AI project WeArealLsMART.pdf

by GAO Yuxin

Submission date: 31-May-2020 10:21PM (UTC+0800)
Submission ID: 1335190546
File name: AI_project_WeArealLsMART.pdf (1.41M)
Word count: 3171
Character count: 15545

M5 Forecasting - Accuracy

Estimate the unit sales of Walmart retail goods

Group Name: WeArealLsMART

Group Members

Student Name	Student ID	Contribution
WANG Su	20659651	Report & PPT
WANG Boyu	20660052	Coding
ZHAO Yinuo	20660399	Coding
GAO Yuxin	20661496	Report & PPT
LI Qian	20662024	Report & PPT

DATE: 2020.05

1. Instruction

In reality, store holders and goods supplier care more about how much specific goods will one store sell each month in a year. This type of forecasting relies on science and historical data. Inaccurate business forecasts could result in actual or opportunity losses. So in this competition, we use hierarchical sales data from Walmart, the world's largest company by revenue, to forecast daily sales for the next 28 days. In following procedures, we will advance the theory and practice of forecasting by identifying the methods that provide the most accurate point forecasts for time series, and provide 28 days ahead point forecasts, as well as the corresponding median and 50%, 67%, 95%, and 99% prediction intervals.

The report will include first data explanation part, where we will show the data we collected and the processing procedures, second the models listing part, which will contain kinds of useful models from time series and machine learning and using the evaluation index to obtain the best one in order to do the forecast, and at last the conclusion of our project.

2. Data Explanation

First, we have grouped unit sales data, starting at the product-store level and being aggregated to that of product departments, product categories, stores, and three geographical areas: the States of California (CA), Texas (TX), and Wisconsin (WI). Second, it includes item level, department, product categories, and store details, ranging from 2011-01-29 to 2016-06-19. Besides the time series data, it includes explanatory variables such as sell prices, promotions, days of the week, and special events (e.g. Super Bowl, Valentine's Day, and Orthodox Easter) that typically affect unit sales. Together, this robust dataset can be used to improve forecasting accuracy. Last, it focuses on series that display intermittency, i.e., sporadic demand including zero.

Then we will show the data we collected: File1: "calendar.csv" contains information about the dates the products sold. File 2: "sell_prices.csv" contains information about price of the products in per store and date, as shown in figure 1 File 3: "sales_train.csv" contains the historical daily unit sales data per product and store, as shown in figure 2.

Figure 1: the data format of calendar.csv and sell_price.csv

date	wm_yr_wk	weekday	wday	month	year	d	event_name_1	event_type_1	event_name_2	event_type_2	snap_CA	snap_TX	snap_WI
2011-01-29	11101	Saturday	1	1	2011	d_1	NaN	NaN	NaN	NaN	0	0	0
2011-01-30	11101	Sunday	2	1	2011	d_2	NaN	NaN	NaN	NaN	0	0	0
2011-01-31	11101	Monday	3	1	2011	d_3	NaN	NaN	NaN	NaN	0	0	0
2011-02-01	11101	Tuesday	4	2	2011	d_4	NaN	NaN	NaN	NaN	1	1	0
2011-02-02	11101	Wednesday	5	2	2011	d_5	NaN	NaN	NaN	NaN	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2016-06-15	11620	Wednesday	5	6	2016	d_1965	NaN	NaN	NaN	NaN	0	1	1
2016-06-16	11620	Thursday	6	6	2016	d_1966	NaN	NaN	NaN	NaN	0	0	0
2016-06-17	11620	Friday	7	6	2016	d_1967	NaN	NaN	NaN	NaN	0	0	0
2016-06-18	11621	Saturday	1	6	2016	d_1968	NaN	NaN	NaN	NaN	0	0	0
2016-06-19	11621	Sunday	2	6	2016	d_1969	NBAFinalsEnd	Sporting	Father's day	Cultural	0	0	0

1969 rows × 13 columns

	store_id	item_id	wm_yr_wk	sell_price
0	CA_1	HOBBIES_1_001	11325	9.58
1	CA_1	HOBBIES_1_001	11326	9.58
2	CA_1	HOBBIES_1_001	11327	8.26
3	CA_1	HOBBIES_1_001	11328	8.26
4	CA_1	HOBBIES_1_001	11329	8.26
...	...	...	...	...
6841116	WI_3	FOODS_3_827	11617	1.00
6841117	WI_3	FOODS_3_827	11618	1.00
6841118	WI_3	FOODS_3_827	11619	1.00
6841119	WI_3	FOODS_3_827	11620	1.00
6841120	WI_3	FOODS_3_827	11621	1.00

6841121 rows × 4 columns

Figure 2: the data format of sales_train.csv

	id	item_id	dept_id	cat_id	store_id	state_id	d_1	d_2	d_3	d_4	...	d_1904	d_1905	d_1906	d_1907	d
0	HOBBIES_1_001_CA_1_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	1	3	0	1	
1	HOBBIES_1_002_CA_1_validation	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	0	0	0	0	
2	HOBBIES_1_003_CA_1_validation	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	2	1	2	1	
3	HOBBIES_1_004_CA_1_validation	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	1	0	5	4	
4	HOBBIES_1_005_CA_1_validation	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	2	1	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
30485	FOODS_3_823_WI_3_validation	FOODS_3_823	FOODS_3	FOODS	WI_3	WI	0	0	2	2	...	2	0	0	0	
30486	FOODS_3_824_WI_3_validation	FOODS_3_824	FOODS_3	FOODS	WI_3	WI	0	0	0	0	...	0	0	0	0	
30487	FOODS_3_825_WI_3_validation	FOODS_3_825	FOODS_3	FOODS	WI_3	WI	0	6	0	2	...	2	1	0	2	
30488	FOODS_3_826_WI_3_validation	FOODS_3_826	FOODS_3	FOODS	WI_3	WI	0	0	0	0	...	0	0	1	0	
30489	FOODS_3_827_WI_3_validation	FOODS_3_827	FOODS_3	FOODS	WI_3	WI	0	0	0	0	...	0	0	0	0	

30490 rows × 1919 columns

3. Exploratory Data Analysis

First when observing sales at the scale of state, we can conclude that California generally has better sells than the other two states. And apart from foods, Texas is better than Wisconsin. The total sales of the category is: Foods > household > hobbies, as shown in Figure 3. Second, at the scale of stores, though California has the best sales, only CA_3 store has outstanding sales. The rest of California stores are just the same as other states, or even less. It is quite interesting, even in lease populated states, Walmart still manage to reach certain sales. Perhaps the location and the number of stores in the area are the real factors, as shown in Figure 4.

Figure 3: Sales of Foods & household & hobbies among three states

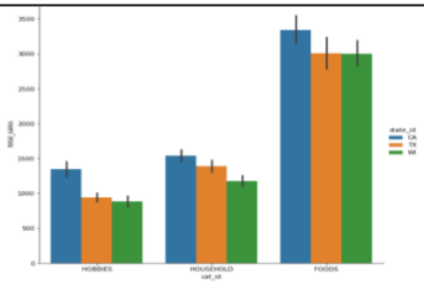
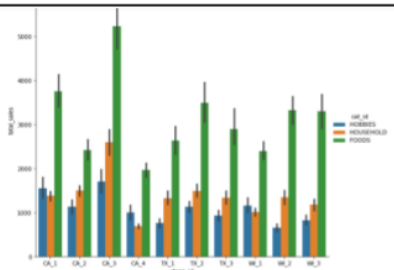


Figure 4: CA_3 store has outstanding sales



Third, in perspective of time, the time length of the following graph is from 2011 to 2016, as shown in Figure 5. It is interesting that there are yearly patterns in the sales. For example, we can see around 360 days there is a day when the sale is 0. As we see, these results are costed by annual events. Like the 0 sales days we just mentioned are cost by Christmas. In the given data sets there are a lot of annual events, some have effects on the sales and some do not. Fourth, we consider some other special events. Since special events like Christmas affects the sales in every state, perhaps there are other events that also make the sales go lower or higher nationally. In file "calendar", there are 30 different events, including Superbowl, Valentines' day, Presidents day, etc. Below shows the stores mean sales in HOBBIES of each states, and points out the special events. It is pretty obvious that there are some events always appear in the same place compare to the sales trend. For instance, there are always two points beside the Christmas points (those equal to 0). Latter there is Figure 6.

Figure 5: Yearly patterns of sales date from 2011 to 2016

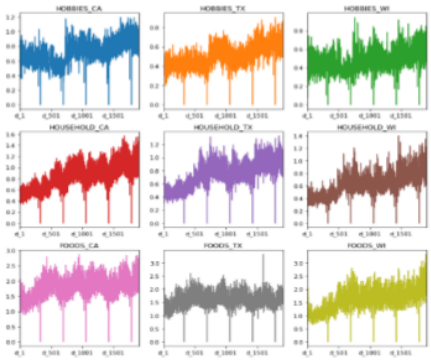
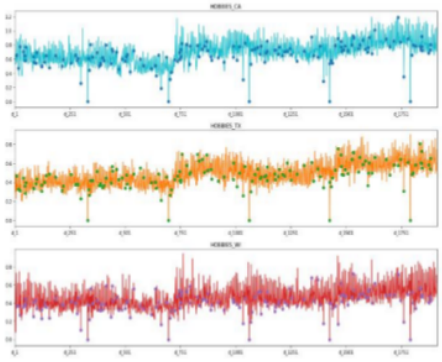


Figure 6: Special Events shown in the sales patterns.



Fifth, here we look at the sales from different perspectives in figure 7. For week, in every state and every product type, all the trends of sales are the same. Peak on Saturday, decrease till Thursday, and rise on Friday forming a deep valley. We think this is the reason why there is a dense oscillation between growth and recession of the sales through the years. For month, there is an obvious hill in the curves of "household" and "food" between May and September. That is, summer vacation starts from June to August. Then start decreasing in September, when summer vacation ends. Perhaps this is why when the sales increase every years-"S" shape trend. For year, as the economy grow in America, yearly sales in every state basically grow every year, except for year 2014. There is quite a bit of a setback in 2014. So here are a few things went on that year: "Ebola Epidemic Becomes Global Health Crisis", "Rise of ISIS", "California facing extreme drought", "World cup", "Ferguson protests". Sixth, for perspective of price, here are a few funny things in figure 8. The upper one shows the prices of hobbies_1 through the time series. The lower one shows the sales through the time series. Here are the funny stuffs: whenever there is a raise in price, there is a drop in sales. After a drop in sales, it slowly climbs back. Then Walmart raise its price again, sales drops again. As this goes, Walmart manages to get more money without losing customers in the long term. Walmart raise its price nationally.

Figure 7: various perspectives of sales data

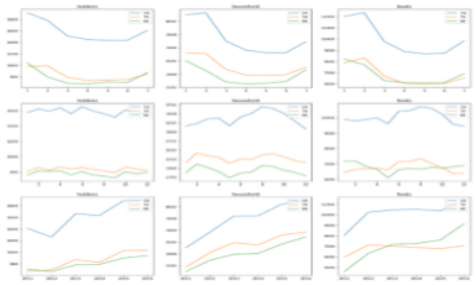
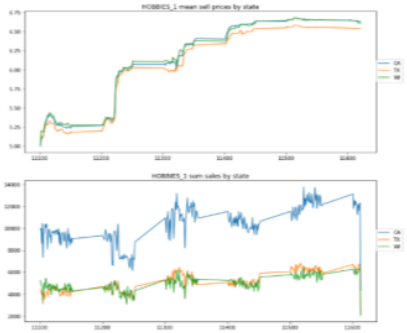
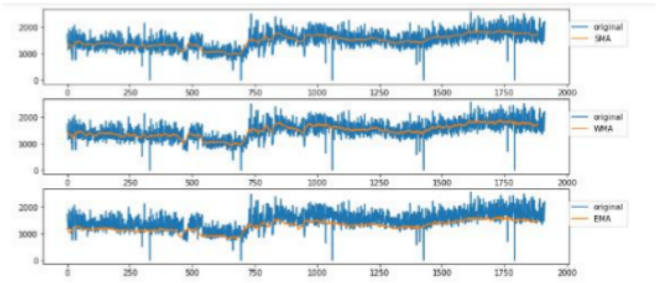


Figure 8: the price of good from 2011 to 2016.



Seventh, we find the sales with different mean. Here we use simple moving average, weighted moving average, and exponential moving average to find the tendency of sales in the close period. As you can see in the figure 9 each mean method revealed different tendency. This might come handy in the feature.

Figure 9: Each mean method revealed different tendencies



Finally, we find that about 8% of days have a special event. Among these events, about 1/3 are Religious (e.g. Orthodox Christmas) and 1/3 are National Holidays (e.g. Independence Day). The remaining third is again split into 2/3 Cultural (e.g. Valentine's Day) and 1/3 Sporting events (e.g. Super Bowl). Looking at the percentage of days where purchases with SNAP food stamps are allowed in Walmart stores, we find that it is the exact same for each of the 3 states: 650 days or 33%.

Figure 10: Days with special events

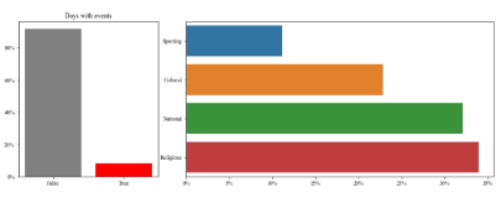
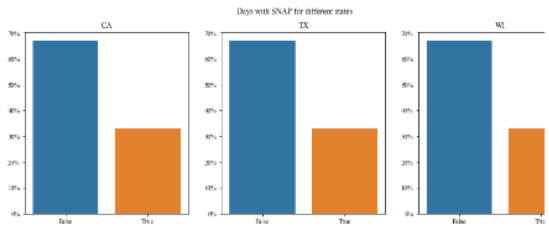
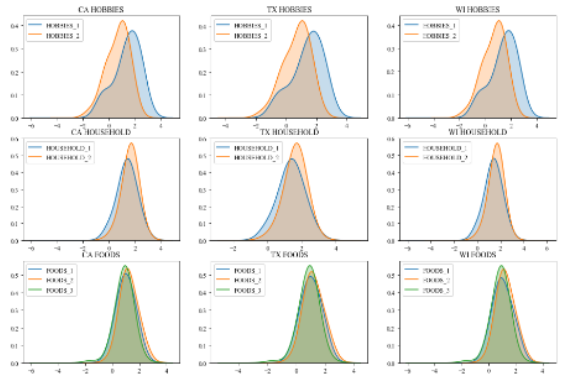


Figure 11: percentage of days purchases with SNAP food stamps.



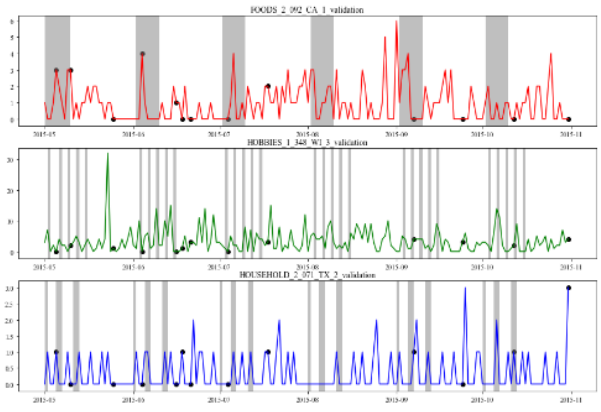
As for the distribution of item prices, they are almost identical between the 3 states. There are notable differences between the categories: FOODs are on average cheaper than HOUSEHOLD items. And HOBBIES items span a wider range of prices than the other two; even showing a second peak at lower prices. Among the three food categories department 3 (i.e. “FOODS_3”) does not contain a high-price tail. The HOBBIES category is the most diverse one, with both departments having quite broad distributions but “HOBBIES_1” accounting for almost all of the items above $10. “HOBBIES_2” has a bimodal structure.

Figure 12: The distribution of item prices



Here we pick 3 random items and then we extract their sales numbers and join calendar events together with SNAP flags. We plot the sales numbers as line charts on top of background rectangles that show the (regular) periods of SNAP days. Then we add event indicators as black points. For the FOOD item, the sales patterns are consistent with more purchases during SNAP periods. For HOBBIES and HOUSEHOLD items there are no immediate indications that SNAP days provide a particular sales boost. The impact of events is more complex. For simplicity we don’t distinguish between different types of events in this plot. We see that sometimes there are sales spikes on the day of the event (e.g. FOOOS for early Jun), while other times those spikes occur prior to an event (HOBBIES late May) or thereafter (FOODS after Jul 4th). Other combinations of events and categories appear to show no particular impact either way.

Figure 13: Sales numbers and join calendar events together with SNAP flags



4. Models

In this part, the structure we have done is as follows: firstly, choose one sample to test the five non-machine learning models, namely naive approach, moving average, Holt linear, ARIMA, exponential smoothing, with training data set, and select the optimal one to do further test. Secondly, test all samples with ES model and calculate equal weighted RMSSE. Thirdly, do all the training dataset tests with LGBM to calculate equal weighted RMSSE. Finally, compare the equal weighted RMSSE of ES and LGBM, and choose one with good results for prediction. Before doing the specific model analysis, we need to do two preprocessing: dataset classification and RMSSE definition introduction. First, we need to create miniature training and validation sets to train and validate our models. We use the last 560 days' sales, about 30%, as the validation data and the sales left, about 70% days as the training data. Below are the sales from the first sample, "HOBBIES_1_001_CA_1_validation", data points. We will use this sample to demonstrate the working of the first five non-machine learning models and compare them through the RMSSE in order to select a better model to do further test.

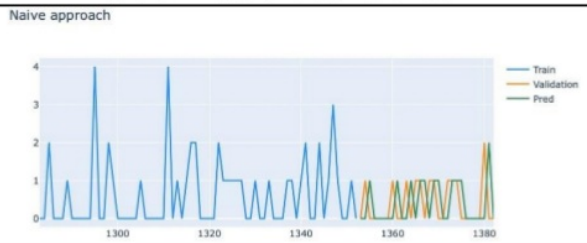
Figure 14: data set of the first sample

Then, we use RMSSE to evaluate the adaptability of models. And the calculation formula is as follows: $\mathrm{RMSSE} = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}(Y_t-\bar{Y}_t)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(Y_t-\bar{Y}_{t-1})^2}}$, where $Y_t$is the actual future value of the examined time series at point t, $Y^t$the generated forecast, n the length of the training sample(number of historical observation), and h the forecasting horizon.

4.1 Naïve Approach

The first approach is the very simple naive approach. It simply forecasts the next day's sales as the current day's sales. The model can be summarized as follows: $\widehat{Y_{t+1}} = Y_t$ In the above equation, $\widehat{Y_{t+1}}$ is the predicted value for the next day's sales and Y_t is today's sales. The model predicts tomorrow's sales as today's sales. Now let us see how this simple model performs on our miniature dataset. The training data is in blue, validation data in orange, and predictions in green.

Figure 15: The performance of Naive model



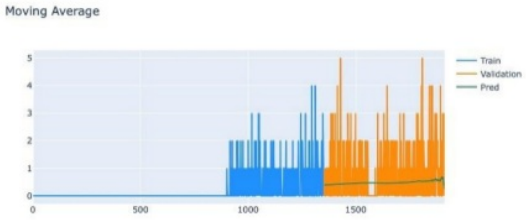
4.2 Moving Average

The moving average method is more complex than the naive approach. It calculates the mean sales over the previous k (here we use 560) days and forecasts that as the next day's sales. This method takes the previous 560 time steps into consideration, and is therefore less prone to short term fluctuations than the naive approach. The model can be summarized as follows:

$$\widehat{Y_{t+1}} = \frac{\sum_{i=1}^{k} Y_{t+1-i}}{K}$$

In the above equation, $\widehat{Y_{t+1}}$ is tomorrow's sales. On the right hand side, all the sales for the previous 560 days are added up and divided by 560 to find the average. This forms the model's prediction, $\widehat{Y_{t+1}}$. Now let us see how this new model performs on our miniature dataset. The training data is in blue, validation data in orange, and predictions in green.

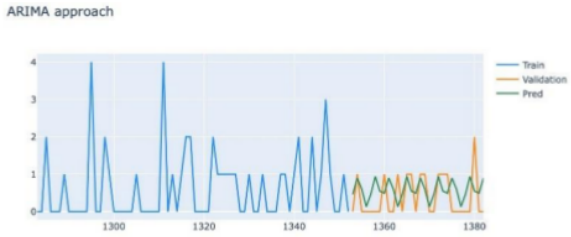
Figure 16: The performance of moving average model



4.3 ARIMA

ARIMA stands for Auto Regressive Integrated Moving Average. While exponential smoothing models were based on a description of trend and seasonality in data, ARIMA models aim to describe the correlations in the time series. Now let us see how ARIMA performs on our miniature dataset. The training data is in blue, validation data in orange, and predictions in green.

Figure 17: The performance of ARIMA model



4.4 Holt linear

The Holt linear is completely different from the first two methods. Holt linear attempts to capture the high-level trends in time series data using a linear function. The method can be summarized as follows:

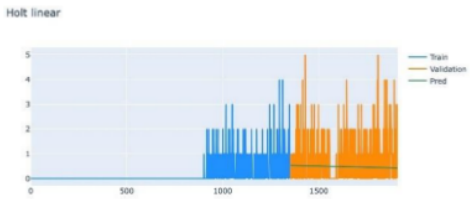
Forecast equation: $\widehat{Y_{t+h}} = L_t + hb_t$

Level equation: $L_t = \alpha Y_t + (1-\alpha)(l_{t-1} + b_{t-1})$

Trend equation: $b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$

In the above equations, α and β are constants which can be configured. The values L_t and b_t represent the level and trend values respectively. The trend value is the slope of the linear forecast function and the level value is the y-intercept of the linear forecast function. The slope and y-intercept values are continuously updated using the second and third update equations. Finally, the slope and y-intercept values are used to calculate the forecast, $\widehat{Y_{t+h}}$ (in equation 1), which is h time steps ahead of the current time step. Now let us see how this model performs on our miniature dataset. The training data is in blue, validation data in orange, and predictions in green.

Figure 18: The performance of Hot linear model



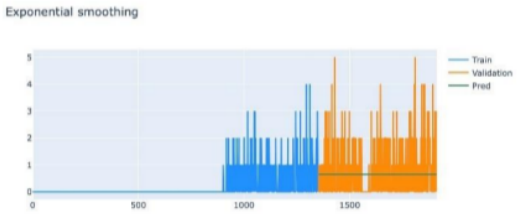
4.5 Exponential smoothing

The exponential smoothing method uses a different type of smoothing which differs from average smoothing. The previous time steps are exponentially weighted and added up to generate the forecast. The weights decay as we move further backwards in time. The model can be summarized as follows:

$$\widehat{Y_{t+1}} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \cdots + \alpha(1-\alpha)^{t-1}Y_1$$

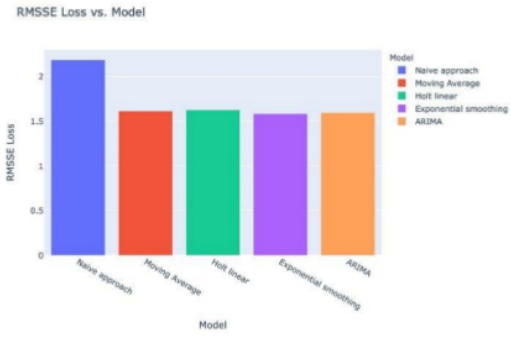
In the above equations, $\alpha\alpha$ is the smoothing parameter. The forecast $\widehat{Y_{t+1}}$ is a weighted average of all the observations in the series $Y_1 \ldots \ldots Y_t$ The rate at which the weights decay is controlled by the parameter $\alpha\alpha$. This method gives different weightage to different time steps, instead of giving the same weightage to all time steps (like the moving average method). This ensures that recent sales data is given more importance than old sales data while making the forecast. Now let us see how this new smoothing method performs on our miniature dataset. The training data is in blue, validation data in orange, and predictions in green.

Figure 19: The performance of Exponential smoothing model



Next, the Picture 20 shows the RMSSE loss comparison among these five models. It can be seen that the experimental smoothing model is the best one. Therefore, we select the exponential smoothing model to do the estimation for the whole 30490 samples.

Figure 20: RMSSE loss of the five models



Finally, we calculate the equal weight RMSSE of the whole sample under exponential smoothing, the result is about 1.3648. Then we would compare it with the one under LGBM, choose the better one to do further 28 days' forecasting.

4.6 LGBM

LGBM is a gradient boosting framework, which uses decision tree based on learning algorithm. And LGBM has the following features: decision tree algorithm based on histogram, leaf wise growth strategy with depth limitation, histogram difference acceleration, directly support category feature, cache hit rate optimization, histogram based sparse feature optimization and multithreading optimization. The features we used in LGBM are item ID, department ID, store ID, category ID, state ID, weekday, month, year, event name, event type, week, and quarter. Besides, via the EDA process, we spotted a strong weekly and monthly pattern. Thus, the 7 days' lag and mean in sales, as well as 28 days' lag and mean in sales are extracted as additional features for the LGBM mode, denoted as lag_7, lag_28, rmean_7_7, rmean_28_7, rmean_7_28, rmean_28_28.

Figure 21: monthly and weekly pattern of sales data

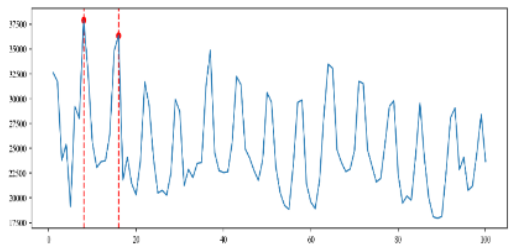


Figure 22: time series data of sales among three states

Under LGBM, we calculate the equal weight RMSSE of the whole sample is about 1.0207. Obviously, it is smaller than 1.3648, which is the one under exponential smoothing model. Therefore, we would choose LGBM to do further 28 days' forecasting next.

5 Forecasting and Conclusion

We choose the LGBM model for the final forecasting task. The training period lasts from day1 to day1885, with the last 28 days left as validation/test set, in the lack of real out-of-sample data. The evaluation method for point forecasts is chosen to be the RMSSE, which is defined as follows. The h is the prediction horizon and n is the number of items. The out of sample prediction RMSSE for the LGBM model is 1.922313

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

Randomly selected 2 prediction results are shown in figure 23. The overall prediction result is shown in figure 24.

And the forecasting result are showing in the figure 25, we did a 28 days ahead forecast using the best model with the lowest RMSSE, which is LGBM.

Figure 23: Prediction results for 2 randomly selected items

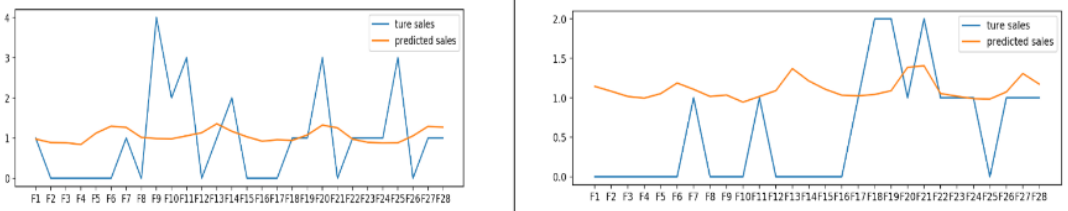


Figure 24: Overall prediction results for all items

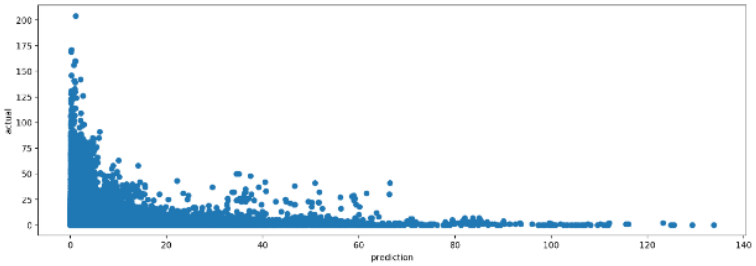
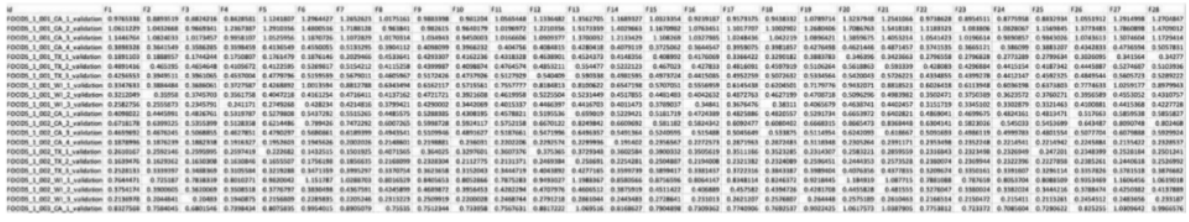


Figure 25: Forecasting result of LGBM model (28 days ahead)



This article focuses on data processing, model building and prediction for M5 competition requirements. We first analyzed the data of the selected subject, observed the characteristics of the data from different regions and different product types, and selected the feature points of the time dimension such as week, month and year, the feature points of special festivals, and price trends From various angles, preliminary feature extraction and analysis of the data were carried out. After conducting preliminary data exploratory analysis, we selected multiple models for fitting based on the data characteristics of the project. On the traditional time series model, we selected five models for fitting, including NAIVE, MA, HOT linear, Exponential Smoothing, and ARIMA; After dividing the training set and test set, we choose the exponential smoothing model that performs best in the traditional model according to the loss function. At the same time, we also tried a new model approach, LGBM. Using the characteristics of the LBGM model to select the appropriate features to fit the model, the fitting results found that the LGBM model is significantly better than other models. Finally, we used the LGBM model to achieve a 28-day forecast of the data. The overall RMSSE of the model is only 1.922. The model has a good effect and successfully achieved an effective forecast of sales data.

Presentation with video for this report can be found at website:

https://www.bilibili.com/video/BV1ie411W7J3

Codes & output results can be found at website:

https://drive.google.com/drive/folders/1h1x8mOA_5FlZlDzvzcytH8ZUAHHeGub9?usp=sharing

AI project WeArealLsMART.pdf

ORIGINALITY REPORT

11%	5%	2%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Sydney Student Paper	2%
2	mofc.unic.ac.cy Internet Source	2%
3	Submitted to The University of Manchester Student Paper	2%
4	Bo Liu, Mengmeng Huang, Kelu Yao, Lan Wei, Xiaolu Fei, Wang Qing. "Prediction and Study of the Applicability of Medical Gels to Patients", 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2019 Publication	1%
5	www.analyticsvidhya.com Internet Source	1%
6	mltrainings.ru Internet Source	1%
7	Submitted to University of Edinburgh Student Paper	1%
	Submitted to City University of Hong Kong	

8	Student Paper	<1%
9	Submitted to National College of Ireland Student Paper	<1%
10	Submitted to University of Hertfordshire Student Paper	<1%
11	Submitted to University of Rome Tor Vergata Student Paper	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		