Demand Forecasting: From Machine Learning to Ensemble Learning

Yang Zhang*
Department of Science
Shanghai University
Shanghai, China
remlin08042@shu.edu.cn

Hongyi Zhu
Department of Traffic and Transportation
Beijing Jiaotong University
Beijing, China
19311076@bjtu.edu.cn

Yujing Wang
Department of Public Administration
Guangdong University of Foreign Studies
Guangzhou, China
1067936051@qq.com

Tianchen Li
Department of Economy and Management
Nanjing Agriculture University
Nanjing, China
laurali0810@163.com

2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS) | 978-1-6654-7053-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/TOCS56154.2022.10015992

***Abstract*—Across all industries and supply chains, the demand from the market is always of great concern for decision-makers. This research focused on solving the problem of forecasting demand for various products. To express demand, the project took sales as the forecasting target according to the supply and demand theory. We worked with the retail data of Walmart over five years and aimed to forecast the total weekly sales of different departments in different states. We first used four different time series models to forecast demand, including Naïve, Moving average, Prophet, and Exponential Smoothing (ETS). Then, we chose "Stacking" as the Ensemble method to optimize the forecast outcomes and tried three models: Linear Regression, Simple Average, and Weighted Average. The performances on test sets showed that: Firstly, Ensemble Learning made sense in forecasting. Secondly, as stacking models, the Weighted Average using Random Forest had a strong generalization ability, while the Linear Regression method was unstable and easily overfitted. Apart from what we've got from Ensemble Learning, we combined sales with other factors to analyze the demand forecasting and gave enterprises some advice to use models and solve supply chain problems.**

Keywords—supply chain, demand forecast, Machine Learning, Ensemble Learning

I. Introduction

With the rise of big data and the Internet technology rapidly merging into various fields, it has made new changes in the structure of the supply chain and reclaimed some new directions for the development of the supply chain. Supply chain management (SCM) focuses on the flow of goods, services, and information from points of origin to customers through a chain of entities and activities that are connected. In typical SCM problems, it is assumed that capacity, demand, and cost are known parameters. However, this is not the case in reality, as there are uncertainties arising from variations in customers' demand, supplies transportation, organizational risks, and lead times. Demand uncertainties, in particular, have the greatest influence on supply chain performance with widespread effects on production scheduling, inventory planning, and transportation. Therefore, demand forecasting is an essential part of the whole supply chain.

Data of supply chains can be categorized into customers, shipping, delivery, order, sale, store, and product data.

As such, supply chain data originates from different sources such as sales, inventory, manufacturing, warehouse, and transportation. In this sense, competition, price volatilities, technological development, and varying customer commitments could lead to underestimation or overestimation of demand in established forecasts. However, traditional forecasting methods have been unable to obtain satisfactory prediction accuracy under the background of big data, which has brought great challenges to the SCM. As a result, more and more researchers start to focus on the efficiency of different forecasting models.

In this paper, we put all the sell-in and sell-out data along with relevant demand casuals together to gain the most complete, joined-up picture of demand possibility. This also provides the foundation for highly demand forecasting that frees up planners to apply the business knowledge to further improve forecasting and customer service.

II. Literature Review

Demand forecasting models have been widely applied in precision marketing to understand and fulfill customer needs and expectations[1]. The characteristics of demand data in today's ever-expanding and sporadic global supply chains make the adoption of big data analytic methods (including machine learning) a necessity for demand forecasting. The digitization of supply chains[2] and the incorporation of Blockchain technologies[3] for better tracking of supply chains further highlight the role of big data analyses. Supply chain data is high dimensional generated across many points in the chain for varied purposes (products, supplier capacities, orders, shipments, customers, retailers, etc.) in high volumes due to plurality of suppliers, products, and customers and in high velocity reflected by many transactions continuously processed across supply chain networks. In the sense of such complexities, there has been a departure from conventional (statistical) demand forecasting approaches that work based on identifying statistically meaningful trends (characterized by mean and variance attributes) across historical data[4] towards intelligent forecasts that can learn from the historical data and intelligently evolve to adjust to predict the ever-changing demand in supply chains[5].

Uncertain point forecasts of product sales exert a vital influence on supply chain management. To identify ways to improve the accuracy of forecasting, the M competitions empirically evaluate several forecasting methods as well as identify the most accurate.

The findings obtained in M competitions have significantly influenced the theory and practice of forecasting by providing valuable insights into how forecasting accuracy

can be improved[6]. The M5 competition extended the objectives of the previous four competitions by focusing on retail sales forecasting applications and using real-life, hierarchically structured sales data with intermittent and erratic characteristics[7][8].

Issues remain to carry on further research in this field. In this paper, the first aim is related to understanding how Machine Learning methods produce their forecasts on sales. The second issue concerns how Machine Learning can be improved to reduce the inaccuracy in forecasting. As for the third, Machine Learning is always suitable for concrete values like sales instead of uncertain variables in demand forecasting, and how we can combine sales with demand forecasting.

III. Model Building

A. Data understanding

The data we used in our project all comes from the M5 competition. In datasets, there was a Calendar that recorded dates and events of nearly 5 years (from 2011.2-2016.6), a Sell_Price with prices of a certain product in the whole period, and a Sales_Train_Validation that told sales of a product every day[9].

To the explanation of the M5 Official, the framework and aggregation levels of datasets are shown in Fig.1 and Fig.2.

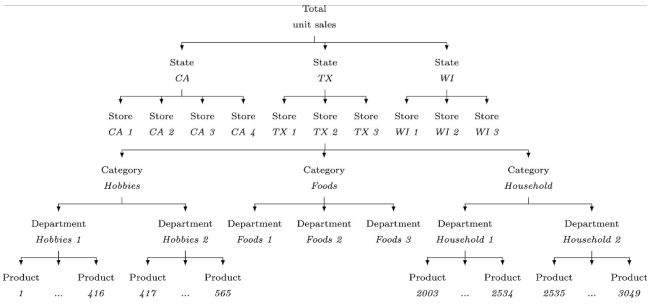

Fig. 1. The framework of datasets

Level id	Level description	Aggregation level	Number of series
1	Unit sales of all products, aggregated for all stores/states	Total	1
2	Unit sales of all products, aggregated for each state	State	3
3	Unit sales of all products, aggregated for each store	Store	10
4	Unit sales of all products, aggregated for each category	Category	3
5	Unit sales of all products, aggregated for each department	Department	7
6	Unit sales of all products, aggregated for each state and category	State-category	9
7	Unit sales of all products, aggregated for each state and department	State-department	21
8	Unit sales of all products, aggregated for each store and category	Store-category	30
9	Unit sales of all products, aggregated for each store and department	Store-department	70
10	Unit sales of product i, aggregated for all stores/states	Product	3,049
11	Unit sales of product i, aggregated for each state	Product–state	9,147
12	Unit sales of product i, aggregated for each store	Product–store	30,490
Total			42,840

Fig. 2. Aggregation levels of datasets

We first chose Level 10 to start our work and did our exploratory data analysis. As an example, we took sales of a certain product, HOBBIES_1001 in the whole country in 2014. After the STL decomposition, the decomposition results are shown in Fig.3.

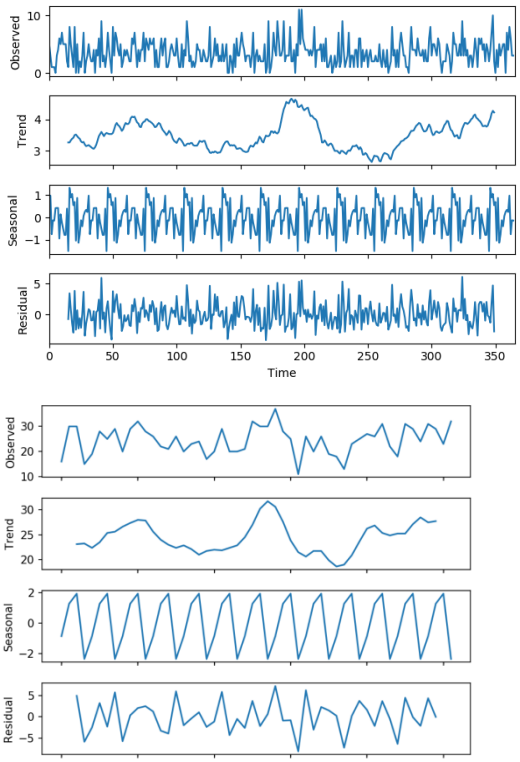

Fig. 3. Decomposition of sales of HOBBIES_1_001 in 2014 per day and per week.

The figure on the top represented sales of HOBBIES_1_001 in 2014 every day, whose seasonality seemed very noisy, while the one on the bottom showed sales every week. When observed values were between 10 and 40, residual values even reached 5, which took a large part of the original data (more than 10%). In previous cases of time series decompositions using STL, residual values were always small compared to observed ones, where the ratio was only 0.001 or even 0.0001.

These 2 results inspired us to set up the criteria for data filtering: (1) Reduce the number of points we would use to forecast to avoid the noisy seasonality; (2) Enlarge values so that it can decrease the error of forecasting. For (1), we still used sales every week instead of every day, and we considered aggregated from products to departments for (2). So finally, we determined to forecast on Level 7 in Figure 3.

During aggregating, we noticed that the prices of different products varied. To keep units the same, we transferred sales into dollars by multiplying prices.

B. Data preparation

The following criteria were followed to clean the data. We got 21-time series after preparation

First, choose one state (CA\TX\WI) and add sales of a product every day in all stores here to those every week. Multiple prices every week to sales, then get turnovers of a product every week (from Week_11149 to Week_11564).

Add up turnovers of all products every week in a department (e.g. HOBBIES_1), then we would get turnovers of a department.

Record the events that happened corresponding to the specified week, the state, and the category to which this department belonged.

Repeat the above steps until all departments in 3 states have been prepared

To verify the performance of the Ensemble learning, we divide our datasets into 3 parts, as shown in Fig.4.

Fig. 4. After the decomposition

C. Basic models

1) Naive

The Naïve approach considers what happened to the last observed value and predicts the same value in the future. It is the simplest forecasting model and provides a benchmark against which more sophisticated models can be compared. This method sometimes works well for economic and financial time series, which often have patterns that are difficult to reliably and accurately predict. The approach can be written in the time series notation as:

$$\hat{y}_{T+h} = y_T \quad (1)$$

2) Moving Average

In statistics, moving averages (rolling average or running average) are calculations to analyze data points by creating a series of averages of different subsets of the full data set. There are 2 kinds of moving averages: Centered and Trailing.

For centered moving averages, given a series of numbers and a fixed subset size, then we make the average of previous and future values in this subset. A centered moving average of order can be written as:

$$\hat{T}_t = \frac{1}{m}\sum_{j=-k}^{k} y_{t+j} \quad (2)$$

3) Prophet

Prophet, which was open-sourced by Facebook's data scientists in 2017, is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, considering holiday effects. It works best with time series with solid seasonal effects and several seasons of historical data.

4) ETS

ETS contains several state space models[10]. Each of them has a measurement equation that describes the original data, and some state equations that show how the hidden components (e.g. level, trend, seasonal) change over time. To denote every state space model, remark ETS (Error, Trend, Seasonal), where Error= {A, M}, Trend= {N, A, Ad} and Seasonal= {N, A, M}. And A stands for additive, M for multiplicative, N for none, and Ad for additive damped.

What is worth mentioning is that forecasting results given by ETS include not only the exact values but also 2 confident intervals (80% and 90%), which tells an entrepreneur more uncertainties to adjust their plans.

D. Ensemble Learning

1) Basic concept

Ensemble learning is to build and combine multiple learners to complete forecasting tasks. The process is that firstly we generate a group of "individual learners." Then we combine them with some strategies. Generally speaking, individual learners are some common machine learning algorithms, such as decision trees, SVM, neural networks, etc. There are generally two kinds of integration here: homogeneous and heterogeneous. Homogeneity means that all individual learners are of the same type. The individual learners in this homogeneous integration are also called "base learners." Heterogeneity means that individual learners contain different types of learning algorithms, such as both decision trees and neural networks. Generally, what we often use is homogeneous, which means individual learners are of the same type.

Although there are varieties of ways to combine different models and learners to predict, there are three kinds of ensemble techniques that we most commonly use and discuss in practice. Because of their great popularity and performance, we refer to them as "standard learning strategies"; they are Bagging, Stacking, and Boosting.

2) Linear Regression (LR)

Linear Regression (or Multiple Linear Regression) is a statistical technique that linearly uses numerous explanatory variables to predict the result of the response variable. The fitting of the datasets to the LR is a process that models the linear relationship between independent variables and dependent variables.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in} + \varepsilon \quad (3)$$

In our research, the validating outcomes of the four algorithms used (Naïve, Moving Average, Prophet, ETS) were mutually independent. Moreover, the actual values in validation sets were regarded as the only dependent variable. So here $i = 1; k = 1,2,3,4; p = 1,2,3,4$.

3) Simple Average

For continuous numerical output, the typical strategy is averaging. The easiest of the average methods is Simple Average. The formula and calculation of the Simple Average for our problem are as follows:

$$H(x) = \frac{1}{T}\sum_{i=1}^{T} h_i(x) \quad (4)$$

4) Weighted Average

On the basis of the above, we changed the same coefficient of 0.25 into different weights according to the characteristics of 4 basic learning methods. We chose the Random Forest to solve weights. Here we just easily explain how the Random Forest works and omit the detailed calculations.

The Random Forest consists of many decision trees, and trees should be independent of each other. When we input information, n results of classifications will be got from n trees. The Random Forest ensembles the functions of "voting" and designate the result with the most votes as the final output. Thus, it can describe how different factors contribute to the output. To form a random forest, the Bootstrap Sample is acted on validation sets, and the out-of-bag error is relied on to estimate whether a random forest is valid.

The Scikit-learn library in Python is equipped with the Random Forest Classifier function. With the help of it, we can quickly get the weights of 4 basic learning methods.

E. Evaluation indicators

To evaluate how basic learning methods and ensemble learning perform, we introduced 2 indicators: MAPE and RMSE. Both of them act on forecasting values and real values.

1) MAPE

The mean absolute percentage error (MAPE), also known as the mean absolute percentage deviation (MAPD), is a measure of the prediction accuracy of a forecasting method in statistics. It usually expresses the accuracy as a ratio defined by the formula:

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad (5)$$

where A_t is the actual value and F_t is the forecast value. Their difference is divided by the actual value A_t. The absolute value in this ratio is summed for every forecasting point in the period and divided by the number of fitted points n.

2) RMSE

The RMSE represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. It is a measure of accuracy, comparing forecasting errors of different models for a particular dataset but not between datasets, as it is scale-dependent.

The RMSE serves to aggregate the magnitudes of the errors in predictions for various data points into a single measure of predictive power with the formula:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(x_{1,t} - x_{2,t})^2}{T}} \quad (6)$$

where $x_{1,t}$, $x_{2,t}$ are respectively the real value and the forecasting value at t, and there are T pairs of values altogether to be calculated.

F. Forecasting Steps

Choose a certain department. Use the Naïve method, Moving Averages, Prophet, and ETS respectively to give the forecasting of 16 points (Week_11533 to Week_11548) by training 193 history points (Week_11149-11532).

Calculate the MAPE and RMSE of each basic learning method.

Construct Ensemble Learning methods for this department using the 3 approaches mentioned above. Choose the approach with the least MAPE and RMSE as the final Ensemble Learning method of it.

Repeat the above steps until 21 departments are all finished forecasting.

IV. MODEL SOLVING AND ANALYSIS

A. Evaluations on validation sets

Take the department HOBBIES_1_CA (HOBBIES_1 in CA) as an example. We followed the steps above, and we got the results as shown in TABLE I:

TABLE I. FORECASTING RESULTS OF HOBBIES_1_CA AND ENSEMBLE LEARNING CONSTRUCTING

Week_id	Naïve	Moving Average	Prophet	ETS	Ensemble_1	Ensemble_2	Ensemble_3	Test
W_11533	73,099	66,839	62,235	67,321	72,892	67,374	65,471	71,663
W_11534	73,099	70,116	62,431	67,488	67,910	68,284	66,682	62,352
W_11535	73,099	69,038	62,783	67,654	69,151	68,144	66,496	76,783
W_11536	73,099	70,266	61,962	67,820	67,503	68,287	66,687	67,870
W_11537	73,099	69,002	62,081	67,987	69,099	68,042	66,362	69,085
W_11538	73,099	71,246	63,278	68,153	65,236	68,944	67,562	64,195
W_11539	73,099	67,050	65,597	68,319	70,243	68,517	66,993	71,075
W_11540	73,099	68,118	66,149	68,486	68,315	68,963	67,587	69,485
W_11541	73,099	68,251	66,772	68,652	67,709	69,194	67,894	68,613
W_11542	73,099	69,724	65,307	68,819	65,947	69,237	67,953	66,911
W_11543	73,099	68,336	65,126	68,985	67,832	68,887	67,487	65,681
W_11544	73,099	67,069	65,586	69,151	69,307	68,726	67,274	66,432
W_11545	73,099	66,342	66,287	69,318	69,913	68,761	67,321	70,056
W_11546	73,099	67,390	65,271	69,484	68,596	68,811	67,388	66,629
W_11547	73,099	67,706	64,630	69,650	68,196	68,771	67,336	68,159
W_11548	73,099	68,281	63,978	69,817	67,426	68,794	67,366	70,283

Ensemble Learning 1: Linear Regression	
Coefficient	0 -1.44211115 -0.3714299 -1.09738067
Intercept	266273.5877
Error	2.70%

Ensemble Learning 2: Making Average	
Coefficient	0.25 0.25 0.25 0.25
Intercept	0
Error	3.84%

Ensemble Learning 3: Weight Average	
Coefficient	0 0.33250239600570786 0.33226586666239475 0.3352317373703413
Intercept	0
Error	4.09%

We got forecasting values from each learning method. Based on basic learning methods, we constructed the Ensemble approach and found that Method 1(Linear Regression) acted best with the least MAPE. Then we calculated the MAPE for basic learning methods and repeated for 21 times. The results are shown in TABLE II and TABLE III:

TABLE II. THE MAPE OF EACH LEARNING METHOD ON VALIDATION SETS

MAPE	Naive	Moving average	Prophet	ETS	Ensemble	Ensemble to be chosen
HOBBIES_1_CA	7.62%	4.45%	6.69%	6.01%	2.70%	Method 1
HOBBIES_1_TX	6.09%	7.10%	7.58%	7.58%	6.02%	Method 1
HOBBIES_1_WI	12.10%	7.20%	9.78%	8.83%	6.36%	Method 1
HOBBIES_2_CA	19.80%	21.17%	13.12%	11.15%	10.83%	Method 1
HOBBIES_2_TX	24.56%	22.64%	19.28%	13.87%	11.71%	Method 1
HOBBIES_2_WI	24.78%	21.89%	10.93%	15.06%	10.72%	Method 1
FOODS_1_CA	7.52%	8.83%	6.79%	7.54%	6.53%	Method 1
FOODS_1_TX	4.87%	5.49%	6.64%	6.84%	5.51%	Method 2
FOODS_1_WI	8.70%	9.20%	12.90%	9.19%	6.20%	Method 1
FOODS_2_CA	9.86%	8.94%	8.37%	9.70%	6.66%	Method 1
FOODS_2_TX	13.23%	13.44%	9.26%	12.50%	6.42%	Method 1
FOODS_2_WI	19.43%	25.38%	17.11%	21.19%	6.51%	Method 1
FOODS_3_CA	13.21%	4.41%	8.48%	5.67%	2.97%	Method 1
FOODS_3_TX	7.22%	7.69%	9.61%	6.77%	5.59%	Method 1
FOODS_3_WI	12.05%	11.48%	8.19%	11.20%	4.18%	Method 1
HOUSEHOLD_1_CA	4.71%	4.47%	3.61%	3.67%	1.83%	Method 1
HOUSEHOLD_1_TX	3.43%	3.88%	6.58%	4.23%	2.46%	Method 1
HOUSEHOLD_1_WI	6.12%	7.83%	6.22%	7.56%	3.61%	Method 1
HOUSEHOLD_2_CA	10.70%	4.60%	10.19%	3.85%	2.78%	Method 1
HOUSEHOLD_2_TX	3.01%	3.55%	4.18%	5.01%	3.44%	Method 1
HOUSEHOLD_2_WI	4.82%	5.64%	5.10%	6.56%	3.70%	Method 1

TABLE III. THE RMSE OF EACH LEARNING METHOD ON VALIDATION SETS

RMSE	Naive	Moving average	Prophet	ETS	Ensemble
HOBBIES_1_CA	5658.93	3962.06	6657.81	3977.62	2745.69
HOBBIES_1_TX	3404.44	3713.19	3855.04	3126.3	2966.57
HOBBIES_1_WI	3932.989	2564.81	3361.83	2810.03	2217.71
HOBBIES_2_CA	749.43	692.88	472.25	296.24	351.49
HOBBIES_2_TX	782.07	604.08	639.09	297.68	334.4
HOBBIES_2_WI	554.46	439.91	297.22	219.59	232.59
FOODS_1_CA	2640.35	3120.96	2229.34	2518.09	1998.7
FOODS_1_TX	1254.88	1518.57	1448.71	1479.23	1303.38
FOODS_1_WI	1593	1905.6	2265.52	1512.21	1125.45
FOODS_2_CA	5231.55	5689.96	5593.59	4261.84	3713.97
FOODS_2_TX	4709.49	4612.83	3484.85	3884.84	1888.77
FOODS_2_WI	9239.41	12385.06	9533.79	7303.88	3658.72
FOODS_3_CA	18679.32	7008	12293.9	8996.34	4973.88
FOODS_3_TX	6921.57	7388.1	9521.44	7169.6	5176.07
FOODS_3_WI	11758.79	10505.5	9392.96	10819.79	4122.16
HOUSEHOLD_1_CA	5717.26	5033.08	3715.93	3435	1958.56
HOUSEHOLD_1_TX	2732.99	3275.67	4486.18	2757.34	1765.63
HOUSEHOLD_1_WI	3641.74	4436.82	3520.07	3755.94	2150.76
HOUSEHOLD_2_CA	3500.06	1648.89	3295.73	1423	1060.4
HOUSEHOLD_2_TX	623.38	723.79	766.29	914.61	663.21
HOUSEHOLD_2_WI	667.8	805.86	714.62	946.86	504.91

From TABLE II, we saw that Ensemble Learning always had the least MAPE, and Method 1 won almost in every department except for FOODS_1_TX, which meant Ensemble

Learning performed well on validation sets, and our innovative Method 1 is useful.

From TABLE III, sometimes the RMSEs of Ensemble Learning were larger than those of basic learning methods, which meant values from Ensemble methods were more unstable than those from other methods. But differences in their RMSEs were too small to influence their results. So, we still considered the Ensemble one.

B. Evaluations on validation sets

To avoid the error brought by the overfitting of Ensemble Learning on validation sets, we used it to forecast the values on test sets to see if it was still well-performed. Here we only showed detailed situations in 2 departments.as shown in TABLE IV.

TABLE IV. FORECASTING ON TEST SETS OF HOBBIES_1_CA

Week_id	Naive	Moving Average	Prophet	ETS	Test	Ensemble_1
W_11549	70, 283	68, 357	69, 697	69, 045	70, 874	66, 039
W_11550	70, 283	69, 772	71, 989	69, 201	62, 676	62, 976
W_11551	70, 283	67, 944	73, 975	69, 357	65, 068	64, 703
W_11552	70, 283	66, 206	74, 055	69, 513	66, 654	67, 009
W_11553	70, 283	64, 799	72, 771	69, 669	72, 493	69, 343
W_11554	70, 283	68, 072	72, 280	69, 824	65, 428	64, 636
W_11555	70, 283	68, 191	73, 961	69, 980	71, 730	63, 668
W_11556	70, 283	69, 884	76, 849	70, 136	68, 234	59, 983
W_11557	70, 283	68, 464	78, 629	70, 292	70, 784	61, 199
W_11558	70, 283	70, 250	78, 064	70, 448	70, 675	58, 663
W_11559	70, 283	69, 898	76, 222	70, 604	64, 149	59, 683
W_11560	70, 283	68, 536	75, 260	70, 760	65, 619	61, 833
W_11561	70, 283	66, 814	76, 197	70, 915	66, 086	63, 797
W_11562	70, 283	65, 285	78, 019	71, 071	68, 455	65, 155
W_11563	70, 283	66, 720	78, 895	71, 227	58, 556	62, 589
W_11564	70, 283	64, 365	78, 021	71, 383	63, 639	66, 138

The best method in Ensemble learning: Method 1	
The best method in all learning: Moving Average	
MAPE	5. 00%
RMSE	4119. 05

For HOBBIES_1_CA, moving average won surprisingly, while the MAPE of LR method (Method 1) in ensemble learning is 6.23%, which was a bit larger than that of moving average.

Simple average method (Method 2) in ensemble learning was the best approach for HOBBIES_1_TX. When we did LR method, something unexpected happened. The result is shown in Fig.5 and the prediction results of the test set are shown in TABLE V.

```
Out[6]: array([-193542.76122715, -194203.67581119, -194450.03676011,
               -195191.70528168, -197098.4939659 , -200019.57930051,
               -202807.07089787, -203659.61402036, -202140.51921957,
               -199846.20205786, -199355.46383324, -201786.20108768,
               -205804.07957552, -208506.62756816, -208959.14069397,
               -207764.87180063])
```

Fig. 5. Use Method 1 to forecast

TABLE V. FORECASTING ON TEST SETS OF HOBBIES_1_TX

Week_id	Naive	Moving Average	Prophet	ETS	Test	Ensemble_2
W_11549	44, 092	41, 324	46, 184	42, 234	43, 651	43, 459
W_11550	44, 092	40, 953	46, 515	42, 234	46, 012	43, 448
W_11551	44, 092	44, 585	46, 779	42, 234	47, 338	44, 422
W_11552	44, 092	45, 667	47, 206	42, 234	43, 040	44, 800
W_11553	44, 092	45, 464	48, 192	42, 234	51, 277	44, 996
W_11554	44, 092	47, 218	49, 781	42, 234	52, 438	45, 831
W_11555	44, 092	48, 918	51, 298	42, 234	46, 187	46, 636
W_11556	44, 092	49, 967	51, 782	42, 234	48, 499	47, 019
W_11557	44, 092	49, 042	50, 955	42, 234	48, 732	46, 581
W_11558	44, 092	47, 806	49, 713	42, 234	44, 782	45, 961
W_11559	44, 092	47, 338	49, 439	42, 234	49, 916	45, 776
W_11560	44, 092	47, 810	50, 725	42, 234	45, 782	46, 215
W_11561	44, 092	46, 827	52, 784	42, 234	46, 918	46, 484
W_11562	44, 092	47, 539	54, 219	42, 234	41, 195	47, 021
W_11563	44, 092	44, 632	54, 347	42, 234	45, 105	46, 326
W_11564	44, 092	44, 406	53, 716	42, 234	50, 521	46, 112

The best method in Ensemble learning: Method 2	
The best method in all learning: Ensemble_2	
MAPE	5. 50%
RMSE	2563. 84

Forecasting values were all negative using Method 1. When we recalled Method 1 on validation sets, we found that the absolute values of some coefficients and the intercept from the Linear Regression were so big that even a slight disturbance would make the results change drastically. The result is shown in Fig.6.

Ensemble Learning 1: Linear Regression	
Coefficient	0 0.07 -1.91 279.71
Intercept	0-11921214.745767144
Error	6.02%

Fig. 6. Parameters of the Linear Regression

We inferred that Method 1 performed well when it had relatively proper coefficients and intercepts, otherwise methods of averages worked better.

Forecasting results of 21 departments on test sets are shown in TABLE VI:

TABLE VI. RESULTS OF 21 DEPARTMENTS ON TEST SETS

Collection	The best method in Ensemble Learning(1)	MAPE of (1)	RMSE of (1)	The best method in all Learning(2)	MAPE of (2)	RMSE of (2)
HOBBIES_1_CA	Method 1	6.23%	5456.92	Moving Average	5.00%	4119.05
HOBBIES_1_TX	Method 2	5.50%	2563.84	Ensemble_2	5.50%	2563.84
HOBBIES_1_WI	Method 1	6.97%	2419.18	Prophet	4.91%	1786.55
HOBBIES_2_CA	Method 2	7.86%	193.05	Ensemble_2	7.86%	193.05
HOBBIES_2_TX	Method 3	12.07%	275.3	Ensemble_3	12.07%	275.3
HOBBIES_2_WI	Method 2	10.99%	169.74	Ensemble_2	10.99%	169.74
FOODS_1_CA	Method 3	5.04%	1606.51	ETS	4.19%	1194.63
FOODS_1_TX	Method 2	6.57%	1130.75	ETS	6.48%	1138.48
FOODS_1_WI	Method 3	4.85%	1073.32	Ensemble_3	4.85%	1073.32
FOODS_2_CA	Method 3	8%	5393.71	Prophet	6.77%	4330.39
FOODS_2_TX	Method 3	12%	4675.74	Prophet	11.08%	4312.21
FOODS_2_WI	Method 3	20.28%	15306.62	ETS	17.85%	13792.49
FOODS_3_CA	Method 3	4.19%	7317.27	Ensemble_3	4.19%	7317.27
FOODS_3_TX	Method 3	6.09%	6287.32	Ensemble_3	6.09%	6287.32
FOODS_3_WI	Method 3	3.18%	3904.76	Ensemble_3	3.18%	3904.76
HOUSEHOLD_1_CA	Method 3	3.55%	4402.82	Ensemble_3	3.55%	4403.82
HOUSEHOLD_1_TX	Method 3	3.34%	2633.55	Ensemble_3	3.34%	2633.55
HOUSEHOLD_1_WI	Method 3	6.35%	4252.21	Prophet	4.42%	3319.48
HOUSEHOLD_2_CA	Method 3	2.54%	993.26	Ensemble_3	2.54%	933.26
HOUSEHOLD_2_TX	Method 3	4.02%	1058	Ensemble_3	4.02%	1058
HOUSEHOLD_2_WI	Method 3	4.53%	771.09	Prophet	4.18%	771.09

We marked the results of departments whose best method wasn't one of Ensemble Learning in blue and some conclusions were got:

In Ensemble Learning, weighted average method (Method 3) had the absolute advantage. Although Method 1 had the best results in most departments, it became unstable when we change the datasets.

More than 57% of departments fit best by Ensemble methods. Although sometimes basic methods won, the differences between basic methods and Ensemble methods were less than 3%. This error is small and acceptable, so Ensemble methods are recommended to use.

For results with big errors (MAPE ≥ 10%), ETS would be the best choice because it gave confident intervals, so at least entrepreneurs can estimate the possible range of sales. Here was an example of ETS, with the deep blue for the 80% confident interval, and light blue for the 95% confident interval. An example of ETS is shown in Fig.7.

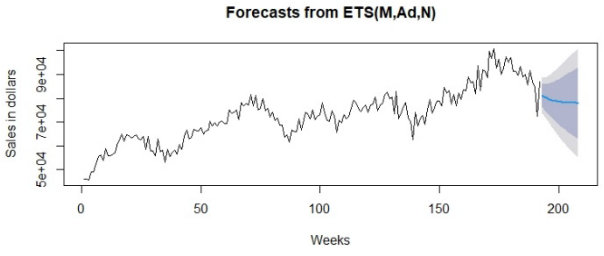

Fig. 7. An example of ETS

We recommended the sequence to use learning methods for entrepreneurs: Ensemble_3, Ensemble_2, ETS, and then Prophet.

V. CONCLUSION

Companies worldwide are trying to forecast the demand for their products from the market because they can base their production plan on the forecast results and get more profits.

Therefore, this research focused on forecasting the weekly retail sales using datasets provided by Walmart concerning their retail data over five years. We initially classified the forecast subject into 21-time series according to different departments and states. Then we finished the data cleaning and prepared 225 weeks' data for each time series (first 193 weeks for training, next 16 weeks for validation, and last 16 weeks for testing). Next, we used the first 193 weeks' data to train four Machine Learning models (Naïve, Moving average, Prophet, ETS) and got the forecast outcomes for the next 16 weeks. Afterward, we chose "Stacking" as the ensemble technique and did some ensemble learning: We took those four models as primary learners and chose Linear Regression, Simple Average, and Weighted Average (Random Forest) as three secondary learners. We then fit the results of the four primary learners to the Ensemble model on validation sets and got the corresponding parameters. At last, we verified and compared the performance of those four Machine Learning models and three Ensemble models on test sets.

Our research shows: In the validation stage, the Ensemble model using Linear Regression fit the best in almost every department according to the MAPE and RMSE lists. However, the Weighted Average method supported by Random Forest behaved the best in Ensemble learning in the testing stage, while the Linear Regression was extremely unstable with significant errors. Therefore, we concluded that the Linear Regression was overfitting in the validation stage. On the contrary, the Random Forest method showed its strong generalization capability and anti-over-fitting ability as a "bagging" method in the test set. What is more, although simple Machine Learning models behave better than Ensemble Learning in half of the departments in the testing stage, their differences are less than 3%, so we recommend Ensemble Learning, especially Weighted Average, to be a universal method.

We propose that the Machine Learning model for time series and Ensemble models used in the research can be adopted for demand forecasting. Enterprises should select the most appropriate model according to the historical data of relevant products. Furthermore, choosing a secondary learner with a strong generalization capability is more advisable to stack primary models for Ensemble Learning to avoid over-fitting.

Results from the model will hopefully serve as useful feedback information to improve sales forecasting. Furthermore, future studies are needed to overcome the difficulties of combining sales with external factors from the supply chain. Not only considering the different sources of data such as inventory, manufacturing, warehousing, and transportation to reduce the uncertainties arising from variations in customers' demand, supplies transportation, organizational risks, and lead times, but also incorporating existing driving factors outside the historical data such as economic instability, inflation, and purchasing power, to help adjust the predictions considering unseen future scenarios of demand. Combining predictive algorithms with optimization or simulation can equip the models with prescriptive capabilities in response to future scenarios and expectations.

REFERENCES

[1] You Z, Si Y-W, Zhang D, Zeng X, Leung SCH, Li T. A decision-making framework for precision marketing. Expert Syst Appl. 2015;42(7):3357–67.

[2] Büyüközkan.G. & Göçer.F.(2018). Digital Supply Chain: literature review and a proposed framework for future research. Comput Ind. 97,157–77.

[3] Kshetri.N.(2018). Blockchain's roles in meeting key supply chain management objectives. Int J Inf Manage. 39,80–9.

[4] Michna.Z, Disney. S.M, & Nielsen.P.(2019) The impact of stochastic lead times on the bullwhip effect under correlated demand and moving average forecasts.

[5] Zhu.Y, Zhao.Y, Zhang.J, Geng.N, & Huang.D.(2019). Spring onion seed demand forecasting using a hybrid Holt-Winters and support vector machine model. PLoS ONE. 14(7).

[6] Hyndman, R. J. (2020). A brief history of forecasting competitions. International Journal of Forecasting, 36, 7–14.

[7] Syntetos, A. A, & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. International Journal of Forecasting, 21, 303–314.

[8] Syntetos, A. A., Boylan, J. E., & Croston, J. D. (2005). On the categorization of demand patterns. Journal of the Operational Research Society, 56, 495–503.

[9] Spyros, M, Evangelos, S, &Vassilios, A. (2022). The M5 competition: Background, organization, and implementation. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2021.11.013.

[10] Rob, J.H, &George, A. (2021). Exponential smoothing. In Rob, J.H, &George, A. (Ed.), Forecasting: Principles and practice. Otexts.