiday, the store type, how far is the nearest competitor store and promotion details. However, some data was missing because some stores were temporarily closed for refurbishment, so we will need to do some data cleaning and preprocessing before applying any model.

Proposed ML techniques you are proposing on applying to solve the problem

Tree Models:
A time series problem can be solved using a tree-based approach. It requires that the time series data be transformed to a supervised learning problem first and then a tree-based approach can be used. While using a tree-based model we need to use a special validation technique called walk-forward validation as using k-fold cross validation will give biased results.

FB Prophet:
Prophet is known to work best with the time series data that have strong seasonal effects and several seasons of historical data. In our project we would also be considering seasonality as one of the important factors and hence this methodology seems to be resourceful. In addition to the same, Prophet is robust to the missing data and handles outliers well. 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.

ARIMA (Autoregressive integrated moving average):

Autoregressive integrated moving average, one of the predictive analysis methods of time series.

AR(p): Autoregressive, it indicates that the value of the current point in time is equal to the regression of the value of the past points in time (lagged values). p is the order (number of time lagged values) of the autoregressive model.

I(d): Integrated. Because time series analysis requires stationarity, unsteady series need to be transformed into stationary series by using integration. d is the degree of differencing (the number of times the data have had past values subtracted).

MA(q): Moving average model, it indicates that the value of the current time point is equal to the regression of the prediction error of the past several time points. q is the order of the moving average model, the number of the prediction error of the past several time points.