Corporación Favorita Grocery Sales Forecasting

Proposal

Hemant N Yadav
November 25, 2017

Domain Background

This project is based on a challenge posted on kaggle.com. It is about prediction product sales for Corporación Favorita(CF) . CF is one of the Ecuadorian-based grocery retailers, operating hundreds of stores, with over 200,000 different products which includes perishable products also such as egg,milk etc. They need good estimated sale of each product so that they can purchase and update inventory accordingly. This estimate eventually benefits two ways to FC. One, help FC not to overstock under popular products specifically perishables and another is have enough inventory in anytime for each product so that customers do not disappointed because not having stock of product.

I have enrolled in this in MLND because I want to learn machine learning and participate on kaggle.com and that is why I started to explore latest featured challenges. Finally, I got in touch with Grocery Sales Forecasting challenge and I selected it for my capstone project as I thought it would be nice experience to compare different machine learning models on it, specifically LSTM to predict sales. Also, this would be a current problem of almost all supermarket chain in India which than can be explored.

Problem Statement

Sales forecasting is remained the one of the main requirement of any grocery stores. Here, predicting the sales for each product is considered as main problem to be solved. Generally, grocery stores use traditional models for prediction with little data taken in consideration. Here, it is important to evaluate which would be better for forecasting sales of CF among two different kinds of models. One is time series based model and another is regression based on causality. Here, comparison is using LSTM for time-series forecasting and Random forest for causal forecasting. This will allow as understanding which is better for forecasting of sales.

Datasets and Inputs

Following data are given by CF for competition. It includes information of each product sold, store information, city and state wise store information, information of promotions running

during period of time on each product. Also, information of oil price changes during some period of time is give which can affect sales.

Training data

- It includes product sale data for each store product id wise for four years.
- It also has information on whether the product is on promotion on specific data and store.

Stores Data

- Details about each stores of CF that includes city, state, type, and cluster.
- Cluster is a grouping of similar stores.

Items Data

- Details about each product id which includes family, class, and perishable.

Transactions

- It is number of sales on each store on each day of train training data.

Oil data

- Daily oil price as fluctuation in oil price affect sales.

Holiday Events

- Details of holidays for a period of time in consideration in training data. It also has information on how many holidays are transferred.

Solution Statement

Sales forecasting can be done using two approaches. As said before, LSTM is used for timeseries based forecasting and Randomforest is for causality based forecasting. As, additional details are in data (more than target and time) multi-variant time series is also worth to compare against single variant time-series.

Benchmark Model

Benchmark model is not considered but we are comparing different models with each other. Also, the competition is still going on and there no well proven methods exist for this data.

Evaluation Metrics

The method given in competition is being used as Evaluation metrics. It is the Normalized Weighted Root Mean Squared Logarithmic Error (NWRMSLE), calculated as follows:

$$NWRMSLE = \sqrt{\frac{\sum_{i=1}^{n} w_i(\ln(\hat{y}_i + 1) - \ln(y_i + 1))^2}{\sum_{i=1}^{n} w_i}}$$

Where for row i,

- yhat(i) is the predicted unit_sales of an item
- y(i) is the actual unit_sales
- n is the total number of rows in the test set.
- W(i) is a weight for each item. Perishable items are given a weight of 1.25 where all other items are given a weight of 1.00

Project Design

Project is about understanding which kind of model would be better to predict sales of CF so here first requirement is to understand data.

Tentative workflow would be as below.

1. Dataset Size reduction: As training data is quite large consisting of 125497040 observations which when loaded in memory takes more than 50 GB. So, first step would be to reduce dataset using data transformation or sampling.
2. Data analysis: This includes study the effect of each factor on sale such as,
    a. Kind of change in sale due to promotion
    b. Effect of oil price changes
    c. Sale volume on each store
    d. Highest consumer's city
3. Data preprocessing: It includes,
    a. Treatment of null, negative or NaN values in each column.
    b. Removal of unnecessary features such as transaction id.
    c. Concatenation of data such as store number and sales data to find sales at each store.
4. Experimenting with models: It includes
    a. Data preparation which comprises onehotencoding for categorical features, time series data modeling etc to train in LSTM. For random forest division of test, train is needed.

   b. Network development for LSTM.
   c. Implementing LSTM with reasonable accuracy and testing with changes in hyper-parameters that include number of epoch, batch_size, activation function and optimizer.
   d. Also, building randomforest with different estimators.
5. Testing the Models: Over here I will compare model with themselves for different hyperparametes and also with each other. Also, modifications to step 4 is expected by this comparison.