Project 3: M5 Forecasting - Accuracy
Estimate the unit sales of Walmart retail goods

Team Member: Fu Qiyin, Liu Enjie, Yu Xintong, Zhao Encong {qfuab, ezhaoab, eliuac,xyubi} @connect.ust.hk
Kaggle ID: math6010z_Fu_Yu_Liu_Zhao

1. Introduction

The objective of this project is to estimate the unit sales of Walmart retail goods based on hierarchical time series data. To better understand this complicated project, we first did **exploratory data analysis** from various perspectives such as data structure, overall trend, time rule and seasonal trends. Then we used **LSTM and LGBM** algorithm to build our model and see their performance. Then we analyzed 3 types of important features and see how they affect the sales data.

2. Exploratory data analysis

2.1 Data structure: The training set comes from 10 Walmart stores in 3 American states (CA, TX and WI) between 2011.01.29 and 2016.05.22, which includes the sales volume and commodity prices of 3049 goods about food, household and personal hobbies. There are lots of 0 value in the sales data , which means a majority of goods has no sells on the vast majority of days.

2.2 The overall trend & Time pattern: Figure 1: Overall sales are rising every year, showing a strong seasonality and periodicity. At the end of each year, there is a sudden drop in sales, which we suspect is caused by Christmas holiday. **Figure 2:** As we can see from the color blocks, people prefer to go shopping on weekends and sales are significantly lower in Nov. and Dec. in winter, as well as Jun. and Jul. in summer.

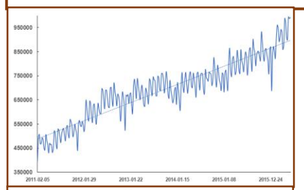


Figure 1: All aggregate sales

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Sat	997833	1041717	1136936	1095021	1135555	902349	826864	984441	905580	936699	880504	867084
Sun	929319	1012259	1152665	1040798	1136582	905140	884666	963438	952934	889029	904110	818722
Mon	699629	823590	845361	830608	868762	754775	733643	748332	765193	718843	666887	753065
Tue	628697	722176	807303	796957	822896	661240	739307	658649	688428	673914	624137	689964
Wed	629219	749709	763004	782400	832704	637255	736361	682468	640696	691833	633310	652671
Thu	675517	734972	806011	752817	868768	643164	722279	698013	639712	700701	578785	661639
Fri	785297	805166	875930	899192	992190	746289	800855	836389	729346	769570	705891	668827

Figure2: Sales in different day of week and month of year

2.3 Seasonal breakdown: Figure 3: We adjusted the sales data seasonally: first extract the trend and then used HP filtering to obtain the cycle. The year-on-year change is calculated to reflect a more realistic sales change.

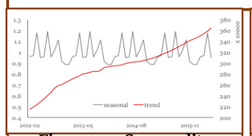


Figure 3.1: Seasonality adjustment

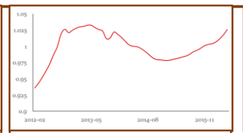


Figure 3.2: Cycle

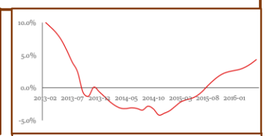


Figure 3.3: Growth YOY

2.4 Some comparisons among states and categories: Figure 4: People in CA are more prosperously and they buy more goods. Sales in WI and TX are similar; **Figure 5:** The top selling category is food, followed by household. Both two are well above hobbies.

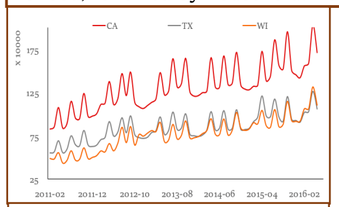


Figure 4: Monthly sales per state

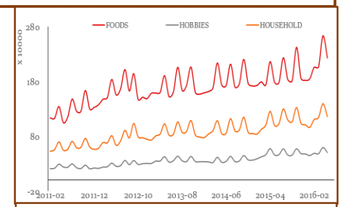


Figure 5: Sales per category

3. Data preprocessing & Feature engineering

3.1 Data preprocessing: (1) Re-code some nominal variables through one-hot encoding; (2) **Data memory reduction**: Consider the computer memory and operation efficiency limit, we transfer some datatype to reduce its size. For example, change int from 64 ints to 32 ints to save 50% memory.

3.2 Feature engineering: (1) Data standardization and 0-1 normalization; (2) **Create new features** based on time series data: We create 6 more features based on some caluation in time dimension. i. "Price_change_week" : Measure price change by dividing the previous week's price by the current week's price; ii. “Rolling_sales_mean_7”: Use sales data in a week to calculate moving average value. There are only two examples due to space limitation, and other feature calculation principles are similar as above.

4. Model selection

4.1 LGBM: We found that LGBM algorithm may be a good choice in this forecasting questions for it can grasp the nonlinear input/output data relationship, and can well **fit the fluctuation term**. In addition, GBDT can grasp the period term because this algorithm always find the optimal split point of features.

4.2 LSTM: We also use LSTM (Long short-term memory) to build our model because it is very effective for sequential data and can mine **temporal information**.

5. Model performance & Feature importance

5.1 Model performance: To judge our model performance, we first calculate the Rmse in our training data and submit our forecasting result in Kaggle website to see the score in testing data. From the results listed in the right table, we can see LSTM performed better than LGBM in both the training set and Kaggle testing set.

Model	Training Rmse	Kaggle test score
LSTM	0.103	0.83
LGBM	1.735	4.99

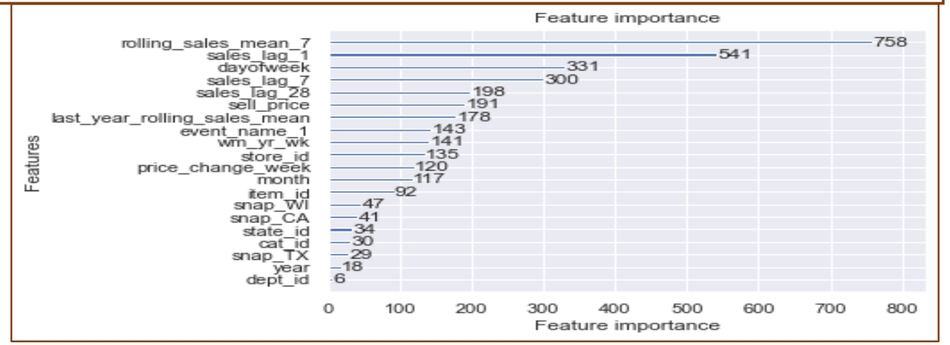
5.2 Feature importance:

① **Recent past data**: It can be seen from the figure that the average value of last week, the sales lag of 1,7, and 28 days all play an important role. Combined with the overall increasing trend mentioned in the exploratory data analysis, we can infer that the sales data in the recent past has a great influence, while the data with a long time distance does not.

② **Weekdays, weekends, or holidays**: The day of the week and some special events can have a big impact on people's shopping decisions.

③ **Individual heterogeneity**: We can see that the sales volume of different specific product, different stores and even different states vary greatly, these characteristics of individual attributes also play a big role

Figure 6 : Feature Importance



6. Conclusion

In this project, we analyzed Walmart retail goods hierarchical time series data and used machine learning method to do forecasting. Considering the characteristics time series data, we use LGBM and LSTM. For model performance, **LSTM** have a better Kaggle score of 0.83, compared with 4.99 of **LGBM**. And **3 types features** (recent past feature, holidays or not, individual heterogeneity) are the most influential for forecast accuracy.

7. Contribution

Data process & Modeling: Fu Qiyin, Yu Xintong **Poster:** Liu Enjie **Data Analysis & PPT**: Zhao Encong