proposal.md

Machine Learning Engineer Nanodegree

Capstone Proposal

Takayuki Sato
August 15th, 2019

Proposal

Domain Background

Sales forecasting is very important for retail business. An accurate sales forecast helps companies to

- Understand companys' performance / customer demand
- Change strategy for the companys' growth
- Conduct beneficial Marketing promotion
- Conduct better management of Inventory
- Adjust price adequately

Many factors, such as, Past sales result / Seasonality / Marketing promotion effect / Consumer Behavior / Price / Competitors' situation / weather / Economic indicators may affect sales forecasting result. There are also many methodologies for time series forecast [1, 2]. Companies need to identify important factors and select an adequate methodology for an accurate sales forecasting.

Problem Statement

Walmart Inc. is an American multinational retail corporation that operates a chain of hypermarkets, discount department stores, and grocery stores. It is the world's largest company by revenue and the largest private employer in the world.

In their kaggle competition(Walmart Store Sales Forecasting), they provide historical sales data for 45 Walmart stores located in different regions. Each store contains many departments. The objective of this competition is to project the sales for each department in each store. To add to the challenge, selected holiday markdown events are included in the dataset. These markdowns are known to affect sales, but it is challenging to predict which departments are affected and the extent of the impact.

This is a typical time series problem. General time series modeling framework(e.g. ETS, ARIMA), Gradient Boosting framework(e.g. xgboost), and Neural Network framework(e.g. LSTM) are expected to be potential solutions.

Datasets and Inputs

I will use the dataset provided by the kaggle competition. (Walmart Store Sales Forecasting)

stores.csv

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

train.csv / test.csv

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

- Store - the store number
- Dept - the department number
- Date - the week
- Weekly_Sales - sales for the given department in the given store
- IsHoliday - whether the week is a special holiday week

Test data is the historical weekly data, which covers to 2012-11-02 to 2013-08-02, with the same fields except for sales.

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
- Temperature - 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

I will confirm the relationship between features and sales, also condct feature engineering if needed.

After that, I will build the following moedls for sales forecast;

- ARIMA
- Decision Tree Regressor(Bench mark)
- Random Forest Regressor
- Decision Tree Regressor
- Graient Boosting Regressor
- Long Short Term Memory

I will also use the ensemple model if needed.

Benchmark Model

I will use Decision Tree Regressor as a bench mark model and make comparison between all models performance and the benchmark.

Evaluation Metrics

I will use the weighted mean absolute error (WMAE) as evaluation metric.

$$WMAE = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i \left| y_i - \widehat{y_i} \right|$$

where

n is the number of rows ŷ i is the predicted sales yi is the actual sales wi are weights. w = 5 if the week is a holiday week, 1 otherwise

Project Design

- Data exploration
  - Confirm basic statistics / trend / seasonality / correlation btween features.
- Data cleaning / Feature engineering
  - Conduct removing missing value / scaling / removing outliers / dimention reduction of features.
- Splitting the data into training(80%) and testing(20%)
- Training
  - Tuning hyper-parameters if needed.
- Forecast / Confirm the performance / Select the best model

References

[1] R. Adhikari and R. K. Agrawal, 'An introductory Study on Time Series Modeling and Forecasting', arXiv:1302.6613, 2013.
[2] K. Bandara, P. Shi, C. Bergmeir, H. Hewamalage, Q. Tran and B. Seaman, 'Sales Demand Forecast in E-commerce using a Long Short-Term Memory Neural Network Mthodology', arXiv:1901.04028v2, 2019.