M5 Forecasting - Uncertainty

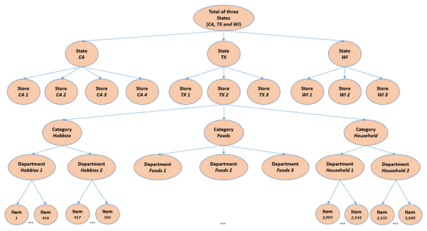
LU Fei 20797386, XIONG Yi 20787642, ZHENG Hao 20797594

Abstract

If a store can predict the future sales of products, it will be able to plan its business operation more rationally and make greater profits. In the M5 forecasting project, we use Walmart store sales data to estimate the uncertain distribution of the unit sales of various products. We first preprocess the data and observe the sales characteristics. Then, three models are utilized and compared, and the corresponding results are given and analyzed. The best score of our models is 0.11617 on the public leaderboard and 0.20020 on the private leaderboard.

1 Problem Definition

Time series prediction is a popular topic in scientific research and the real world. The prediction usually relies on historical data and current features. In this project, detailed sales data in 1914 days from Walmart are given, and the goal is to do the uncertainty estimation of the sales for the next 28 days. Uncertainty estimation means that we should estimate the uncertainty distribution of the sales instead of the exact sales. According to the M5 participant's guide, we are required to provide the median, and the 50%, 67%, 95%, and 99% prediction intervals, so the corresponding quantiles are 0.005, 0.025, 0.165, 0.25, 0.5, 0.75, 0.835, 0.975, and 0.995. Besides, we should not only predict the sales of each product in every store but also predict on the different hierarchical levels. According to the M5 participant's guide, the dataset involves the unit sales of 3,049 products, classified in 3 product categories (Hobbies, Foods, and Household) and 7 product departments, in which the above-mentioned categories are disaggregated. The products are sold across ten stores, located in three States (CA, TX, and WI). In this respect, the bottom-level of the hierarchy, i.e., product-store unit sales can be mapped across either product categories or geographical regions, as shown in 1:



Level id	Aggregation Level	Number of series
1	Unit sales of all products, aggregated for all stores/states	1
2	Unit sales of all products, aggregated for each State	3
3	Unit sales of all products, aggregated for each store	10
4	Unit sales of all products, aggregated for each category	3
5	Unit sales of all products, aggregated for each department	7
6	Unit sales of all products, aggregated for each State and category	9
7	Unit sales of all products, aggregated for each State and department	21
8	Unit sales of all products, aggregated for each store and category	30
9	Unit sales of all products, aggregated for each store and department	70
10	Unit sales of product x, aggregated for all stores/states	3,049
11	Unit sales of product x, aggregated for each State	9,147
12	Unit sales of product x, aggregated for each store	30,490
	Total	**42,840**

Figure 1: hierarchy of the uncertainty problem

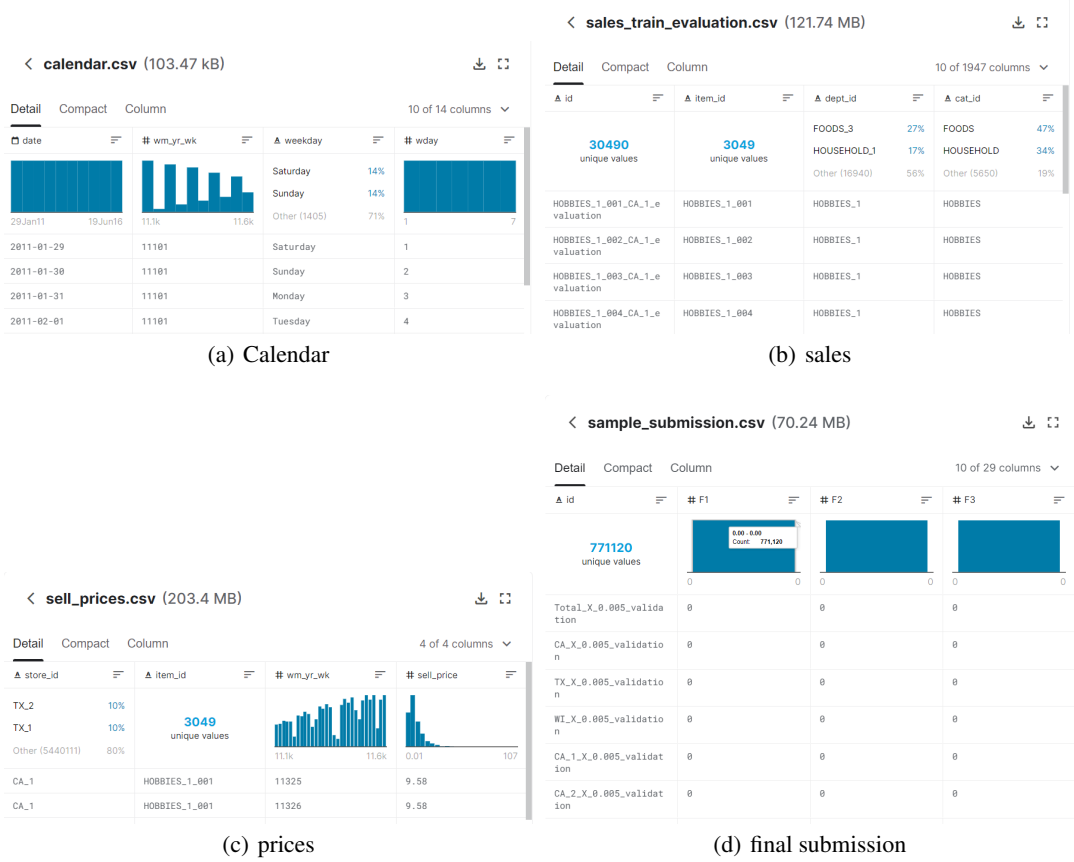
2 Data Analysis, Preprocessing and Feature Engineering

The data used in this project are sales data from the real world. In order to find the most suitable algorithm for the later prediction, data engineering is of great importance. Firstly, we examine the data set. Then, some data visualization and preprocessing works are performed to see the sales characteristics under different hierarchies. Finally, we engineer the features for the later model training.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

2.1 Dataset Description

The data set is provided by Walmart. It covers sales of stores in three US States, i.e. California, Texas, and Wisconsin, and includes the item, department, product categories, and store details. In addition, it has explanatory variables such as price, promotions, day of the week, and special events. Also, it provides the calendar data to make it possible for us to join the sell prices table and sales_train_evaluation tables together with the attributes "wm_yr_wk" and "d_i". Here's the outline of the data 2. The sales_train_evaluation contains 30490 unique item values and for each item, we can get the sales data from d_1 to d_1941, which is from 2011-01-29 to 2016-05-22. The goal is to predict the following 28 days' sales.



(a) Calendar (b) sales

(c) prices (d) final submission

Figure 2: Dataset outline

2.2 Data Analysis

Since the more detailed exploratory data analysis of this data set is thoroughly described in our mid-project report, Accuracy for M5 prediction, here we only present the sales trend under different hierarchies.

As is shown in 2, the sales data presents a strong periodic feature. At state level, sales in CA is the highest while other states' sales trends are similar. At store level, CA_3 has the highest sales. At category level, food category has the highest sales. And as for the department level, FOOD_3 is the most outstanding one. Besides, the sales show a similar trend in different aggregation level, and it shows a sudden drop at the end of a year, since the stores close during national holidays.

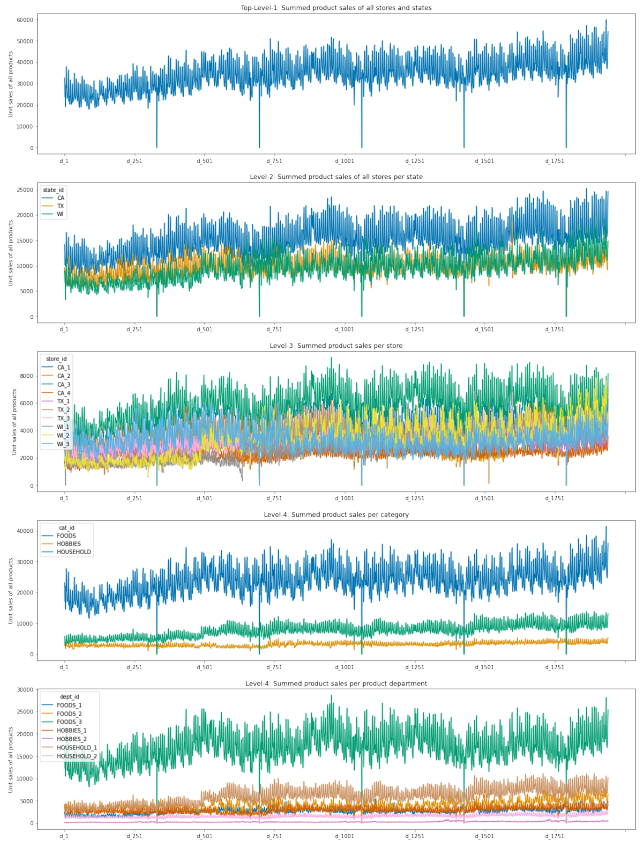


Figure 3: Sales trend under different hierarchies

2.3 Data Preprocessing

We preprocess the original data from 3 perspectives: Firstly, we reduce memory usage. Then, since the final estimation is based on different hierarchies, we should preprocess the original table and aggregate the sales by different levels. Finally, some missing values are replaced.

Memory usage reduction: In this project, we change the data type of every numerical data, nearly half of memory are reduced.

Hierarchy aggregation: According to the M5 participant's guide, there are 12 hierarchies 1 in the final prediction. The original data should be preprocessed to the same schema as the final submission file for the later work. The submission file requires us to get 42840 rows for each quantile, and each row stands for one hierarchy level, containing the total sales, state sales, store sales, category sales, department sales, item sales, and their combinations. Therefore, we use the group by function to aggregate the values at different hierarchies for training. For each quantile, we train a model and predict corresponding values, then all of the predictions are concatenated to the final submission.

Missing values: As we analyzed in the analysis part, there are some missing values since the stores closed during some special days or for some specific items not-selling. Especially in the national holidays like Thanksgiving or Christmas, almost all the items are not sold. In order to reduce the impact of these missing values during national holidays, we calculate the average of the sales in the day before and the day after these days, and the mean values are treated as the sales.

2.4 Feature engineering

Categorical data: In the original data, there are some categorical attributes in the calendar table. For example, the event and SNAP related attributes. The Ordinal Encoder function is utilized to help with

the categorical data encoding.
Feature scaling: Since in different aggregation level, the sum sales are in different scale. Feature scaling is of great importance to fasten the training process during the gradient descent process. The MinmaxScaler function is utilized to do the scaling.

3 Methodologies and Models

In this section, the methodologies and models for uncertainty prediction are introduced. The models for accurate prediction are thoroughly described in the mid-project report [2], and we are not going to repeat them here.

3.1 Methodology

3.1.1 Quantile Regression

Regression analysis is often used when encountering prediction problems, including linear regression, polynomial regression and so on. The basic idea is to find a function to fit the training data as precise as possible, and then use the function to do some predictions. The parameters adjusting process is usually realized by minimizing Mean Squared Error as the loss function.

All the regression functions mentioned above are essentially a conditional expectation function. Under the condition that x is equal to a certain value, find the expectation of Y according to the data. But more often, we don't want to just study the expectation of Y, but also to explore the complete distribution of Y, or maybe in some cases, we prefer to know a quantile of Y, so quantile regression[3] occurs. For example, in the figure4, the distribution range of Y increases with the increase of X. If classical regression is carried out, the function can not reflect the change of Y range. If the 0.9 quantile regression is carried out so that 90% of the points are below the function line, the change range of Y is reflected. If we further draw different quantile regression curves, it can more clearly reflect the distribution of y with different X. This is a conclusion that the other regression analysis cannot obtain.

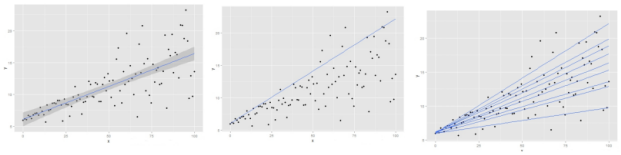


Figure 4: example of quantile regression

In the accuracy competition, we predict the exact value of the unit sales of various products sold in the USA by Walmart in the future 28 days. In the uncertainty competition, instead of single point estimation, the goal is to predict the intervals, which can be also seen as the uncertainty in our single point predictions. To solve this problem, quantile regression method is utilized. The quantile loss differs depending on different quantile, for higher quantiles, more negative errors are penalized and for lower quantiles, more positive errors are penalized. And the quantile loss function is:

$$\begin{cases} e = Value_real - Value_predict \\ quantileloss = \max(q \cdot e, (q-1) \cdot e) \end{cases} \tag{1}$$

According to the M5 participants' guide, we are asked to provide the median, and the 50%, 67%, 95%, and 99% prediction intervals, so the quantiles are: 0.005, 0.025, 0.165, 0.25, 0.5, 0.75, 0.835, 0.975, and 0.995.

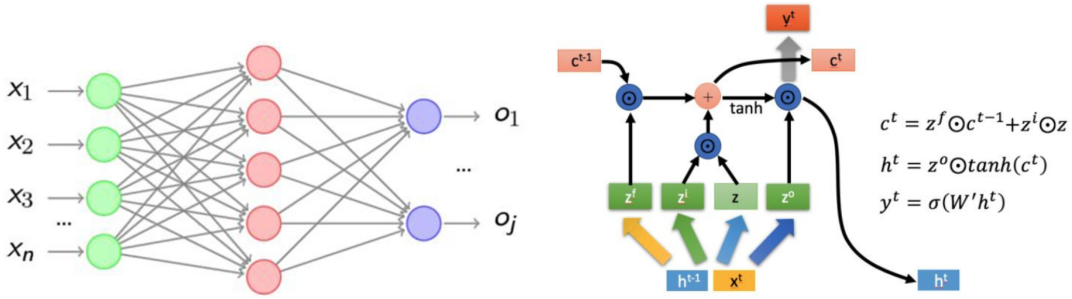
3.1.2 NNs and LSTM

Artificial neural network is an algorithmic mathematical model that imitates the behavior characteristics of animal neural network and carries out distributed parallel information processing. This kind of

network depends on the complexity of the system, and achieves the purpose of processing information by adjusting the interconnected relationship between a large number of internal nodes5(a).

When it comes to data with time series properties, we sometimes choose to use LSTM model. LSTM is a special RNN, which is mainly used to solve the problems of gradient disappearance and gradient explosion in the process of long sequence training. In short, compared with ordinary RNN, LSTM can perform better in longer sequences. The specific structure is as figure5(b).

The specific principle steps of LSTM are as follows: First, the input of time step T will enter the forgetting gate to decide what information to discard from the previous cell state. Then, the output of the forgetting gate is introduced into the input gate to determine which new information will be stored in the cell state. Next, update the cell state, update the cell state, and forget and update the cell at the same time. Finally, in the output gate, which part of the cell state will be output as a result is determined by the activation function based on the current cell state.



(a) NNs (b) LSTM

Figure 5: model architecture

3.2 Models

3.2.1 From point to uncertainty estimation

In the accuracy competition, we used LGBM to estimate the exact sales. In the uncertainty competition, the previous estimated sales in the last 28 days (from d_1914 to d_1941) could be utilized to do the range prediction. Actually, the 0.5 quantile is exactly what we had estimated in the accuracy competition. We assume it is a normal distribution with the 0.5 quantile as the mean value and the other 9 quantiles can be esitimated. Scipy toolkit is utilized to calculate the corresponding quantiles. [1]

3.2.2 NNs with Quantile Regression

In addition to obtaining the uncertainty task directly from the accuracy task, since the data set contains many features, we also consider directly using the basic neural network to predict the future sales situation and make the uncertainty estimates.[4]

For quantitative regression, we set a special loss function, in which the quantiles are 0.005, 0.025, 0.165, 0.250, 0.500, 0.750, 0.835, 0.975 and 0.995 respectively.

In the neural network model, because the dataset is large, we embed the data first. Then, the features of the dataset are concatenated and compressed into one-dimensional data. The main part of the model includes four Dense layers, in which the activation function of the first three layers is Relu and the activation function of the last layer is Linear. Because there are a lot of data, we added Dropout layer between the Dense layer. The optimization function of the model is Adam. 6

We used file sales_train_evaluation.csv for training and cross validation. We input the preprocessed training data into the model for training. Among them, each line of training data includes sales data information under a specific hierarchy, including store commodity information, date information, sales, etc. Then, for different quantiles, we use the trained model to predict respectively and save it to the file submission.csv.

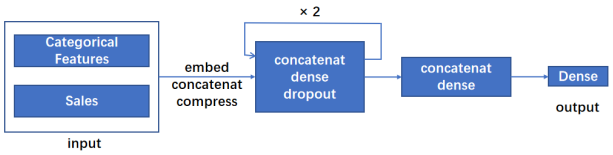


Figure 6: NNs model structure

3.2.3 LSTM with Quantile Regression

Since the goal is to predict the sales in the future 28 days, which are time-series data, LSTM model could be utilized to deal with this kind of task.[5] The input of the model is the sales in 14 days, and the output is the sales of the next one day. We are given data from d_1 to d_1913 and the goal is to predict the sales in d_1914 to d_1941. Thus, the training set contains 1899 samples. And for each training sample, it contains 14 days sales. Besides, the other features such as the calendar-related attributes are also utilized. The neural network is constructed as 7:

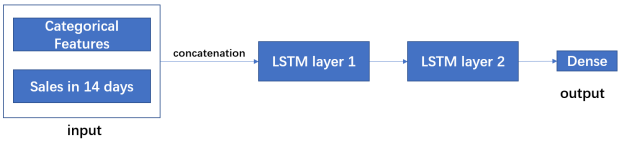


Figure 7: LSTM model structure

4 Results and submission

This section is to present the result of our models and have some discussion about them. Since the details about the accuracy problem's solution is on our mid-project report, here we simply present the result of accuracy problem to make comparison.

4.1 Result of M5 Forecasting - accuracy

The evaluation is based on the Weighted Root Mean Squared Scaled Error (RMSSE) between the true value and prediction value. We used the "sales_train_evaluation.csv" data which includes the sales from d_1 to d_1941 to train our LGBM model, and predicted the sales on d_1942 - d_1969. According to the scoring system of Kaggle, our score is 0.62812, and ranked nearly 205th of more than 5500 groups on the private leader board.

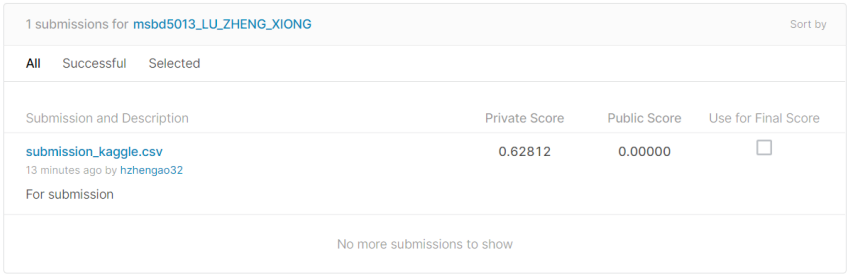


Figure 8: M5 Forecasting - accuracy score

4.2 Result of M5 Forecasting - uncertainty

In the uncertainty competetion, Weighted Scaled Pinball Loss (WSPL) is utilized for the evaluation. Our results are shown in table1, in which we found that the best results were obtained by using NNs with Quantile Regression. Our team name is "msbd5013_LU_ZHENG_XIONG".

Methods	Public score	Private score
LGBM model from accuracy to uncertainty	0.05561	0.30780
NNs with Quantile Regression	0.11617	0.20020
LSTM with Quantile Regression	0.51769	0.49638

Table 1: Final Results

4.3 Result Analysis

4.3.1 from accuracy to uncertainty

Since in the accuracy competition, we got good grades by performing LGBM to do the estimation, the uncertainty estimation from accuracy got not bad results too. However, we can not know the actual variance of the distribution. This kind of method can be a little tricky, since we test different parameters to get the best performance.

4.3.2 NNs with Quantile Regression

The neural network is fed with a large amount of data with multiple features and time series characteristics. Besides, we added dropout layers in the network structure to avoid over fitting. It effectively predict the future sales after iterative training. And it shows the best results among all of the methods.

4.3.3 LSTM with Quantile Regression

The model with LSTM layers show a bad result, we did some analysis as followed:
The loss of periodic feature. Since the input of the model is the sales of 14 consecutive days, which the sales in 15th day estimated is based on. However, although we had taken the calendar related features into account, but it still loses some periodic characteristics, for example, the sales at the end of the year are always higher than other months.
Using our prediction sales to do the further predicting. In this model, the sales of the former 14 days are utilized to predict the sales of the next day, which means the error could be accumulated.
Model construction problem. In this model, the input is the concatenation of categorical features and sales, and the model is constructed with two LSTM layers. It seems some more adjustment of the structure and parameters should be applied to avoid over-fitting problems.

5 Further improvements

More feature engineering should be applied. In the future work, we should explore more about the features provided, maybe using better categorical attribute encoder or combining different attributes.
Neural network structure adjustment. In the future work, more adjustments of parameters and neural network structure should be applied to gain a better performance.

6 Contributions

Name	Contribution Descript
LU Fei	Data Preprocessing; Model training and prediction; Results Analysis; Report writing
ZHENG Hao	Problem Analysis; Exploratory Data Analysis; Report writing
XIONG Yi	Data analysis and visualization; Model training and prediction; Report writing

References

[1] *From Point to Uncertainty*. URL: `https://www.kaggle.com/code/kneroma/from-point-to-uncertainty-prediction`.

[2] *Midterm-Project Report*. URL: https://github.com/yao-lab/yao-lab.github.io/blob/master/course/msbd5013/2022/project1/group4_LU_ZHENG_XIONG/report.pdf.

[3] *Quantile Regression*. URL: https://towardsdatascience.com/quantile-regression-from-linear-models-to-trees-to-deep-learning-af3738b527c3.

[4] *Quantile Regression with Basic Neural Network*. URL: https://www.kaggle.com/code/ulrich07/quantile-regression-with-keras/notebook.

[5] *Quantile Regression with LSTM*. URL: https://www.kaggle.com/code/olafko/m5-uncertainty-prediction-lstm-nn-feature-embedd#LSTM-Modeling-.