Sales Time Series Forecasting

Applied Forecasting in Complex Systems: Group 9

Jop Rijksbaron
11685514
University of Amsterdam

Lex Poon
11031530
University of Amsterdam

Robin Spiers
11829494
University of Amsterdam

ABSTRACT

For many domains, the ability to produce accurate forecasts can be highly valuable. In this study, multiple forecasting methods are applied to predict the product sales of a California based Walmart store, using historical unit sales time series data. Methods ranged from statistical methods like ARIMA to machine learning methods like Prophet. The results of the comparative study show that the Moving Average (4) model and the Auto ARIMA model delivered the best accuracy scores.

1 INTRODUCTION

For modern business intelligence, predicting sales numbers has become an important component [3]. Especially for retailers, as they are often faced with the challenge of choosing the right amount of inventory stock while trying to maximize profit. Retail sales data can be considered as time series data, which allows for the opportunity to apply certain statistical methods in order to produce sales forecasts. Accurate forecasts on the expected demand can provide new insights and assist retailers in making wise decisions and thus increase profit. Contrary, inaccurate forecasts could lead to several unwanted consequences, such as inventory shortage, product back-orders and unsatisfied customers.

For this research, a dataset containing information about the historical sales of a Walmart store was provided with the intention to predict future sales. This dataset was derived from the M5 Forecasting Kaggle competition [6]. The original dataset contained hierarchical sales data from multiple Walmart stores across three different US states (California, Texas and Wisconsin). The goal of the corresponding Kaggle competition was to produce daily predictions for sales over the next 28 days. Variables in the data consisted of attributes like product categories, store details and item levels. On top of that it also included some explanatory variables such as promotions and special events. For this research project only the data from the California based store CA_3 were provided. Simply put, the robust dataset could be used to improve the accuracy of forecasts.

Related Work

Forecasting retail sales using historical time series data is a problem many researchers in the field of data science have tried to tackle. Different relevant experiments have been performed in previous projects, like the comparison of different statistical models [7]. There are also studies that take a more detailed look at very specific machine learning models, such as Random Forest [2]. Moreover, the two most commonly used approaches for sales data forecasting problem are machine learning techniques that use artificial neural networks and statistical approaches that use Autoregressive Integrated Moving Average (ARIMA) models [2]. In domains and settings where there are little to no sudden behavioral changes, ARIMA models are known for producing highly effective and accurate financial time series forecasts [1].

Research Goals

Previous studies have examined the problem settings from various angles. The specific research goal of this project is to explore the effectiveness and performance of different forecasting techniques, in order to get better insights on the modelling techniques and hyperparameters that play an important role when it comes down to predicting the daily sales of Walmart products. To achieve the defined goal, the sales data time series forecasting problem was tackled by examining different statistical methods as well as some machine learning methods.

2 DATASET

The data were derived from the Kaggle competition [4] and contained information about the sales of the California based store CA_3 of the American retail organization Walmart Inc. The data consisted of multiple Comma-Separated Values (CSV) files, the daily unit sales were already split into separate training, validation and test sets. Firstly, the sales training dataset contained the historical daily unit sales of 149 unique products over 1913 days. This dataset was predominantly used as a training set for the models. All 149 unique products in the dataset belonged to the 'HOBBIES_2' category. In figure 1, a histogram can be seen which represents the daily unit sales data of product 7.

Next, the sales evaluation dataset contained the exact same data as the sales training dataset, complimented with daily unit sales information of an additional 28 days, resulting in a total of 1941 days. This sales evaluation dataset served as a validation set in order to measure the performance of models before using them on the test data. The test set was

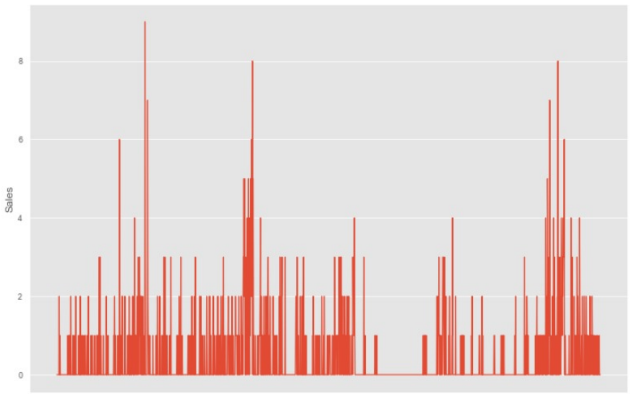

Figure 1: Daily unit sales for product 7

only available for measurement through the Kaggle page. Predictions for the test set had to be formatted in such a way that every product had a forecast for the 28 days after the 1941st day in the validation set.

In addition to the historical daily unit sales data, two explanatory datasets were provided. The sell prices dataset contained information about the prices of products throughout the entire time-span of the historical sales data. Figure 2 shows how the sell price of products such as product 20 can change over time.

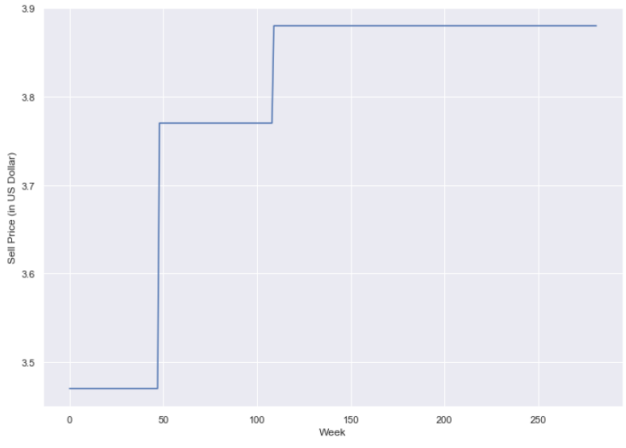

Figure 2: Sell price of product 20 throughout time

Lastly, the calendar dataset contained information about the dates on which products were sold. This dataframe contained information such as the actual dates and the specific weekdays. Besides that, the dataframe also contained information about special events which could affect product sales. These events were categorized under four different types, more information on these event types can be found in table 1. The calendar dataframe also contained the variable 'snap-CA', which represents the governmental program which allows people with a low income to receive discounts on certain days [8].

Table 1: Frequency of event types

Event type	Frequency	Examples
Religious	56	Easter, Christmas
National	52	Independence Day, Thanksgiving
Cultural	41	Cinco De Mayo, Halloween
Sporting	18	Superbowl, NBA finals

3 METHODS

The aim of the Kaggle competition was to predict the sales of products of the store CA_3 for the next 28 days. The retailer's challenge is to have the right amount of products in stock, to satisfy the customer demand without ordering too much inventory. In order to forecast the total stock, it is necessary to forecast the sales volumes for many products that have trends and seasonal patterns. The sales of the products can have sudden changes due to sell price fluctuation and upcoming events. Thus we need to find a forecasting method that allows for the presence of trends and seasonality, while remaining to be robust and adaptive for sudden changes.

The following forecasting methods were used:

- Single Naive
- Linear Model
- Moving Average (1)
- Moving Average (4)
- Autoregressive (1)
- Autoregressive (4)
- Auto ARIMA
- Prophet
- Prophet + holiday
- Long Short-Term Memory (LSTM)

Simple forecasting methods.
The firstly selected approach to select a forecasting model was essentially blind. The first models were some simple forecasting methods, mainly used to discover the data and to learn from the ways models can be used on the data. These forecasting methods were extremely simple, yet surprisingly effective. For the retailers challenge, four simple forecasting methods were used: two Moving Average methods with a order of 1 and 4, two auto regressive methods with a order of 1 and 4, a Single Naive model and finally a Linear model. All these methods only use historic data to forecast, except for the linear model which uses external data from the calendar data set. The linear model uses the predictor event and snap.

These methods would serve as benchmarks rather than the method of choice.

ARIMA models.
After the simple forecasting methods, ARIMA forecasting models were used. ARIMA uses past time series values (such as lags) and the derived forecasts errors from those values as predictors. ARIMA is a combination of autoregressive (AR) and moving average (MA) models. This model works well on the data, as the time series of the daily unit sales are stationary. The auto ARIMA function in R is used for each product to obtain the best fit. This fit is used to forecast the 28 days in the future. Only historic data was used to forecast with the ARIMA models.

Machine learning models.
Finally, two machine learning forecasting methods were used. Prophet was chosen since the individual products consist of univariate time series. Prophet automatically finds a good set of hyperparameters for the model, in effort to forecast the 28 days. Prophet can use other predictors like snap days and other event days as input to predict sales. The effect of adding these predictors will also be tested. Therefore two prophet models will be created, one prophet model will only use historic data. The other will use two predictors. The first predictor is if there is an event on that day and the second predictor is if there is a snap day at that day. The prophet model with extra predictors is called Prophet + Holiday. The last forecasting method was the Long Short-Term Memory (LSTM) model, a recurrent artificial neural network that can deal with long sequences of data. LSTM uses other predictors than only the historic data. Also data for the events is used, this input for LSTM is augmented with data when a specific day is close to an event.

Evaluation

The Sales Time Series Forecasting competition on Kaggle evaluates based on the Root Mean Squared Error (RMSE) of each product sold in the store. Thereafter the models were fine-tuned using the validation dataset. To measure the performance on the validation data, three different evaluation metrics were used. The first metric was the Root Mean Squared Error (RMSE). This metric was used to measure the results on the test set on Kaggle. The RMSE score is a scale-dependent metric and is commonly used to compare the forecasting errors of different models for a particular dataset.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2}$$

The Mean Absolute Error (MAE) was also used as a measurement metric for evaluation. The MAE is commonly used for comparing forecasting errors of different models, as it measures the errors between paired observations that express the same phenomenon.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

Lastly, the Symmetric Mean Absolute Percentage Error (SMAPE) was used as the third measurement metric. SMAPE is widely used to compare performance of algorithms used for the M3, M4 and M5 competitions.

$$SMAPE = \frac{100\%}{n}\sum_{t=1}\frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

Based on the accuracy and evaluation results, decisions were made to determine which forecasting method performs the best, which data should be included and how the hyperparameters should be adjusted. Afterwards, all the data including the validation data will be used to train the same models. The new models will be used to forecast 28 days in the future. Thereafter the forecast output was submitted to the Kaggle competition to get the RMSE score on the test predictions.

4 RESULTS

The best submission for the Kaggle competition, which consisted of predictions for the test set, had a RMSE score of around 0.895 [5]. To get a better image of this, if all predictions were set to 0, the RMSE score would be around 1.091. This best RMSE score was achieved by the Moving Average (4) model, closely followed by the Auto ARIMA model with a RMSE score of around 0.897.

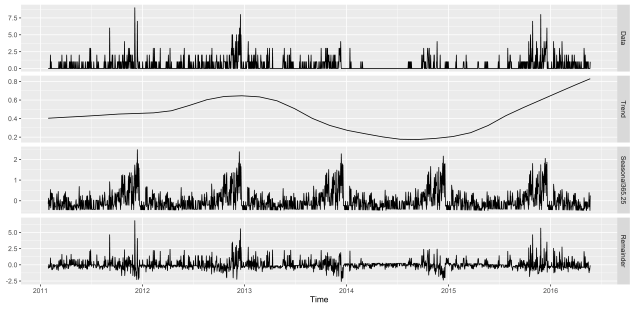

Figure 3: MSTL decomposition of the sales of product 7

In figure 3 a Multiple Seasonal and Trend decomposition using Loess (MSTL decomposition) is displayed of hobby product 7 from Walmart store CA_3. The decomposition shows a clear yearly seasonality, this can be seen in the third row of figure 3. There are no other seasonality patterns found

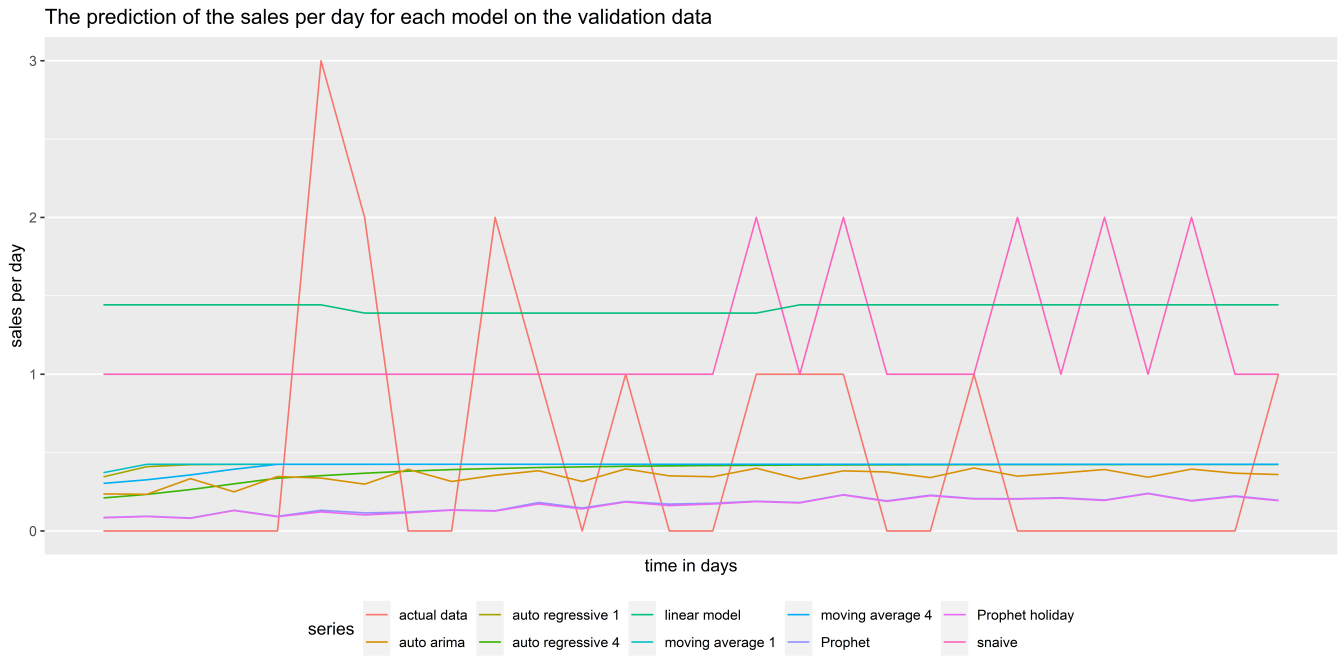

Figure 4: Prediction of all models on the validation data

in the data. Each year the sales increase at the end of each year with a peak around the end of December. This is most likely the effect of events like Christmas and Thanksgiving. In the second row of figure 3 the trend is displayed, the trend has a peak in 2013 but decreases until June 2014. After June 2014 the trend keeps increasing steadily. Lastly, the remainder is displayed in the last row of the MSTL decomposition. The remainder gets larger if the amplitude of seasonality is larger, which is around the end of each year. The height of the remainder is also dependent on the size of the trend, the larger the trend the higher the remainder.

Figure 4 shows a visual representation of the predictions on the validation data by the different models. In this diagram, the red line represents the actual sales of product 7. One can see how the linear model and the SNAIVE methods try to predict the actual sales per day, the other techniques calculated a more stable value throughout the days, leading to more accurate predictions overall.

Results of all models on the validation dataset can be seen in table 2. Note that these scores are not on the test data, as this was only available through the Kaggle competition and could only return the RMSE score. This factor combined with the unfortunate fact that not every model could be evaluated on the test data due to the restricted deadline lead to the validation dataset being used to compare the forecasts of every model. Despite the Moving Average (4) model having the best Kaggle score, it is actually the Auto ARIMA model which has the best RMSE score on the validation set. It can also be noted that the linear model and the SNAIVE model performed relatively bad on the validation set, as the RMSE values are worse than the baseline of 1.091 which is achievable by setting every prediction to 0. This bad performance is also noticeable for the MAE scores. However, unlike the RMSE scores, the lowest scores for MAE were achieved by the two Prophet models. As for the SMAPE scores, the two Prophet models received the highest scores while the linear model and the SNAIVE model got the lowest scores. It should also be noted that while the Moving Average (4) model had a better Kaggle score than the Moving Average (1) model, the Moving Average (1) model had lower scores for each of the metrics on the validation predictions.

The 28 day forecast of the best model, the moving average with order of 4 for product 7, is visible in figure 5. The forecast has a downward trend for a few days before it becomes a straight line. The straight line shows that the predicted sales for each day lies around 0.13 in terms of unit sales.

5 DISCUSSION

In this studies, multiple statistical models and machine learning models were applied to the daily sales time series data of the California based Walmart store CA_3. Most of the models showed reasonable performances. The linear model and the SNAIVE model performed significantly worse than the rest of the models. The Prophet models performed decently,

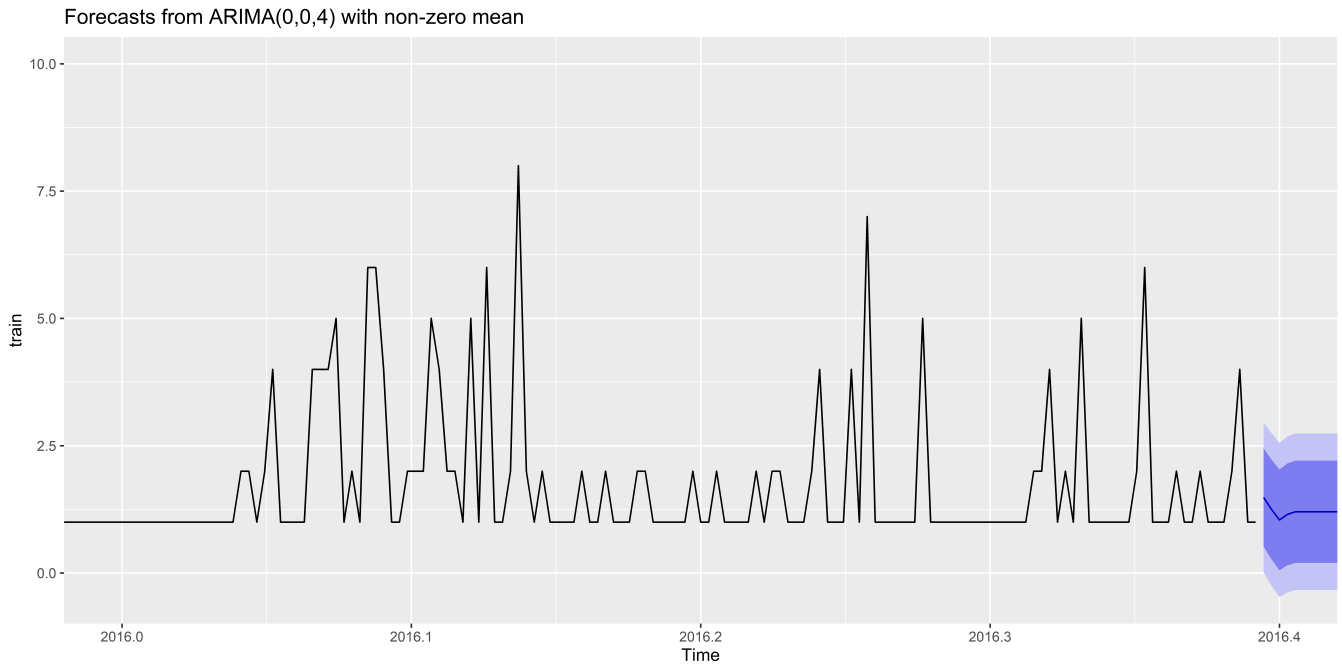

Figure 5: 28 day forecast of the sales with moving average (4) model of product 7

Table 2: Results on validation data

Model	RMSE	MAE	SMAPE
Linear model	1.408583	1.275668	1.567192
SNAIVE	1.828571	1.448466	1.57029
Prophet	0.9000961	0.5367983	1.868004
Prophet + holiday	0.896995	0.5373325	1.868288
Auto ARIMA	0.8483332	0.5746101	1.743956
Moving Average (1)	0.8656341	0.5844628	1.770545
Moving Average (4)	0.8661819	0.5846329	1.771615
Autoregressive (1)	0.8692797	0.5845232	1.771428
Autoregressive (4)	0.9687375	0.5826901	1.768471
LSTM	0.904209	0.6028458	1.786243

but ultimately it were the ARIMA models which lead to the best scores. The Moving Average (4) model attained the best Kaggle score of around 0.895, closely followed by the Auto ARIMA model with a score of around 0.897.

Limitations

The biggest limitation of this research is the fact that the available data wasn't used optimally. Little attention was spent to analyze aspects such as outliers and seasonality patterns. This could have possibly had a direct effect on the performance of the models. For instance, the information from the sell prices and the event information in the calendar dataframe weren't used for every model. This means that only a limited amount of data was used for the forecasting process, which is unfortunate. The main reason why this happened was the limited scope of time in which the project was supposed to be performed. Lastly, only a few machine learning models were part of this research. Previous research has shown the potential of machine learning models such as Random Forest or K Nearest Neighbor for settings like this involving sales data. Thus it is unfortunate that more advanced models like these were not furthermore inspected.

Theoretical Implications

The results of this research have shown how several different models differed from each other when it comes down to projecting forecasts for sales time series data. Simple models such as the linear model and the SNAIVE model appear to be too simple for a task like this, this could be caused by the complexity of the context. More advanced models like Prophet perform better than these simple models, but they don't appear to be the perfect choice. A machine learning model such as LSTM doesn't seem to result in excellent forecasts either, given the circumstances in which it was tested. Although the Moving Average (4) model had a better Kaggle score than the Auto ARIMA model, the case was in reverse for the validation predictions. A possible explanation for this is that the Auto ARIMA method, which chose the best ARIMA model for every product, is more prone to statistical

bias, resulting in a slightly overfitted model. Nevertheless, the results show that ARIMA models which incorporate the moving average of historical sales data produce reasonably accurate forecasts.

Practical Implications

Besides the theoretical implications of the results, it is also important that the results can be interpreted in a practical way that is of added value for involved stakeholders, such as California based Walmart CA_3 in the case of this project. For the store it is important that forecasts are accurate, because an accurate inventory stock should satisfy the customer demand optimally. Looking at the forecasts by the Moving Average (4) model, they are very accurate when looking at the 28 days as a whole. But over the individual days, they might not be perfect since they rely on an average value. This means that the forecasting method is not very resistant to sudden changes in customer behavior. Nonetheless, for an entire month these forecasts can be relevant for retailers.

Future Work

For future research we have a number of interesting suggestions. Firstly, we suggest to further investigate the application of machine learning algorithms on daily sales time series data. Although we were not able to try this out, machine learning models such as Random Forest and K Nearest Neighbor might be interesting for a case like this. Secondly, we highly suggest to explore the remaining data that was provided but not used for modelling in this study. Events like Christmas and Cinco De Mayo might be of added value for the forecasting methods, just like the fluctuation of sell prices throughout time. It could also be interesting to expand the scope of data, exploring other external datasets such as weather information which might be interesting for a case study like this.

6 CONCLUSIONS

This paper dealt with the implementation and comparison of multiple forecasting methods, ranging from statistical models such as ARIMA to machine learning models like Prophet, on the historical daily sales data of the California based Walmart store CA_3. Results lead to the findings that ARIMA models like Moving Average (4) and Auto ARIMA deliver the best scores for this case study. Simple forecasting methods like linear models and SNAIVE models appear to be too inadequate for a context like this. Advanced machine learning models like Prophet and LSTM performed decent, but it might be of interest to look into other models such as Random Forest. Ultimately, these are only the results within the scope of our own applied methods, but the gained information could be of added value to any retailer, big or small, in order to produce more accurate forecasts on future sales for higher commercial profits.

ACKNOWLEDGMENTS

We would like to thank dr. Cristian Rodriguez Rivero for providing guidance and assistance throughout the process of the final project. Many thanks as well to the rest of the staff for the course Applied Forecasting in Complex Systems 2020-2021: Sara Mahdavi Hezavehi, Reshmi Gopalakrishna Pillai and Olivier Sprangers.

CONTRIBUTIONS

Each member of our team contributed to the project equally. The models were developed together while some work was done separately. Theoretical aspects and the writing of the project report was done as a group effort.

REFERENCES

[1] Adebiyi A Ariyo, Adewumi O Adewumi, and Charles K Ayo. 2014. Stock price prediction using the ARIMA model. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*. IEEE, 106–112.

[2] N. Elias and Seema Singh. 2018. FORECASTING of WALMART SALES using MACHINE LEARNING ALGORITHMS.

[3] John T Mentzer and Mark A Moon. 2004. *Sales forecasting management: a demand management approach*. Sage Publications.

[4] University of Amsterdam. 2020. *AFCS2020 - Sales Time Series Forecasting*. Retrieved December 16, 2020 from https://www.kaggle.com/c/sales-time-series-forecasting-ca-afcs2020/overview

[5] University of Amsterdam. 2020. *AFCS2020 - Sales Time Series Forecasting: Leaderboard*. Retrieved December 19, 2020 from https://www.kaggle.com/c/sales-time-series-forecasting-ca-afcs2020/leaderboard

[6] University of Nicosia. 2020. *M5 Forecasting - Accuracy*. Retrieved December 16, 2020 from https://www.kaggle.com/c/m5-forecasting-accuracy/overview

[7] James J. Pao, Danielle Sullivan, and D. G. sullivan. 2017. Time Series Sales Forecasting.

[8] Walmart. 2020. *Customers Using SNAP Benifits Have More Ways to Shop*. Retrieved December 17, 2020 from https://corporate.walmart.com/newsroom/2020/04/13/customers-using-snap-benefits-have-more-ways-to-shop