Major Research Project

Walmart Sales Forecasting

Literature Review and Exploratory Data Analysis

Moeen Bagheri

June 2020

Literature Review

Numerous statistical and deep learning methods have been employed in the past for forecasting sales. Linear statistical models, such as multivariate linear regression, which make predictions based on the historical relationship between different influential factors and the demand, have the advantage of being efficient. However, these linear models perform well for linear problems, where the relationship between the dependent variable and one or more independent variables is linear with a constant rate of change, and hence fail to capture the nonlinear relationships and describe the complexity of the supply chain [1]. Similarly, basic univariate models, such as ARIMA, which consider the data as a time-series, are also unable to describe the complexity of the supply chain since they are only capable of capturing linear relationships in the data [1, 2]. On the other hand, nonlinear statistical models used for sales forecasting include Bayesian networks, support vector machines, and Markov chains. Unlike linear models, these models are able to learn complex nonlinear relationships from the data. In a study done in [3], a Random Forest model was used to forecast Major League Baseball game ticket sales. Their approach consisted of a dynamic month-ahead forecasting strategy, where the data is updated every month. Their results showed that their proposed RF model slightly outperforms their baseline model, which they chose to be an Ordinary Least Squares (OLS) regression model. However, even though nonlinear statistical models have a high capability of solving complex problems, selecting the right model for a certain problem is a difficult task, which requires expert knowledge of statistical models. Moreover, these methods are usually found to perform worse compared to deep learning methods [1].

Deep learning methods are able to automatically extract important features and have been found to obtain better results compared to statistical models [1]. The self-organizing and self-adjusting capabilities of Artificial Neural Networks (ANN) allows them to solve

complex nonlinear problems [4]. In a study done by [5], a Multi-Layer Perceptron (MLP) was used to predict the monthly sale volumes of a Polish company, which imports fabric on a monthly basis, based on the previous three months. The MLP contained three input neurons, a single hidden layer with 15 neurons, as well as one output neuron, and was able to achieve a high accuracy on the data with a Root-Mean-Square Error (RMSE) of 3.34e-11. However, even though ANNs excel at solving complex problems, they lack the ability to interpolate and predict long-term sequences [4]. Alternatively, deep learning methods, such as LSTM, are able to preserve past information and capture the temporal relationships in the data [6]. However, even though deep learning methods can improve the accuracy of the predictions compared to statistical models, it is much more challenging to interpolate and draw conclusions from their results [1, 5].

In a study done in [7], the performances of ARIMA, MLP, and LSTM models were compared for forecasting and predicting cash flow. Interest Opportunity Cost (IOC), which is a measure based on financial concepts and allows finance-specific comparison of the models, was used in MLP and LSTM as the error function to be optimized. According to their results, LSTM was able to obtain the minimum error of 0.09, compared to MLP and ARIMA with an error of 0.10 and 0.23, respectively. The cash flow data exhibited a strong weekly pattern that assisted LSTM in its predictions. However, due to the small amount of data available (3 years), variances caused by holidays and other special events could not be explained.

On a separate note, many methods do not take into account external variables and factors, such as price changes and promotions and have been shown to only perform well in periods without the influence of any external factors [8, 9]. In practice, in order to incorporate the effects of promotions on the sales, many retailers use a base-times-lift approach, where the sales are first forecasted based on a simple time-series and then adjusted based on the incoming promotions [10]. Recent studies have focused on optimizing these adjustments, which are made based on promotions and other external factors [10]. In an alternative approach, hybrid models have been used in order to take advantage of the strengths of different models together, which helps capture both the temporal information in the data, as well as the correlation between the demand and the external factors [1, 11]. The complex behaviour of a time-series cannot be explained by a single model if, for example, the

time-series contains both linear and nonlinear correlations [12]. Hybrid models are usually constructed in a sequential manner, where the first component is fitted to the data first, and then the second component is fitted to the residuals of the first component [12]. The residuals of a model contain the information that could not be captured by that model [11]. However, hybrid models are not guaranteed to perform better than single models and model selection is still a crucial aspect of hybrid models [12].

The study done by [11] presents an example of a hybrid model used for forecasting. In this study, an LSTM model is combined with a Random Forest model to create a hybrid model for forecasting sales of a store with one online and 11 offline sale channels. In the hybrid model, LSTM is applied first to capture the linear and non-linear temporal information from the data. Next, the residuals from the LSTM are used as the dependant variable and the external variables are used as the independent variable in a Random Forest model in order to capture the non-temporal relationships in the data. Another challenge to address is the modeling of sales across multiple channels. One approach would be to model each channel separately, however, this approach will eliminate the aggregate demand information from the data. Therefore, this study forecasted the demand of a product based on its order origin (online vs. offline) instead in order to sustain the aggregate demand information. Their results showed that the hybrid model performed better than its two components, LSTM and RF, individually.

Exploratory Data Analysis

The dataset [13] includes the unit sales of 3,049 products sold by Walmart in the USA in grouped time-series format. The products are classified in three categories (Hobbies, Foods, and Household) and seven departments. Additionally, these products are sold across ten stores, located in the three states of California (CA), Texas (TX), and Wisconsin (WI). Specifically, there are four stores located in California, three in Texas, and three in Wisconsin. The dataset contains the following three data files. Figure A.1 shows a snippet of each file.

- **calendar.csv**: contains information about the dates the products were sold, such as promotions, holidays, and other events [14].
- **sell_prices.csv**: contains the price of the products sold per store and week. The provided prices are an average across seven days [14].
- **sales_train.csv**: contains the daily units sales of each product sold per store for 1913 days [14].

Figure 1 shows the number of products per category. There are 1437, 1047, and 565 products in the `Foods`, `Household`, and `Hobbies` categories, respectively, for a total of 3049 products. Since these products are sold across ten different stores, there is a total of 30,490 items, out of which 22,243 had a price change at some point in time. In addition, from Figure A.2, which shows the price distribution of all

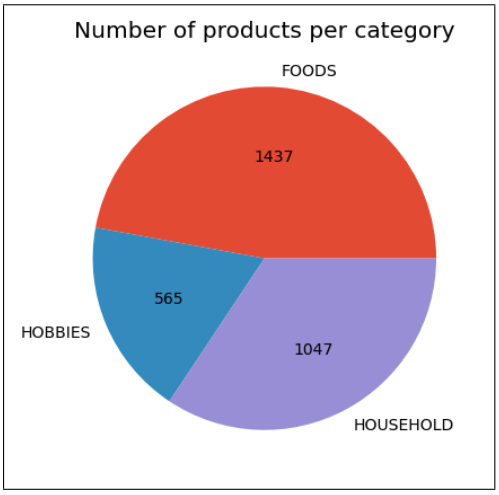

Figure 1: Number of products per category.

products, we can see that most items cost between 3-4 dollars. Furthermore, Figure A.3 shows the price distribution of each product category separately. It is apparent that the `Hobbies` products do not have the same price distribution as `Foods` and `Household` products and tend to have more products that are very cheap.

Furthermore, we examine the number of units sold and revenue earned from each product category. By examining the pie charts in Figure 2, we see that the `Foods` category accounts for most of the units sold and revenue earned, with 68.6% of all products sold belonging to the `Foods` category, which corresponds to 58.0% of the total revenue earned. A possible factor that influences the number of unit sales is the number of products in each category. Since we have a lot more products that belong to the `Foods` category compared to the other two categories, it makes sense that we also have more units being sold in the `Foods` category.

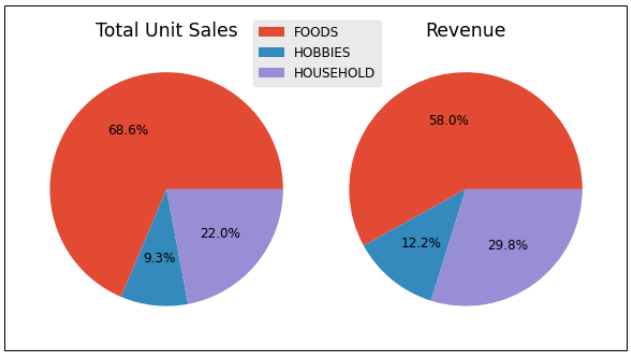

Figure 2: Proportion of units sold and revenue earned in each product category.

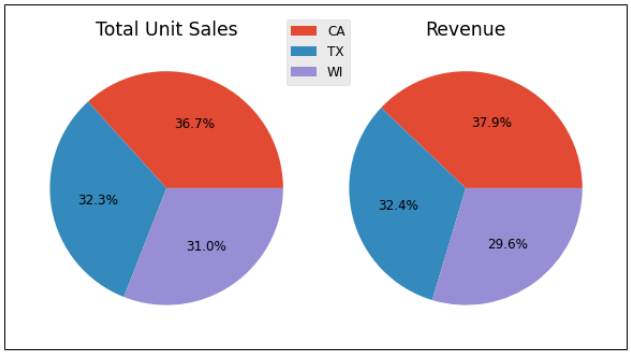

Figure 3: Proportion of total units sold and revenue earned per store in each state.

Another factor to examine is whether there is any difference in the number of units sold and revenue earned between various states. Figure 3 compares the total unit sales and revenue earned in each state. Since the number of stores is not constant between the states, the values for each state have been normalized by the number of stores in that state. By examining the pie charts, we see that, on average, the three states have

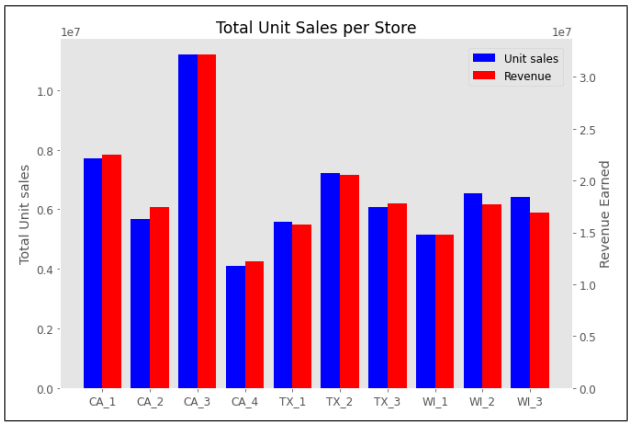

Figure 4: The total number of units sold and revenue earned per store.

about the same number of unit sales and revenue earned per store, however, the store in state of California are, on average, performing slightly better compared to the stores in Texas and Wisconsin. Additionally, we examine the number of unit sales and revenue earned between different stores. From the bar chart in figure 4, we see that store `CA_3` has the highest number of unit sales and revenue earned, and store `CA_4` has the lowest number of unit sales and revenue earned.

On another point, Figure A.4 shows a 4-week moving-average of the revenue earned, number of units sold, and the average price of all items. We see that the average price of products are increasing over time in a linear fashion. Moreover, as expected, revenue and unit sales have increased with almost an identical pattern. Additionally, examining the revenue and unit sales graphs, we can see some form of seasonality in the curves. Hence, we further explore for any seasonality patterns. From Figure A.5, which shows the number of unit sales per day for the year 2011, we see a clear weekly pattern in the number of unit sales. Specifically, the number of units sold tends be the highest during the weekends and the lowest in the middle of the week. This can be confirmed by examining Figure A.6, which shows the average number of units sold per day of week for all years. Additionally, Figure A.5 shows national holidays and other special events, which are considered as external factors and may affect the number of unit sales. Most national holiday events, such as Thanksgiving, have a negative effect on the number of unit sales, possibly because people tend to spend time with their families rather than shopping. Moreover, Figure A.7 shows the number of unit sales per month for all years. We see that the number of unit sales tends to increase as we approach the middle of the year and falls off as we approach the end of the year. This is confirmed by examining Figure A.8, which shows the average number of units sold in each month of the year. Additionally, we can see a spike in unit sales during March.

Appendices

	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
0	2011-01-29	11101	Saturday	1	1	2011	d_1	NaN	NaN	NaN	NaN	0	0	0
1	2011-01-30	11101	Sunday	2	1	2011	d_2	NaN	NaN	NaN	NaN	0	0	0
2	2011-01-31	11101	Monday	3	1	2011	d_3	NaN	NaN	NaN	NaN	0	0	0
3	2011-02-01	11101	Tuesday	4	2	2011	d_4	NaN	NaN	NaN	NaN	1	1	0
4	2011-02-02	11101	Wednesday	5	2	2011	d_5	NaN	NaN	NaN	NaN	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1964	2016-06-15	11620	Wednesday	5	6	2016	d_1965	NaN	NaN	NaN	NaN	0	1	1
1965	2016-06-16	11620	Thursday	6	6	2016	d_1966	NaN	NaN	NaN	NaN	0	0	0
1966	2016-06-17	11620	Friday	7	6	2016	d_1967	NaN	NaN	NaN	NaN	0	0	0
1967	2016-06-18	11621	Saturday	1	6	2016	d_1968	NaN	NaN	NaN	NaN	0	0	0
1968	2016-06-19	11621	Sunday	2	6	2016	d_1969	NBAFinalsEnd	Sporting	Father's day	Cultural	0	0	0

1969 rows × 14 columns

(a) calendar.csv

	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

(b) sell_prices.csv

	id	item_id	dept_id	cat_id	store_id	state_id	d_1	d_2	...	d_1906	d_1907	d_1908	d_1909	d_1910	d_1911	d_1912	d_1913
0	HOBBIES_1_001_CA_1_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	0	0	...	0	1	1	1	3	0	1	1
1	HOBBIES_1_002_CA_1_validation	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	0	0	...	0	0	0	1	0	0	0	0
2	HOBBIES_1_003_CA_1_validation	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	0	0	...	2	1	1	1	0	1	1	1
3	HOBBIES_1_004_CA_1_validation	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	0	0	...	5	4	1	0	1	3	7	2
4	HOBBIES_1_005_CA_1_validation	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	0	0	...	1	0	1	1	2	2	2	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
30485	FOODS_3_823_WI_3_validation	FOODS_3_823	FOODS_3	FOODS	WI_3	WI	0	0	...	0	0	0	0	1	0	0	1
30486	FOODS_3_824_WI_3_validation	FOODS_3_824	FOODS_3	FOODS	WI_3	WI	0	0	...	0	0	0	0	0	0	1	0
30487	FOODS_3_825_WI_3_validation	FOODS_3_825	FOODS_3	FOODS	WI_3	WI	0	6	...	0	2	0	1	0	0	1	0
30488	FOODS_3_826_WI_3_validation	FOODS_3_826	FOODS_3	FOODS	WI_3	WI	0	0	...	1	0	0	1	0	3	1	3
30489	FOODS_3_827_WI_3_validation	FOODS_3_827	FOODS_3	FOODS	WI_3	WI	0	0	...	0	0	0	0	0	0	0	0

30490 rows × 1919 columns

(c) sales_train.csv

Figure A.1: A snippet of the dataset.

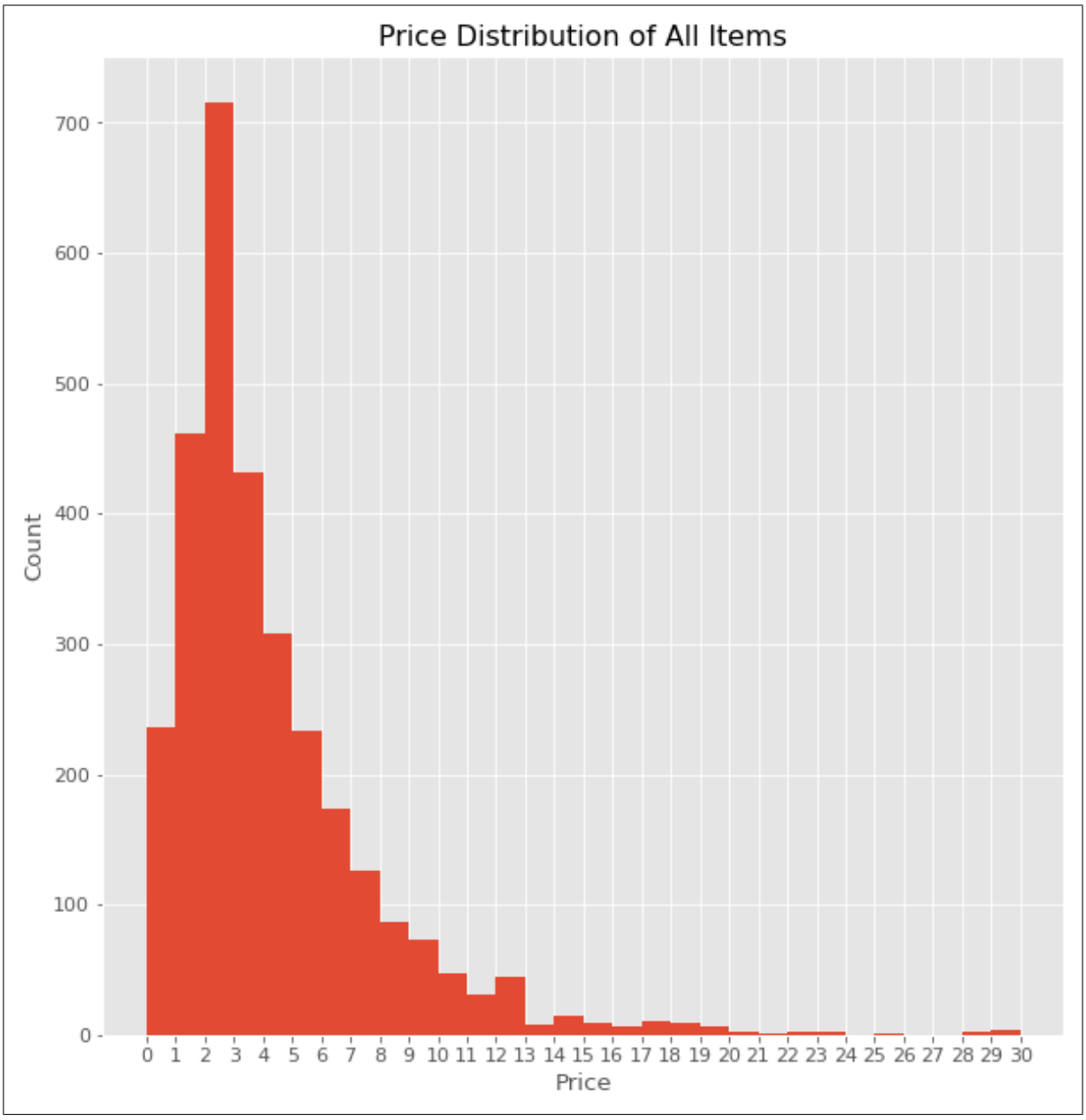

Figure A.2: The distribution of the prices of all products.

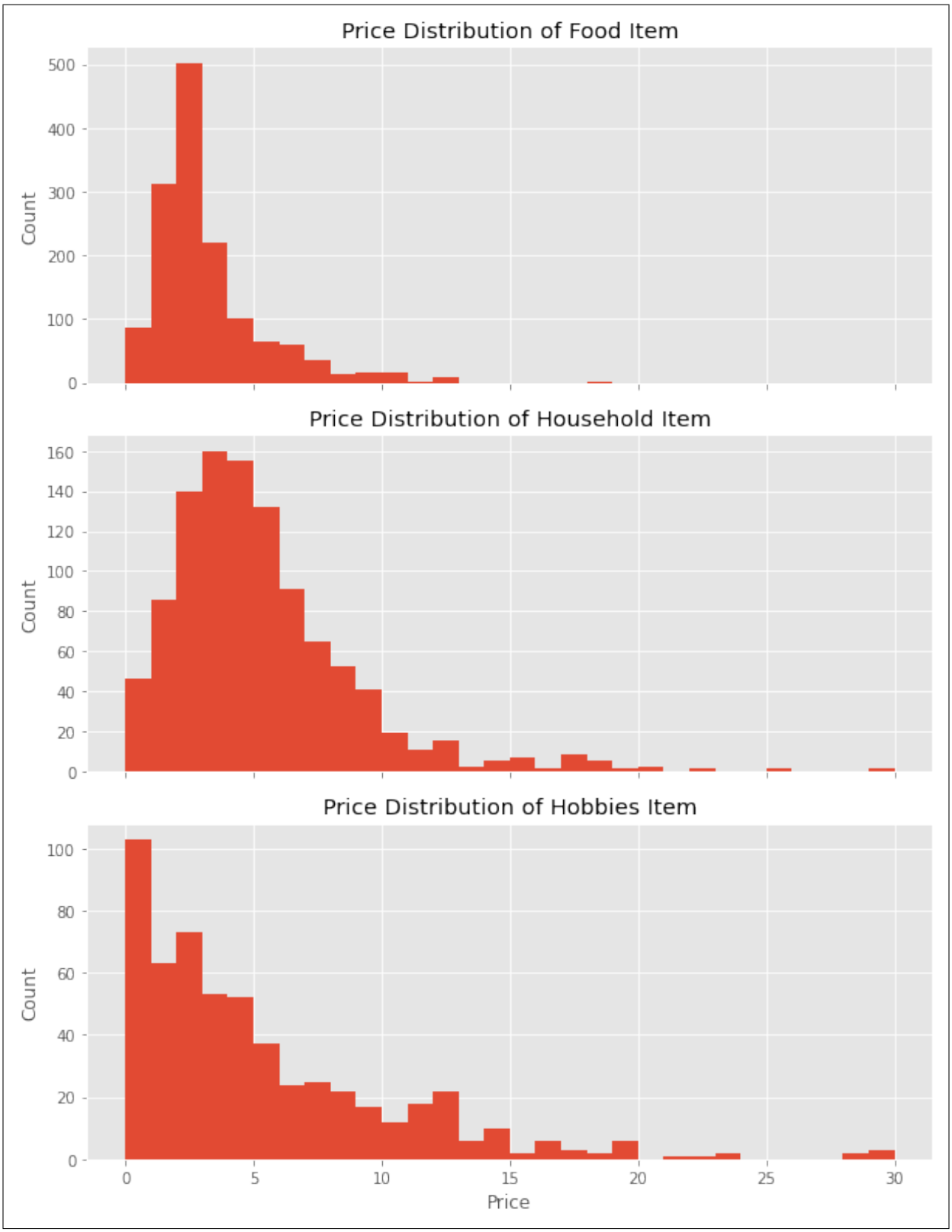

Figure A.3: The price distribution of products in each product category.

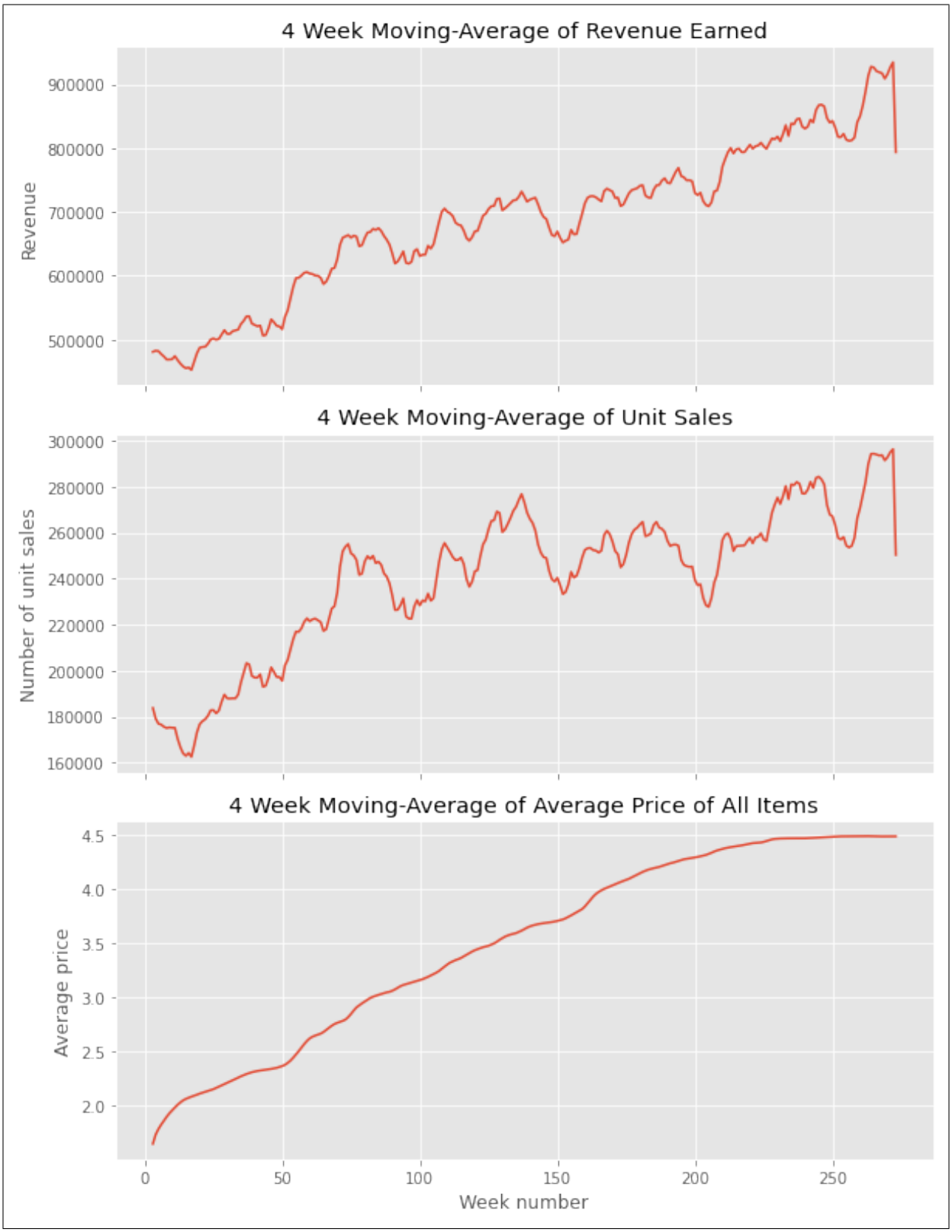

Figure A.4: 4 week moving-average of revenue, unit sales, and average price of items.

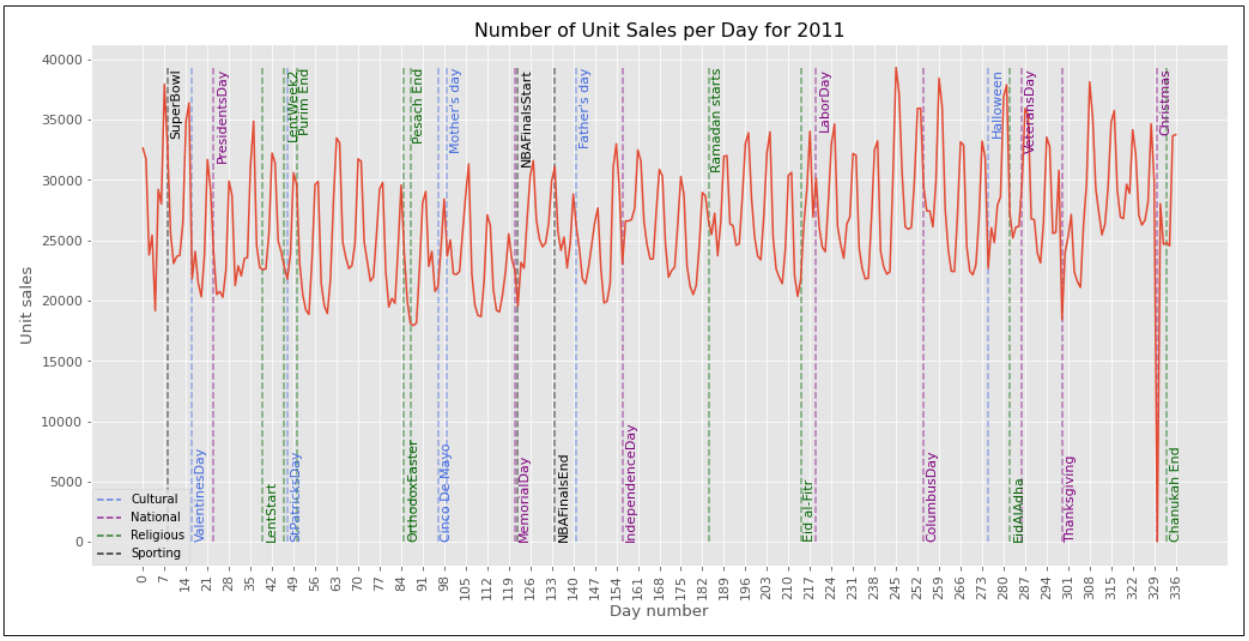

Figure A.5: The number of unit sales per day in 2011.

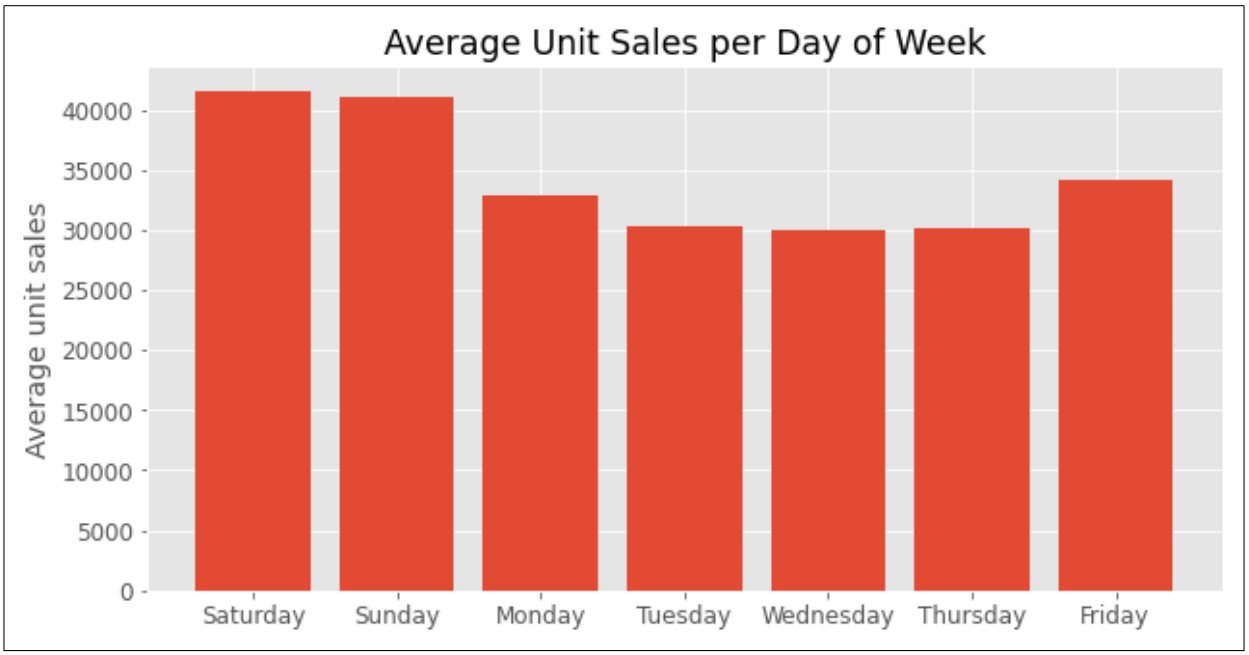

Figure A.6: Total number of units sold per day of week

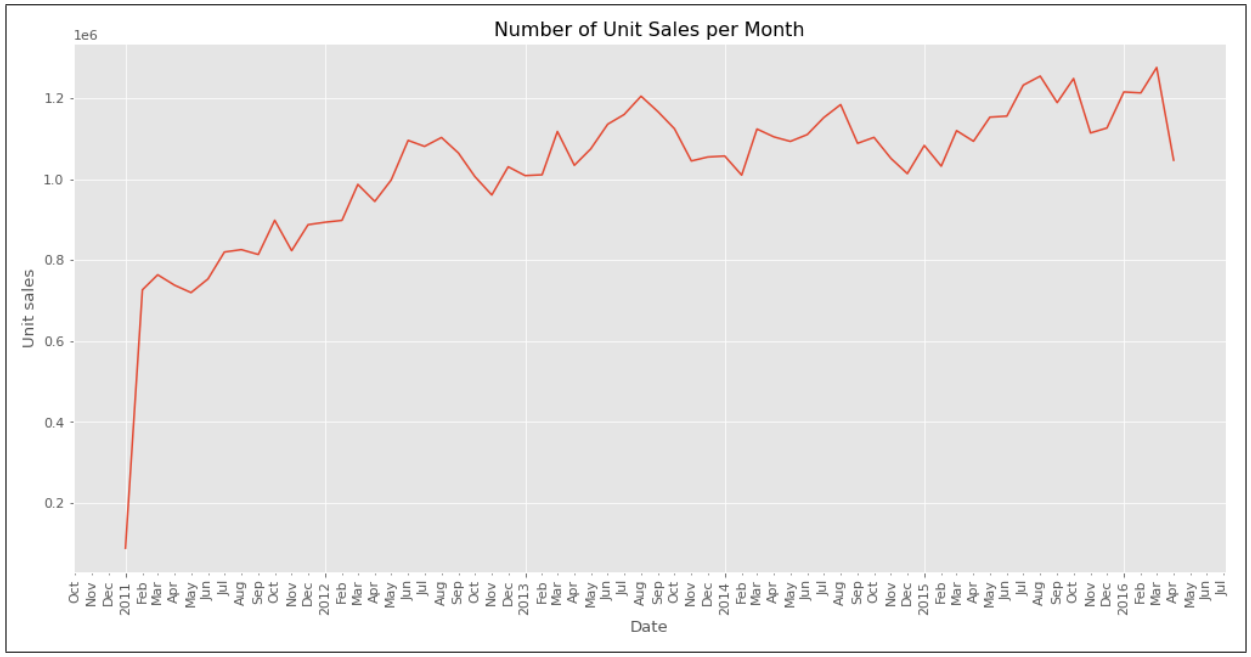

Figure A.7: The number of unit sales per month.

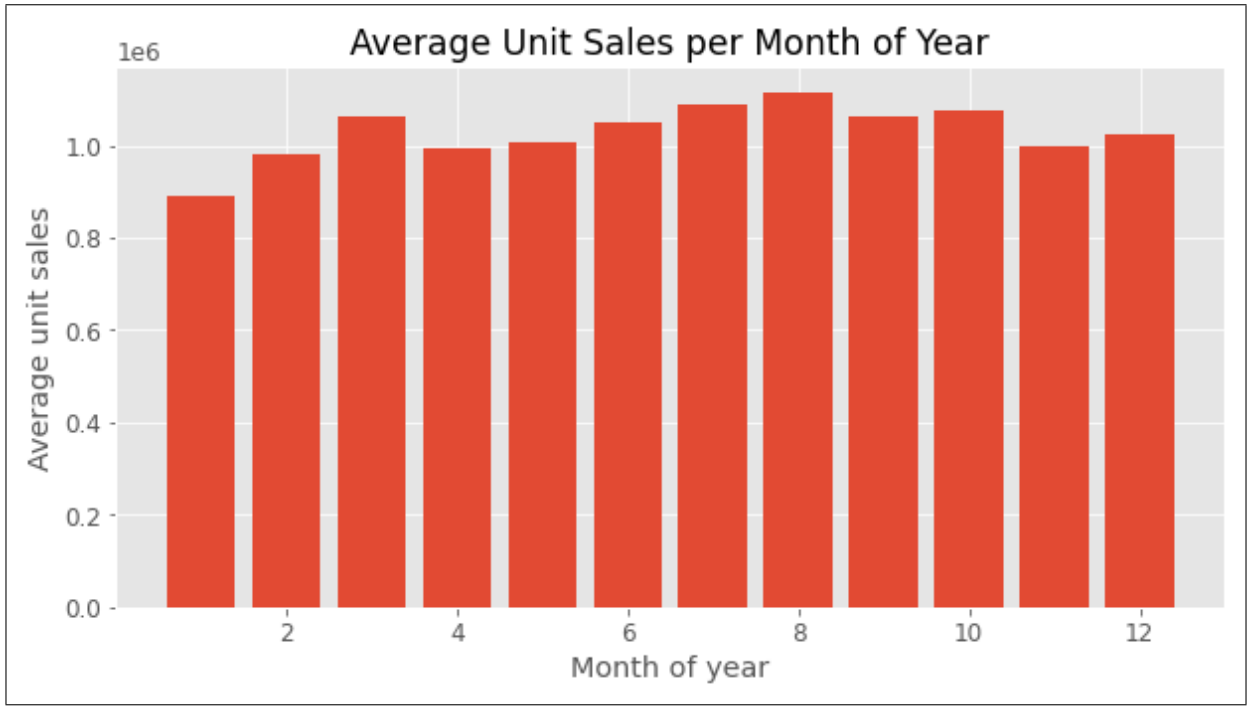

Figure A.8: Average number of units sold in each month.

Bibliography

[1] T. Weng, W. Liu, and J. Xiao, "Supply chain sales forecasting based on lightgbm and lstm combination model," *Industrial Management Data Systems*, vol. 120, no. 2, pp. 265–279, 2019;2020.

[2] Z.-Y. Chen and R. J. Kuo, "Evolutionary algorithm-based radial basis function neural network training for industrial personal computer sales forecasting," *Computational Intelligence*, vol. 33, no. 1, pp. 56–76, 2017.

[3] S. Q. Mueller, "Pre- and within-season attendance forecasting in major league baseball: a random forest approach," *Applied Economics*, vol. 0, no. 0, pp. 1–18, 2020.

[4] B. Shao, M. Li, Y. Zhao, and G. Bian, "Nickel price forecast based on the lstm neural network optimized by the improved pso algorithm," *Mathematical Problems in Engineering*, vol. 2019, pp. 1–15, 2019.

[5] M. Scherer, "Multi-layer neural networks for sales forecasting," *Journal of Applied Mathematics and Computational Mechanics*, vol. 17, no. 1, pp. 61–68, 2018.

[6] S. Helmini, N. Jihan, M. Jayasinghe, and S. Perera, "Sales forecasting using multivariate long short term memory network models," *PeerJ PrePrints*, 2019.

[7] H. Weytjens, E. Lohmann, and M. Kleinsteuber, "Cash flow prediction: Mlp and lstm compared to arima and prophet," *Electronic Commerce Research*, 2019.

[8] C. P. d. Veiga, C. R. P. d. Veiga, W. Puchalski, L. d. S. Coelho, and U. Tortato, "Demand forecasting based on natural computing approaches applied to the foodstuff retail segment," *Journal of Retailing and Consumer Services*, vol. 31, pp. 174–181, 2016.

[9] Ö. G. Ali, S. Sayın, T. Van Woensel, and J. Fransoo, "Sku demand forecasting in the presence of promotions," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12340–12348, 2009.

[10] T. Huang, R. Fildes, and D. Soopramanien, "The value of competitive information in forecasting fmcg retail product sales and the variable selection problem," *European Journal of Operational Research*, vol. 237, no. 2, pp. 738–748, 2014.

[11] S. Punia, K. Nikolopoulos, S. P. Singh, J. K. Madaan, and K. Litsiou, "Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail," *International Journal of Production Research*, pp. 1–16, 2020.

[12] T. Taskaya-Temizel and M. C. Casey, "A comparative study of autoregressive neural network hybrids," *Neural Networks*, vol. 18, no. 5, pp. 781–789, 2005.

[13] Kaggle, "M5 forecasting - accuracy," 2020. https://www.kaggle.com/c/m5-forecasting-accuracy/overview.

[14] MOFC, "The m5 competition," 2020. https://mofc.unic.ac.cy/m5-competition/.