Capstone Project

November 6, 2017

1 Proposal -- Corporación Favorita Grocery Sales Forecasting

1.1 Domain Background

This project is a Kaggle challenge launched on Oct. 20, 2017. Here is the website of the competition: https://www.kaggle.com/c/favorita-grocery-sales-forecasting.

Sales forecasting estimates future sales, which helps to make business decisions, such as planning stocking amount, allocating resources, predicting sales revenue, managing cash flow, and so on. In retail industry, especially grocery retail industry, sales forecasting is critical, as they may get stuck with perishable goods if overstocked, or lose custormers if understocked. Corporación Favorita, a large Ecuadorian-based grocery retailer, owns 54 stores in different cities/regions in Ecuador. It has this challenge on Kaggle in order to have a better prediction of product sales.

As a research topic, sales forecasting has been studied for a long time. The forecasts are often based on past sales data, industry-wide comparisons, and economic trends. (ref: https://trackmaven.com/marketing-dictionary/sales-forecasting/) Gnerally speaking, methods of forecasting can be classified into three categories: qualitative techniques, time series analysis and projection, and causal models (ref: https://hbr.org/1971/07/how-to-choose-the-right-forecasting-technique) The first depends on qualitative data, like expert opinion, where past information may or may not be taken into consideration. The second focuses me merely on historical data. It explores patterns and pattern changes to predict future sales. The third uses highly refined and specific information about relationships between system elements.

In this grocery sales forecasting problem, historical datasets are provided. From the perspecitve of machine learning, I will tackle the problem with a supervised regression method, model product unit sales with either a parametric or non-parametric method.

1.2 Problem Statement

The task **T** of this challenge is to forecast product sales of a specific item in a specific store for some given dates. In this project, experience **E**, historical datasets, are provided to tackle this problem: sale records from Jan.1 2013 to Aug.16 2017, along with some other historical data, such as oil price, transaction, etc. Evaluation metric **P** is Normalized Weighted Root Mean Squared Logarithmic Error (NERMSLE). Based on historical sales records and related factors **E**, a supervised machine learning method can be employed to perform task **T** -- predict product sales, mesured with metric **P**.

1.3 Datasets and Inputs

Datasets are provided by Corporación Favorita: https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data , which include:

1.train.csv

- id: This is meaningless for model training, and will be dropped - date: From 2013-01-01 to 2017-08-15 - store_nbr: conintunous integer from 1 to 54 - item_nbr: item id, un-continuous integers - unit_sale: continuous float number with min=-0.000153 max=89440 - onpromotion: bool 0 1, and missing entries

column	date	id	stor_nbr	item_nbr	unit_sales	onpromotion
datatype	int32	date	int16	int32	float32	float32

2.holidays_events.csv

- date: 312 unique dates from 2012-03-02 to 2017-12-26
- type: ['Holiday', 'Transfer', 'Additional', 'Bridge', 'Work Day', 'Event']
- locale: ['Local', 'Regional', 'National']
- locale_name: ['Manta', 'Cotopaxi', 'Cuenca', 'Libertad', 'Riobamba', 'Puyo','Guaranda', 'Imbabura', 'Latacunga', 'Machala', 'Santo Domingo','El Carmen', 'Cayambe', 'Esmeraldas', 'Ecuador', 'Ambato', 'Ibarra','Quevedo', 'Santo Domingo de los Tsachilas', 'Santa Elena', 'Quito','Loja', 'Salinas', 'Guayaquil']
- description:103 entries (don't understand)
- transferred: [False, True]

column	date	type	locale	locale_name	description	transferred
datatype	date	string	string	string	string	bool

3.stores.csv

- store_nbr: integer from 1 to 54 - city: ['Quito', 'Santo Domingo', 'Cayambe', 'Latacunga', 'Riobamba', 'Ibarra', 'Guaranda', 'Puyo', 'Ambato', 'Guayaquil', 'Salinas', 'Daule', 'Babahoyo', 'Quevedo', 'Playas', 'Libertad', 'Cuenca', 'Loja', 'Machala', 'Esmeraldas', 'Manta', 'El Carmen'] - state: ['Pichincha', 'Santo Domingo de los Tsachilas', 'Cotopaxi', 'Chimborazo', 'Imbabura', 'Bolivar', 'Pastaza', 'Tungurahua', 'Guayas', 'Santa Elena', 'Los Rios', 'Azuay', 'Loja', 'El Oro', 'Esmeraldas', 'Manabi'] - type: ['D', 'B', 'C', 'E', 'A'] - cluster: [13, 8, 9, 4, 6, 15, 7, 3, 12, 16, 1, 10, 2, 5, 11, 14, 17]

column	store_nbr	city	state	type	cluster
datatype	int64	string	string	string	int64

4.oil.csv

- date: from 2013-01-01 to 2017-08-31
- dcoilwtico: continuous value from 26.19~110.62

column	date	dcoilwtico
datatype	date	float64

5.transactions.csv
- date: from 2013-01-01 to 2017-08-15 - store_nbr: 54 store numbers - transactions: integers between 5 and 8358 Transactions data in testing set is not provided, so this dataset will not be used in training, although transactions are highly related to unit sales we want to predict.

6.items
- item_nbr: 4100 discrete values - family: [’GROCERY I’ ’CLEANING’ ’BREAD/BAKERY’ ’DELI’ ’POULTRY’ ’EGGS’ ’PERSONAL CARE’ ’LINGERIE’ ’BEVERAGES’ ’AUTOMOTIVE’ ’DAIRY’ ’GROCERY II’ ’MEATS’ ’FROZEN FOODS’ ’HOME APPLIANCES’ ’SEAFOOD’ ’PREPARED FOODS’ ’LIQUOR,WINE,BEER’ ’BEAUTY’ ’HARDWARE’ ’LAWN AND GARDEN’ ’PRODUCE’ ’HOME AND KITCHEN II’ ’HOME AND KITCHEN I’ ’MAGAZINES’ ’HOME CARE’ ’PET SUPPLIES’ ’BABY CARE’ ’SCHOOL AND OFFICE SUPPLIES’ ’PLAYERS AND ELECTRONICS’ ’CELEBRATION’ ’LADIESWEAR’ ’BOOKS’] - class: 337 discrete values - perishable: 0 and 1

column	item_nbr	family	class	perishable
datatype	int64	string	int64	int64

7.test.csv
Similar to train.csv without column ’unit_sales’.

8.sample_submission.csv
This demonstrates the format of submission file.

I explored each data file from these 4 aspects: 1. A peek of sample items (display(df.head(5)) or display(df.tail(5))) 2. A summary of each data file (df.describe()) 3. Data types (df.dtypes) 4. Data type of each column (df.dtypes) 5. Unique values of each variable (display(df[’column_name’].unique()))
Details are here: https://github.com/highhigh/kaggle/blob/master/Favorita_Grocery_Sales_Forecasting/groc

Here oil price data will be merged with train data according to date. I’m still thing about how to use holiday_envents, items, and stores data, or even don’t use them at all.

The training dataset is large: 125,497,040 items across 4 years and 8 months. By setting proper datatypes of columns in Pandas function read_cvs, original 6 GB training data can be reduced to 4GB, which is still very large. There are many approaches to handle this dataset. One approach is to split data according to different stores, and train models for each store. Another one is to split data according to different items, and train models for each item. I’ll run them to see which performs better.

1.4 Solution Statement

There are many factors affecting grocery sales, such as holiday_events and oil prices mentioned above. The first step is to examine which data will be used to form features among many data provided.
The second step is to train a regression model, say linear regression, to model grocery sale, and use cross-validation skill to choose proper model parameters.
The last step is to run the trained model on testing items provided, and upload to Kaggle to evaluate the model’s performance.

1.5 Benchmark Model

A naive model is to predict unit_sale based on historical values. This non-parametric model is an unweighted average of the historical values of the same item in the same store from the same date of previous 4 years. Unit_sales of items that can not be found in history record are filled with 0.
Evaluation from Kaggle website shows prediction from this method scors 0.908 (https://www.kaggle.com/c/favorita-grocery-sales-forecasting/submissions?sortBy=date&group=all&page=1)

1.6 Evaluation Metrics

Evaluation metric for the challenge is provided here https://www.kaggle.com/c/favorita-grocery-sales-forecasting#evaluation.

To be specific, the prediction is evalueated on the Normalized Weighted Root Mean Squared Logarithmic Error (NERMSLE):
$\sqrt{\frac{\sum_{i=1}^{n} w_i (ln(\hat{y}_i + 1) - ln(y_i + 1))^2}{\sum_{i=1}^{n} w_i}}$,
where for row i, $\hat{y}_i$ is thepredicted unit_sales of an item and y_i is the actual unit_sales; n is the total number of rows in the test set. w_i, the weights is given based on if an item is perishable or not: Perishable items have a weight of 1.25 and all other items have weights of 1.00.

This metric is able to avoid penalizing large differences when both the predicted and true numbers are large and it works well to predict values across a large range of orders of magnitudes.

1.7 Project Design

This project is to predict daily grocery unit sale for several stores of a grocery retailer for a period of two weeks in late Auguest in 2017, based on historical sales record.
Grocery unit sale, the target to be predicted, is numerical, so it is a supervised regression problem in machine learning. Training data, historical sale records, are split into two parts: One is used to train a model, say linear regression, the other one is used to validate the model according to the pre-defined evaluation metric. After training, the model achieved is then used for prediction.
The Kaggle competition saves a testing data from public, where true target values are hidden. The trained model is applied to this testing data set to generate prediction of grocery sales, and results are then submitted to Kaggle for evaluation.