Machine Learning Engineer Nanodegree

Capstone Proposal

Walmart Store Sales Forecasting

Nitin Pai

29th Oct 2019

Domain Background

Walmart is an American multinational retail corporation that operates a chain of hypermarkets, department stores and grocery stores. As of Jul 2019, Walmart has 11,200 stores in 27 countries with revenues exceeding $500 billion. A challenge facing the retail industry such as Walmart's is to ensure the supply chain and warehouse space usage is optimized to ensure supply meets demand effectively, especially during spikes such as the holiday seasons.

This is where accurate sales forecasting enable companies to make informed business decisions. Companies can base their forecasts on past sales data, industry-wide comparisons and economic trends. However, a forecasting challenge is the need to make decisions based on limited history. If Christmas comes but once a year, so does the chance to see how strategic decisions impacted the bottom line.

Problem Statement

Historical sales data for 45 Walmart stores located in different regions has been provided. Each store contains many departments, and the sales for each department in each store needs to be projected. Additionally, Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. These markdowns are known to affect sales, but it is challenging to predict which departments are affected and the extent of the impact.

This problem is sourced from a Kaggle Competition and the data source is available here.

Predicting output values (in this case sales) is a regression problem and machine learning can be applied for this problem (Citation #1, #2 & #3).

Datasets and Inputs

The datasets are provided by Walmart on Kaggle's website (dataset URL) and is free to download.

Dataset Details: The dataset contains 4 CSV files –

1. stores.csv

   This file contains anonymized information about the 45 stores, indicating the type and size of store.

2. train.csv

This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file you will find the following fields:

1. STORE – the store number
2. DEPT – the department number
3. DATE – the week
4. WEEKLY_SALES – sales for the given department in the given store
5. IsHoliday – whether the week is a special holiday week

3. test.csv

   This file is identical to train.csv, except weekly sales data is withheld. You must predict the sales for each triplet of store, department, and date in this file.

4. features.csv

   This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

   - STORE – the store number
   - DATE – the week
   - TEMPARATURE – average temperature in the region
   - FUEL_PRICE – cost of fuel in the region
   - MARKDOWN1-5 – anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
   - CPI – the consumer price index
   - UNEMPLOYMENT – the unemployment rate
   - IsHoliday - whether the week is a special holiday week

Solution Statement

As the dataset is labelled, Supervised Learning can be leveraged. The goal is to predict sales for each row in the labelled dataset. This implies the use of regression model(s) for the prediction.

Dataset contains a few non-numeric columns (Categorical) which would need to be converted to numerical. Additionally, if the range of values varies widely, feature scaling would be leveraged to normalize and scale the data.

Multiple regression models would be compared and finally the best model for this problem would be chosen.

Benchmark Model

The given dataset is a typical supervised learning problem and Linear Regression is being chosen as the benchmark model. Hyperparameter tuning and other regression/ensemble methods would be explored and compared with the benchmark.

Also, as this problem is sourced from Kaggle, as a secondary benchmark, the result of the final solution will be compared with this competition's Kaggle Leader board.

Evaluation Metrics

Model prediction for this problem can be evaluated in several ways. However, since Kaggle's evaluation is based on weighted mean absolute error (WMAE), same will be leveraged here:

$$\text{WMAE} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i |y_i - \hat{y}_i|$$

where

- n is the number of rows
- $\hat{y}_i$ is the predicted sales
- y_i is the actual sales
- w_i are weights. w = 5 if the week is a holiday week, 1 otherwise

Project Design

To start with each feature within the dataset would be explored to identify missing values, outliers etc. Data visualization would be leveraged to gain further insight on data distribution. Based on the insight gained, data would be pre-processed – missing values would be filled/dropped, features would be converted from non-numeric to numeric and scaled etc.

To train models, multiple models such as XGBoost, LightGBM and others would be employed for comparison. Comparison would be based on evaluation metrics and computation time. If time permits, Neural Networks and Custom Ensemble models would also be evaluated.

The best model would be further examined and tweaked in order to improve performance. Different set of hyperparameters would be tried out via grid or random search. This result would then be compared to the benchmark.

References

1. Kaggle, Walmart Store Sales Forecasting | https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/overview
2. Walmart.com | https://corporate.walmart.com/our-story/our-business
3. Walmart's Wikipedia Page | https://en.wikipedia.org/wiki/Walmart
4. Trackmaven | https://trackmaven.com/marketing-dictionary/sales-forecasting/
5. Analytics India Mag | https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/
6. Geekflare | https://geekflare.com/choosing-ml-algorithms/
7. Datarobot | https://www.datarobot.com/wiki/regression/