香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

M5 Forecasting - Accuracy

SHANG Zhiheng
CHEN Yuying
LOU Ruoyu

Contents

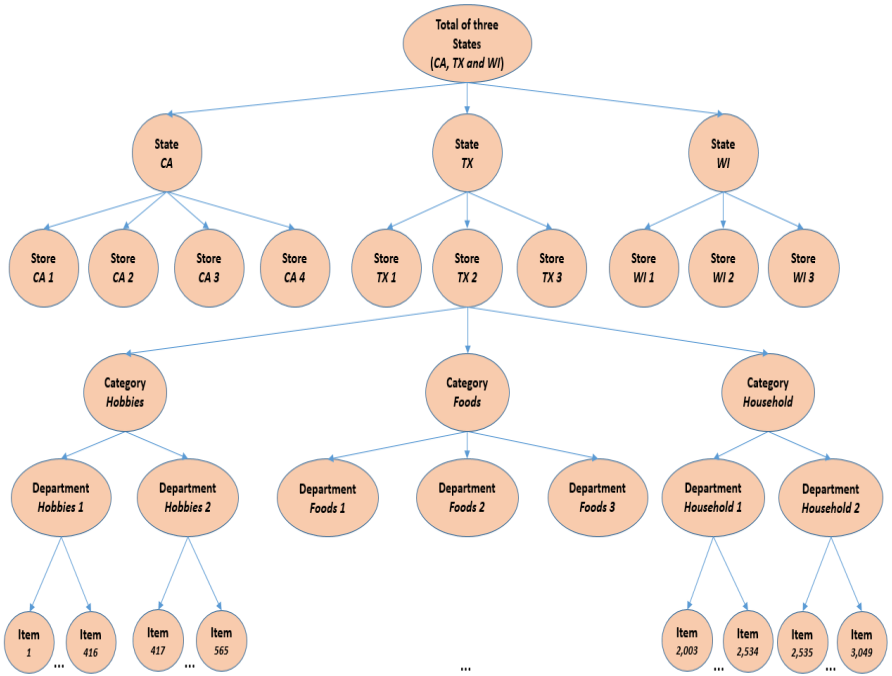
CONTENTS PAGE

01

PART

Introduction

Introduction



- We aim to estimate the point forecasts of the unit sales of various products sold in the USA by Walmart.
- The datasets involves the unit sales of 3049 products sold in three states of the USA (California (CA), Texas (TX), and Wisconsin (WI)), which are organized in the form of 42,840 hierarchical time series.

Contents

CONTENTS PAGE

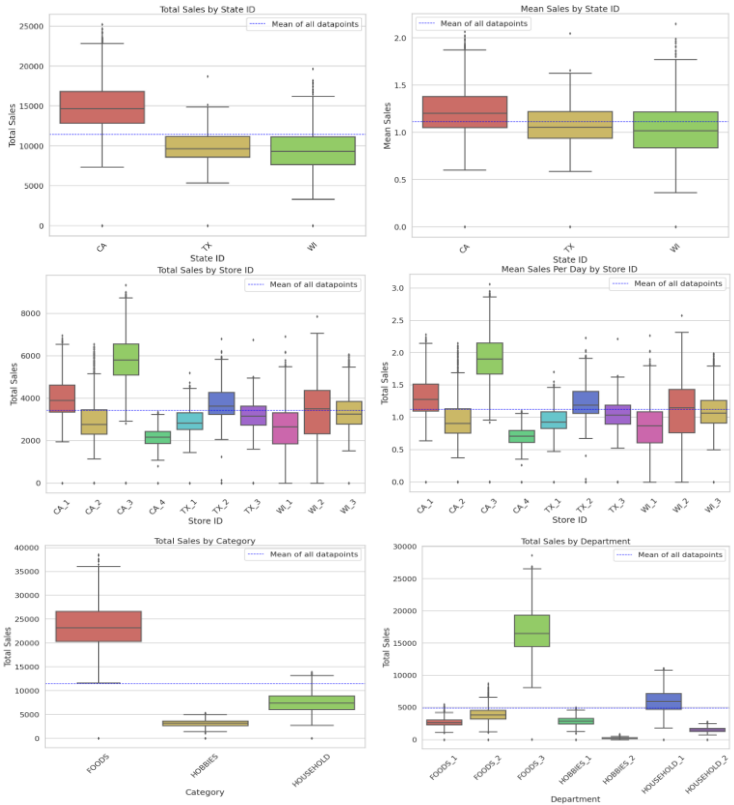
02

PART

Explanatory Data Analysis

Explanatory Data Analysis

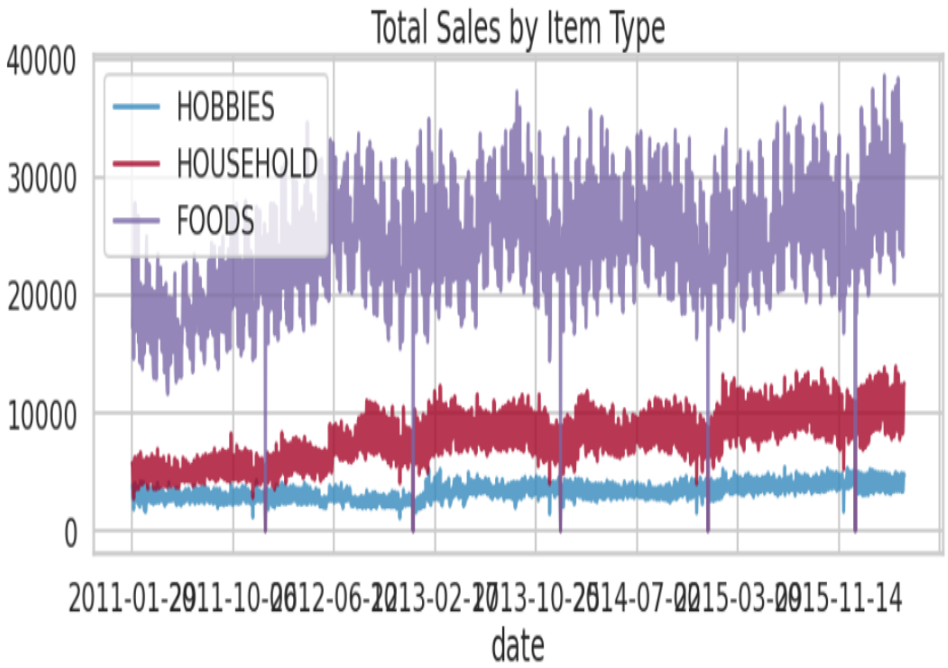
Boxplot



- **State Level**: California has the higher total sales
- **Store Level**: CA_3 store has the highest total sales and mean sales, while CA_4 store is relatively low.
- **Category & Department Level**: FOODS has relatively high total sale score. Particularly, the FOODS_3 is on a high total sale position.

Explanatory Data Analysis

Time Series Plot

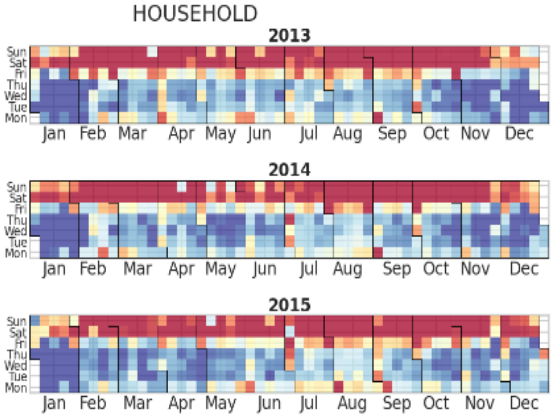


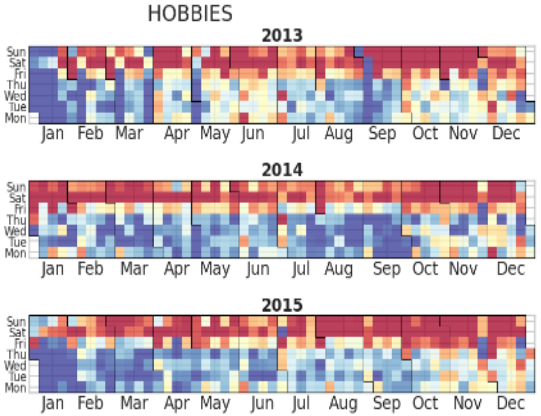
- Food has long been highest position of total sales, then followed by household and then hobbies.
- The unique sales peaks also found in this picture, which means people tend to buy more food in particular time of the year.

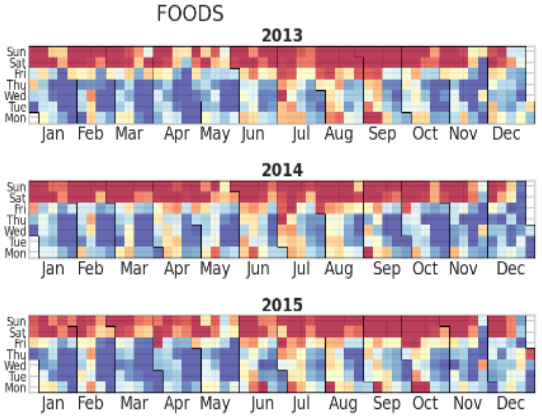
Explanatory Data Analysis

Calendar Heat Maps

- Weekends are the most popular shopping days for customers among all item categories.
- FOODS category tended to take the highest position of sales amount and then experience a long decrease in each month.
- HOUSEHOLD and HOBBY items sold much less in January.







Contents

CONTENTS PAGE

03

PART

Data Processing & Feature Engineering

Deal with the calendar file

- **Reduce the memory usage**
  - Change the type of categorical variables from string to category.
  - Restrict the precision of the numeric variables to be int8.
  - Transform the type of binary variables into bool.
- **Create new variables**
  - Combined “event name 1” and “event name 2” to a comprehensive variable named “event name”.
  - Combined “event_type 1” and “event type 2” to a comprehensive variable named “event_type”
  - Create a variable named “days_to_event” indicating the number of days from the examined date to the next event.
- **Drop some repetitive variables**
  - “year”, “weekday”, “month”, “wday”, “event_name_1”, “event_name_2”, “event_type_1” and “event_type_2” are dropped.

Data Processing & Feature Engeering

Merge the files and Process the whole data

- **Merge the files**
  - Merge the calendar file and the sell price file according to “wm_yr_wk”.
  - “sales_evaluation.csv” is merged in according to ”item_id”, “store_id” and “d”.
- **Reduce the memory usage**
  - The type of the category variables are set to be categorical. (“store_id”, “item_id” and “id”).
  - The numeric variables are set to be float16 and int16.(“demand” and “d”).
  - Create a new variable named “is_snap_avaliable” to replace three variables “snap_TX”, “snap_CA” and “snap_WI”.
- **Created new variables**
  - Some new features can be generated based on the time series to capture the information shown by the dependent variable(demand) itself. (The details are shown in the next page)
- **Encode the Categorical variables**
  - Use the function “CatBoostEncoder” to encode “store_id”, “item_id” and “dept_id”.
  - Binary variables “is_in_CA”, “is_in_TX” and “is_in_WI” are employed to replace “state_id”.
  - Binary variables “is_foods”, “is_household” and “is_hobbies” are employed to replace “cat_id”.

Data Processing & Feature Engeering

variables	meaning
cumulative_mean_demand_29d_ago	the cumulative mean value of the demand after a lag of 29 days for each examined date.
cumulative_md_low_demand_29d_ago	the cumulative median value of the demand after a lag of 29 days for each examined date.
ewm_mean_1w_demand_29d_ago	Exponential moving weighted mean value of the demand after a lag of 29 days at a span equals 7*1.
ewm_mean_4w_demand_29d_ago	Exponential moving weighted mean value of the demand after a lag of 29 days at a span equals 7*4.
rolling_mean_1w_demand_29d_ago	the mean value of the demand after a lag of 29 days in a 7*1 days rolling window.
rolling_mean_4w_demand_29d_ago	the mean value of the demand after a lag of 29 days in a 7*4 days rolling window.
rolling_mean_8w_demand_29d_ago	the mean value of the demand after a lag of 29 days in a 7*8 days rolling window.

Contents

CONTENTS PAGE

04

PART

Model Fitting

Model Fitting

Evaluation Metrics & Model Selection

Evaluation Metrics

$$WRMSSE = \sum_{i=1}^{42,840} w_i * RMSSE$$

$$RMSSE = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}\left(Y_t - \hat{Y}_t\right)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(Y_t - Y_{t-1})^2}},$$

Model Selection

Due to the large number of data features and calculating power limitations, CatBoost is finally chosen as the primary model in this project, since CatBoost converges to a good solution in the shortest time comparing with XGBoost and LightGBM within gradient boosted decision trees (GBDT)

Model Fitting

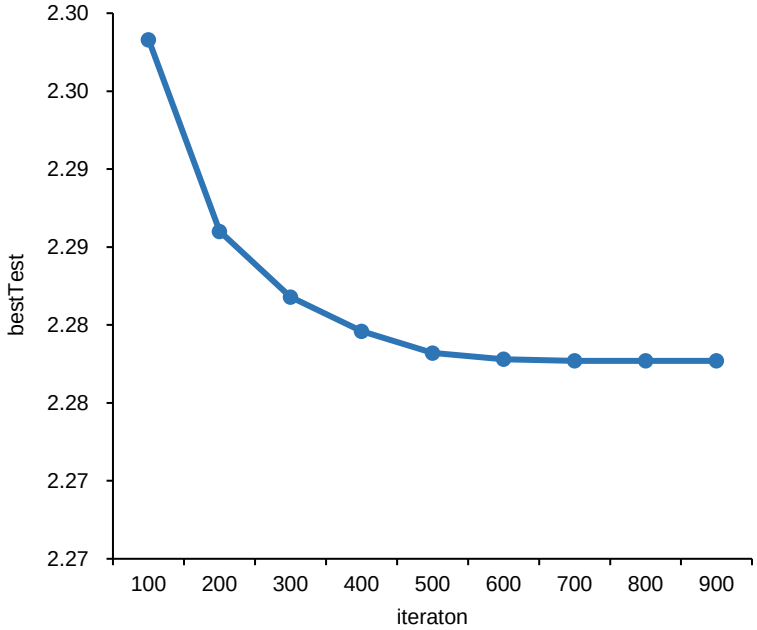
Fiting Strategy

Parameter	Description	Value
Learning Rate	This setting is used for reducing the gradient step. It affects the overall time of training: the smaller the value, the more iterations are required for training.	0.1
Depth	In most cases, the optimal depth ranges from 4 to 10. Values in the range from 6 to 10 are recommended.	10
Loss Function	The metric to use in training. The specified value also determines the machine learning problem to solve.	RMSE
Leaf estimation iterations	This parameter defines the rules for calculating leaf values after selecting the tree structures. The default value depends on the training objective and can slow down the training for datasets with a small number of features	1
Evaluation Metric	Standard of evaluation	RMSE

Split the dataset:
The train part contains combined data before date '2015-05-01', valid part contains combined data laying before sales date ’1942’ and after date '2015-05-01', while test part contains those laying after sales date ’1914’.

Tune the parameters:
The proper parameters are shown in the left table.

Model Fitting



- In the first fitting round, we get best test score at 2.2777 and corresponding best iteration 625th step. The whole model then shrink to first 626 iterations. When it comes to 900 iteration steps, the Overfitting Detector of CatBoost is triggered

- In the final round fitting, the iteration is set as the product of former best iteration step (625) and train-valid-ratio (ratio of valid data size and train data size). In this round, the model shrinkage in combination with learning continuation is not implemented, which means the fitting process will finish all iterations set before.

Contents

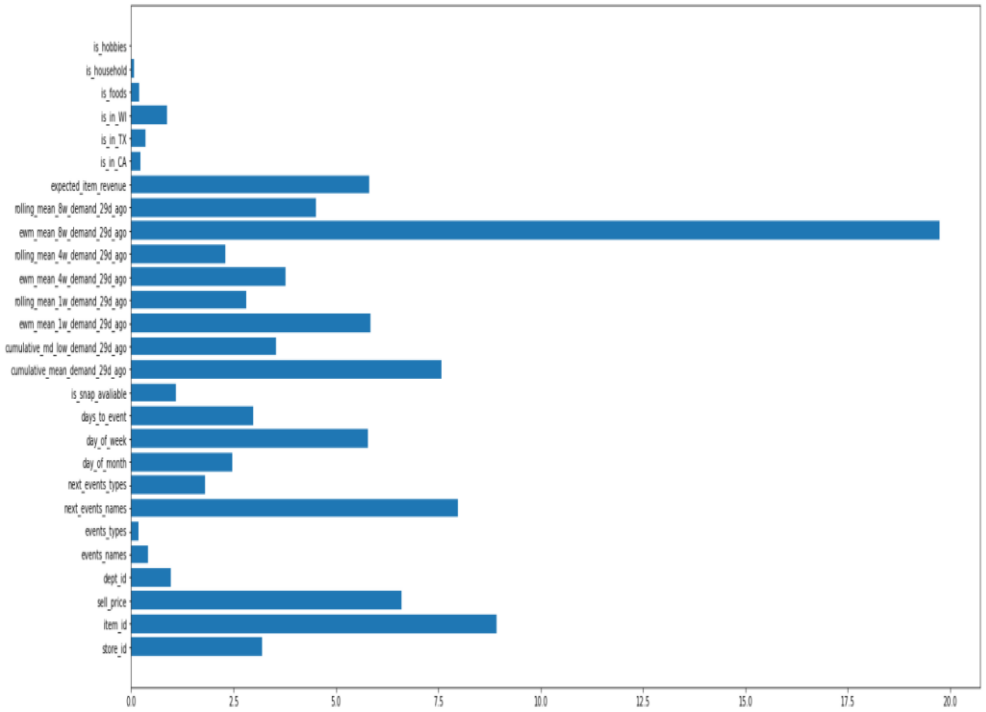
CONTENTS PAGE

05

PART

Results & Discussion

Results & Discussion



- The feature ‘ewm_mean_8w_demand_29d_ago’ has the highest importance.
- Among all exponential moving weighted mean value of the demand at a span level of 7*X, it can be found that the corresponding feature tends to be more important as X increases.

Results & Discussion

- According to the Kaggle competition platform, finally we get 0.603 as private score and 0.5312 as public score.
- Some other models like LightGBM and XGBoost are tried. However, due to the limitation of calculation resources and time, only CatBoost succeed in finishing the result.
- Progress will be made if more effective data conversion methods could be found

	Competition Notebook	Run	Private Score	Public Score
M5	M5 Forecasting - Accuracy	19448.2s	0.60280	0.53116

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

THANK YOU!

SHANG Zhiheng
CHEN Yuying
Lou Ruoyu