Final Project Report

Forecasting Walmart Sales Based on Ensemble Learning

Li Shengshu
Student ID: 20746064

Abstract

In this project, two Machine Learning models, Random Forest (then will be abbreviated as RF) and Gradient Boosting Decision Tree (then will be abbreviated as GBDT) are used to predict the sales units of different merchandises of Walmart in nest 28 days, via an original dataset of 1913 days of data from January 29, 2011 to April 24, 2016, containing different features of those goods.

The report can be divided into 5 parts according to project process order. The first part of the report is data analysis. In this part, several data visualization methods are used to analyze some basic characteristics of the data, such as classifying the sales data according to commodity categories or sales departments, and observing the characteristics of the data in time series under different classifications.

The second part is the data processing part. On the one hand, the data format provided in the Kaggle competition is different from the format required by the selected models. Thus, I need to convert the data format to a certain extent. On the other hand, some logical operation can be processed among the different characteristics of the original data in order to create new features with practical significance to enhance the prediction ability of the model.

The third part is data cleaning. In the data processing of the previous part, some missing values will be generated. These missing values need to be filled in or deleted according to certain logic.

The fourth part is model calculation. The processed and cleaned data are put in the model and the result are analyzed via Kaggle score.

The final part is summary, which make a conclusion of the final project.

Data Analysis

By reading the original data, I find that the commodities are sold in 3 different states: California,

Texas and Wisconsin. In California, the Walmart have 4 different stores and in Texas and Wisconsin, the Walmart have 3 stores receptively. If the goods are classified by their category, they can be divided into 3 different parts: Hobbies, Household and Foods. To specify, Hobbies and Household have 2 different departments and Foods have 3 different departments.

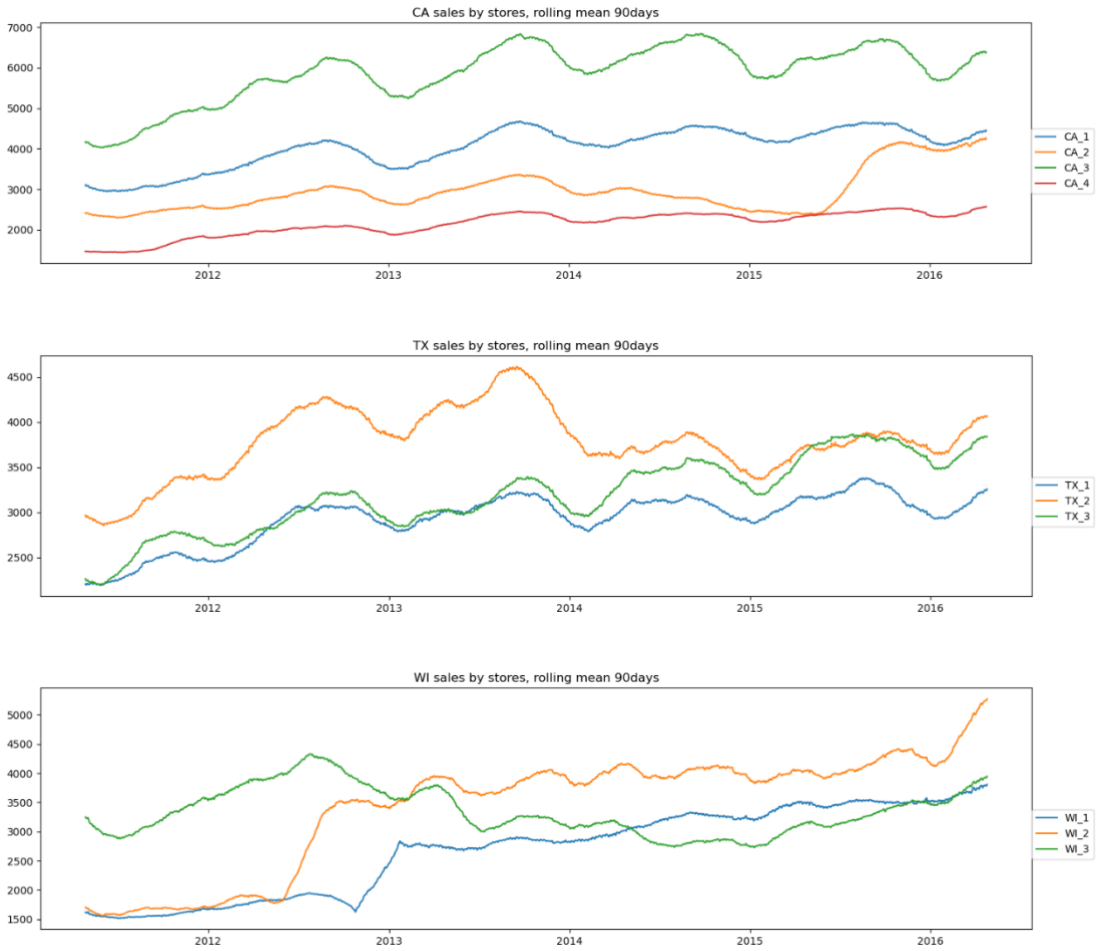


Figure 1: The sales volume of different stores divided by state.

From above figures, it is easy to find that the sales in various states are highly cyclical, peaking at the end of the year and bottoming at the beginning of the year, which suggests that, for later model calculation, I ought to select the data targeted to make time period of train and test data are consistent. In other words, if I chose the train the model by splitting original data randomly, it will have more noise and consume more time to training the model.

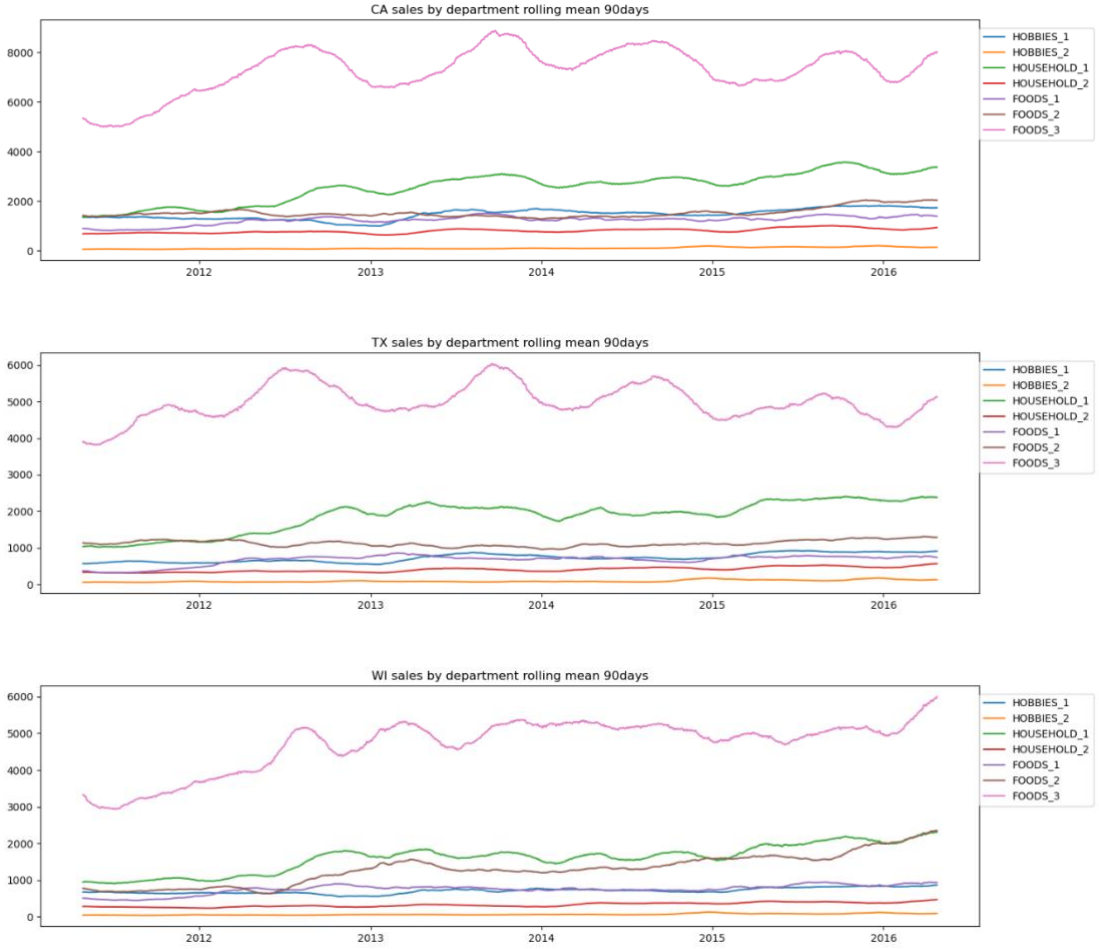
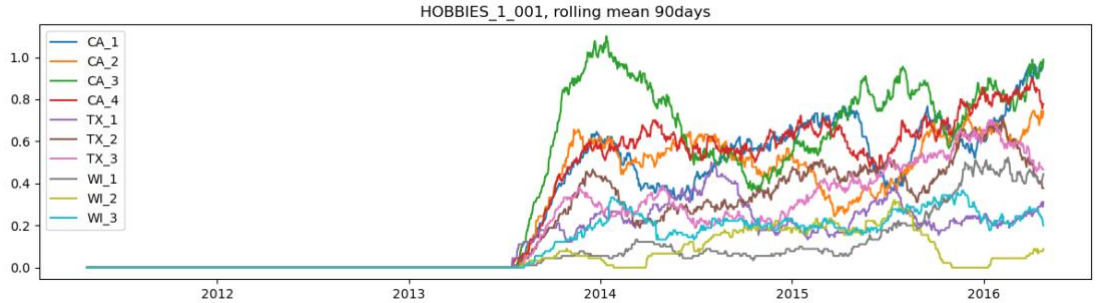


Figure 2: The sales volume of the departments divided by state.

From the above figures, I also find that different departments and categories are cyclical. Furthermore, the sales volume of different departments or categories are significant different. Thus, this information can also be used in data pre-selection.



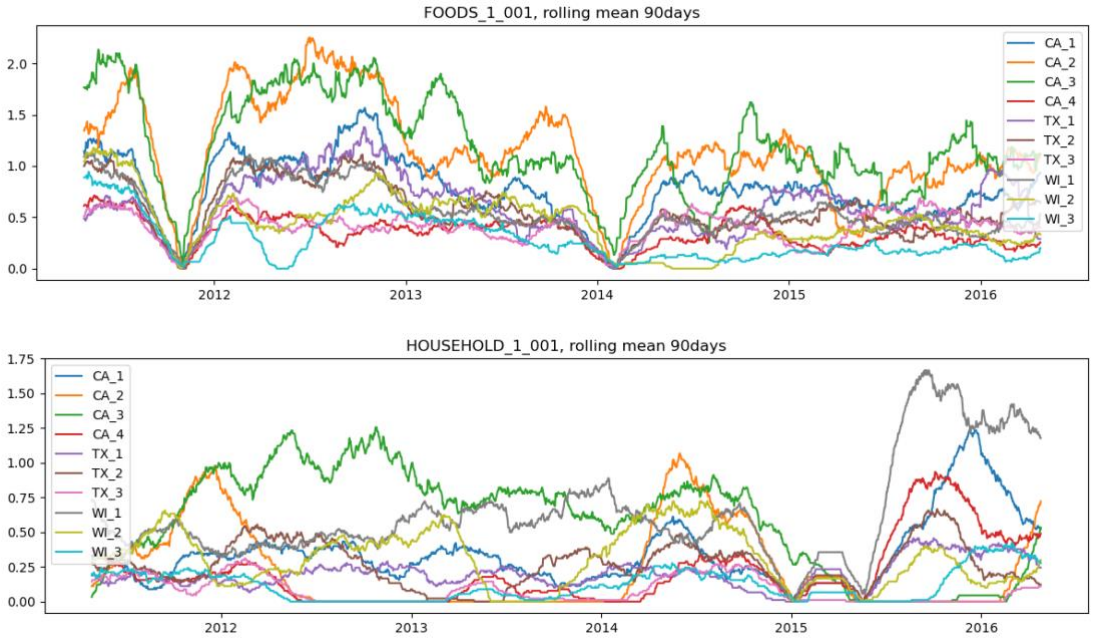


Figure 3: The sales volume of certain items divided by department.

From the above figures, I find that different items in different year have different sales volumes. For example, the item, HOBBIES_1_001, is only sold after second half of 2013, so it might be better to choose recent data for training.

Data Processing

The format of original dataset is highly structured. However, the models I decided to used do not support time-series-data. Or in other world, the dates of each item should be regard as a kind of features.

id part 1	⋯	id part p	day 1	⋯	day n
$item_{1,1}$	⋯	$item_{1,1}$	$sales_{1,1}$	⋯	$sales_{1,n}$
⋮	⋱	⋮	⋮	⋱	⋮
$item_{m,1}$	⋯	$item_{m,p}$	$sales_{m,1}$	⋯	$sales_{m,n}$

⇒

	id part 1	⋯	id part p	day	sales
$item\_day_{1,1,1}$	$id\_1_{1,1}$	⋯	$id\_p_{1,1}$	day_1	$sales_{1,1,1}$
⋮	⋮		⋮		⋮
$item\_day_{m,1,1}$	$id\_1_{m,1}$		$id\_p_{m,1}$		$sales_{m,1,1}$
⋮	⋮	⋱	⋮	⋮	⋮
$item\_day_{m,p,1}$	$id\_1_{m,1}$		$id\_p_{m,1}$		$sales_{m,p,1}$
⋮	⋮		⋮		⋮
$item\_day_{m,p,n}$	$id\_1_{m,n}$	⋯	$id\_p_{m,n}$	day_n	$sales_{m,p,n}$

Figure 4: The sketch map of the transformation from original format to target format.

Besides, the original datasets also provide a calendar, providing more information of each day and also indicate whether a day is an anniversary. Thus, the features can be enlarged by certain calculations.

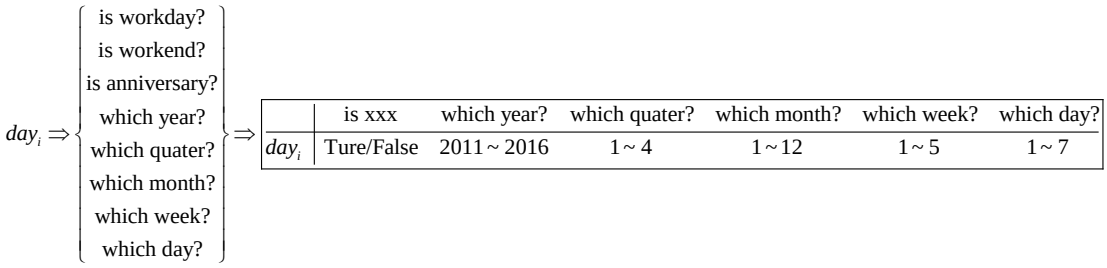


Figure 5: The sketch map that how to enlarge the features.

According to the previous time series analysis, combined with the characteristics already provided by the original data set, the expanded data set should be sufficient for calculation of subsequent model calculation in formation and feature sizes.

Data Cleaning

In the original data, the prices of some commodities are missing at some times. When the data format is converted, these missing elements are still retained. In the face of these missing values, the interpolation is an appropriate way used to calculate the value of such prices. On the other hand, during the process transforming the format of original dataset shown on Figure, the method called “Cartesian Product” have been used, leading to a dilemma that dozens of inexistence items are created, and they ought to be dropped.

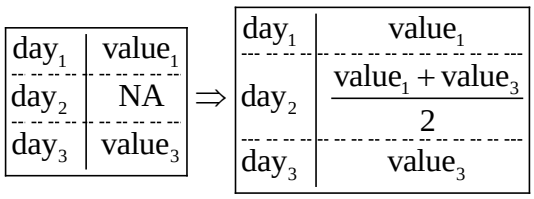


Figure 6: The sketch map that how to interpolate the missing value.

$$A = \{a, b\}, B = \{1, 2\} \qquad \textit{Cartesian Product}: A \times B = \{(a,1), (a,2), (\text{b},1), (b,2)\}$$

$$\textit{Orignal Dataset} = \{(a,1), (a,2), (\text{b},1)\}$$

Figure 7: The sketch map that why inexistence items are created.

Model Calculation

The first model I choose from ensemble learning is RF.

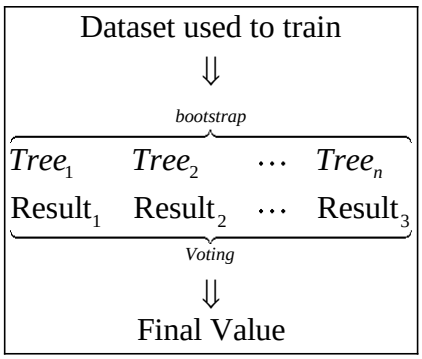


Figure 8: The sketch map of Random Forest algorithm.

In ordinary process, the samples and the features ought to be chosen randomly and classified by Trees. However, in previous data analysis, some periodic data relationships have been found. In chronological order, the next 28 days to be predicted will be in May and June, so the training data can be selected from these two months. More strictly speaking, the data can be chose from the past two years for training.

The result shows that, the pre-selection can significantly reduce both the public and private score.

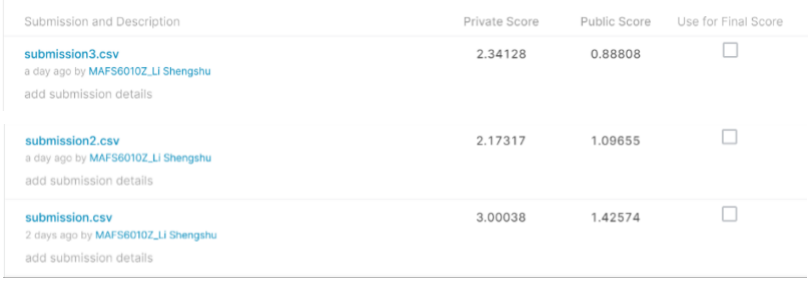


Figure 9: The public and private score of Random Forest model.
("submission.csv" representing training data without pre-selection, "submission2.csv" representing training data with pre-selection of month and "submission3.csv" representing training data with pre-selection of both month and year)

The second model I chose is GBDT.

$$\hat{y}_i = F_M(x_i) = \sum_{m=1}^{M} h_m(x_i)$$

$$F_m(x) = F_{m-1}(x) + h_m(x)$$

$$h_m = \arg\min_h \sum_{i=1}^{n} l(y_i, F_{m-1}(x_i) + h(x_i))$$

$$l(y_i, F_{m-1}(x_i) + h_m(x_i)) \approx l(y_i, F_{m-1}(x_i)) + h_m(x_i)\left[\frac{\partial l(y_i, F(x_i))}{\partial F(x_i)}\right]_{F=F_{m-1}}$$

Figure 10: The formulas used in Gradient Boosting Decision Tree

The result of GBDT has better private score than RF, but has lower public score than

RF.

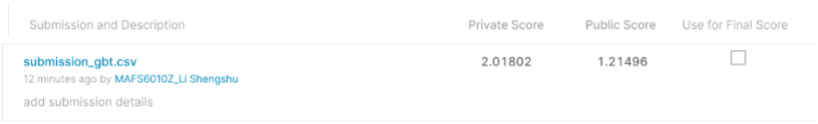


Figure 11: The public and private score of Gradient Boosting Decision Tree model.

Summary

This paper mainly uses the Ensemble Learning method to predict the sales volume of each commodity of Wal Mart in the next 28 days. In the whole prediction process, this paper analyzes, processes and cleans the data in order, and finally uses Random Forest and Gradient Boosting Decision Tree to predict the sales volume. The final result is satisfactory and explainable in some extent.