DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2018

A comparative study between LSTM and ARIMA for sales forecasting in retail

AJLA ELMASDOTTER

CARL NYSTRÖMER

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

A comparative study between LSTM and ARIMA for sales forecasting in retail

AJLA ELMASDOTTER, CARL NYSTRÖMER

Bachelor in Computer Science
Date: June 6, 2018
Supervisor: Pawel Herman
Examiner: Örjan Ekeberg
Swedish title: En jämförande studie mellan LSTM och ARIMA för prognostisering av försäljning i livsmedelsbutiker
School of Electrical Engineering and Computer Science

Abstract

Food waste is a major environmental issue. Expired products are thrown away, implying that too much food is ordered compared to what is sold and that a more accurate prediction model is required within grocery stores. In this study the two prediction models Long Short-Term Memory (LSTM) and Autoregressive Integrated Moving Average (ARIMA) were compared on their prediction accuracy in two scenarios, given sales data for different products, to observe if LSTM is a model that can compete against the ARIMA model in the field of sales forecasting in retail.

In the first scenario the models predict sales for one day ahead using given data, while they in the second scenario predict each day for a week ahead. Using the evaluation measures RMSE and MAE together with a t-test the results show that the difference between the LSTM and ARIMA model is not of statistical significance in the scenario of predicting one day ahead. However when predicting seven days ahead, the results show that there is a statistical significance in the difference indicating that the LSTM model has higher accuracy. This study therefore concludes that the LSTM model is promising in the field of sales forecasting in retail and able to compete against the ARIMA model.

Sammanfattning

Matsvinn är ett stort problem för miljön. Utgångna produkter slängs, vilket implicerar att för mycket mat beställs jämfört med hur mycket butikerna säljer. En mer precis modell för att förutsäga försäljningssiffrorna kan minska matsvinnet. Denna studie jämför modellerna Long Short-Term Memory (LSTM) och Autoregressive Integrated Moving Average (ARIMA) i deras precision i två scenarion. Givet försäljningssiffror för olika matvaruprodukter, undersöks ifall LSTM är en modell som kan konkurrera mot ARIMA-modellen när modellerna ska förutsäga försäljningssiffror för matvaruprodukter.

Det första scenariot var att förutse försäljningen en dag i framtiden baserat på given data, medan det andra scenariot var att förutse försäljningen varje dag under en vecka i framtiden baserat på given data. Genom att använda måtten RMSE och MAE tillsammans med ett T-Test visade resultaten av studien att skillnaden mellan LSTM- och ARIMA-modellen inte var av statistik signifikans i fallet då modellerna skulle förutsäga försäljningen en dag i framtiden. Däremot visar resultaten på att skillnaden mellan modellerna är av signifikans när modellerna skulle förutsäga försäljningen under en vecka, vilken implicerar att LSTM-modellen har en högre precision i detta scenario. Denna studie drar därmed slutsatsen att LSTM-modellen är lovande och kan konkurrera mot ARIMA-modellen när det kommer till försäljningssiffror av matvaruprodukter.

Contents

Chapter 1

Introduction

Environmental problems are becoming more prominent and the need to tackle these is becoming more and more urgent. One such problem is food waste. The processes of growing or raising food, shipping it to the stores, cooling it down and then throwing it away all have a big impact on the environment [1]. Grocery stores throw away thousands of tons of food every year due to products passing their expiration date [2] and because of this they have a big, negative, environmental impact. In addition to the environmental effects, there is an obvious economical implication as the stores has to throw away food that is not sold. Ordering more accurate volumes of food comparing to how much is sold, would benefit the grocery store economically as well as have a positive impact on the environment due to reduced food waste.

As sales can be measured over time intervals, for example on a daily basis, an approach to predict future sales is to look at past sales and their patterns over time to then re-use these patterns for prediction. Sales forecasting is not alone with this kind of problem formulation. For example, stock market prediction [3] and weather forecasting [4] uses historical data to predict how future data will behave. This is a so called time-series forecasting problem and there exists a lot of research in this domain. Formulating the sales prediction problem as a time-series forecasting problem is therefor suitable. [5]. Models such as Autoregressive Integrated Moving Average (ARIMA) [6], [7] and Artificial Neural Networks (ANN) [8], [9] are a few of several methods used in the forecasting domain. Given its prediction accuracy and performance however, ANN approaches have been given a lot of attention [8] especially in recent year due to the rise of AI-research. In particular the Long Short-Term Memory (LSTM) model has been widely used in the forecasting domain in general as it has the capability to remember information far back in the time-series [5], [8], [10]. Hence this study will approach the sales prediction problem using the LSTM and the ARIMA model.

1.1 Research Problem

As previous research suggest that Machine Learning can provide successful and accurate models for forecasting sales of products and thus reducing the expenses for the stores [6], the purpose of this study is to examine state of the art time-series forecasting models used for predicting sales in retail. There are multiple methods used for forecasting prediction, however this study will primarily focus on LSTM networks and the traditionally used ARIMA model. Since the ARIMA model is a linear model, the LSTM model is expected to handle the non-linearity of sales forecasting with higher prediction accuracy. However, ARIMA is well-known for its prediction accuracy, making it interesting to compare the models in the sales forecasting domain. Choosing the more accurate model, giving a more accurate prediction of sales, could give the stores an opportunity to reduce the food waste, hence reducing both economical and environmental impact, by ordering only as much food as necessary for the given time period. The study intends to extend previous research made ([5], [6], [11]) by comparing the LSTM model against the ARIMA model. Since the ARIMA model is widely used in the forecasting domain, this study aims to research if LSTM is a model that could potentially compete against ARIMA in the field of sales forecasting in retail.

1.2 Scope

As there exists multiple state-of-the-art methods for time-series forecasting, the study is limited to comparing LSTM to a frequently used model: ARIMA. The models will be compared in the domain of sales forecasting in retail, using two scenarios: The first scenario being the models predicting one day ahead using given data and the other scenario being the models predicting each day one week ahead, seven days ahead, using given data. Data from an Ecuadorian grocery store was used to conduct the experiments [12], where the data about specific holidays and weather conditions were excluded from this study. As only one dataset is used for this study, it can be considered a limitation.

1.3 Thesis Outline

In chapter 2, the Background, related work and relevant theories will be presented. Following in chapter 3, Method, a description of the approach taken to answer the stated research question is presented. Chapter 4 will present the results of the study. Lastly, chapter 5 and 6 will discuss and conclude the results, limitations of the research and possible future research.

Chapter 2

Background

In the following sections background relevant for the study will be explained and discussed. In the end of this chapter, previous research relevant to our problem statement will be presented.

2.1 Time-series Forecasting

A time-series is a sequence of observations measured sequentially through time. These observations are either continuous through time or documented at a discrete set of equal time intervals. Continuous time-series, such as measurement of brain activity, are usually analyzed by sampling the series at equal time intervals to give a discrete time-series [13, p. 1]. As the report will focus on discrete sets of data, the continuous time-series will not be further analyzed in this report. The observations at different time-steps in the series often correlate in different ways. When successive observations are dependent, future values may be predicted from past observations. Time-series analysis is the research area where these correlations and dependencies are analyzed. The correlation could be the order the data that is recorded, the linearity of the model, repeating patterns etcetera. Time-series analysis provides techniques to analyze the data [14] [15, p. 100]. Time-series forecasting involves using the observations of the time-series, together with the time-series analysis, to establish a model describing the dependencies. This model is then used to predict future values in the series. The forecasting predicts future values using only past information [16].

Sales of a particular product in successive months is an example of a time-series. By using time-series forecasting models it is possible to predict future sales of the product [13, p. 1]. There are multiple models used for time-series forecasting, each used in different domains and context. ARIMA, ANN and Support Vector Machines are a few examples [7], [17]. Most models can be classified

as linear or non-linear. Linear models are limited by their assumption of a linear behaviour, whereas non-linear models can fit a more complex function's shape more closely. A grocery store's sales, for example, may fluctuate greatly depending on season, holidays, days following pay-day and so on. This implies that the grocery store's sales will not be a linear function. This might raise the question if there are benefits using non-linear models trying to predict non-linear functions. It can prove beneficial, but this is not always the case. Linear models, such as ARIMA, are well-known for their accuracy and flexibility in handling several different types of time-series, including non-linear ones [6], [7]. Research does however suggest that non-linear models, particularly ANNs, perform better than linear models to problems concerning sales. ANNs have also shown promising results in areas such as prediction and pattern recognition [6], [10], [18]. In particular research suggests that the LSTM model performs better than the traditional models in the time-series forecasting domain [8], [11].

2.2 Forecasting models

2.2.1 LSTM

To understand what the LSTM, Long Short-Term Memory, model is and how it works, ANNs and RNNs need to be briefly described.

ANN

ANNs are loosely inspired by the neural connections in our brain, trying to mimic the neural pathways and their behaviour. ANNs, being robust and self-adapting, generally provides satisfying solutions to non-linear problems not easily implemented implicitly, and solving tasks such as speech recognition, natural language processing and forecasting [10]. The network has so called layers between input and output, where increasing the number of layers also increases the complexity of the network. By feeding the network with known data, the network is trained by weight selection in relation to the desired output of the input. The weights are scalars adjusted by the network to reduce the error between the, by the network concluded, output and the actual desired output using the gradient of error with regards to the weights. These types of networks are however not suitable for sequential data. As the network has no memory of previous time-steps, it cannot model dependencies, such as the ones described in section 2.1, making it hard to analyze the sequential data. Therefore a memory of some sort is desired.

RNN

RNNs, recurrent neural networks, chain together multiple layers of networks, where information from previous time-steps, in addition to the output, is carried over to future time-steps. As the input and parameters from each layer is processed, output from previous layers are taken into consideration, giving the network a form of memory [19].

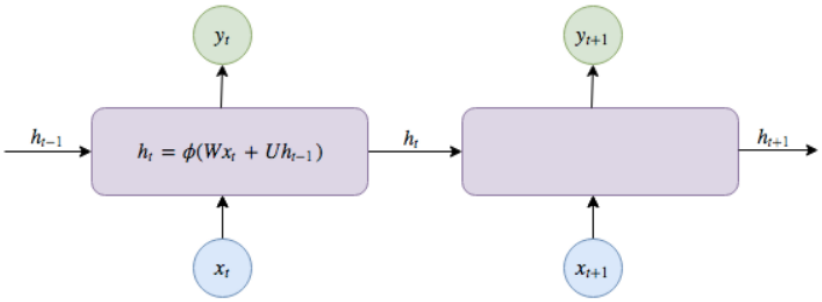

Figure 2.1: An RNN where W is a weight matrix and U is a matrix consisting of so called biases and ϕ is an activation function, making the network non-linear

As a time-dimension has to be taken into consideration for RNNs the gradient can get relatively complex, resulting in that information form previous time-steps either start to vanish or get significantly amplified. These phenomenons are called a vanishing or exploding gradient respectively. Because of these phenomenons, it might be difficult for the network to remember information from previous time-steps correctly if it lies far back in sequence [10]. LSTM is a solution to these problems.

LSTM

The LSTM networks implements a gated cell to store information, similar to computer memory. Unlike the aforementioned networks, the LSTM cells also learn when to allow reads and writes of information from previous time-steps [11], [20]. Hence the LSTM model solves the problem of a vanishing or exploding gradient and makes it possible for the network to correctly remember information far back in the sequence [8], [10].

Figure 2.2 presents the inner components of the LSTM-cell. One cell handles one time step worth of data and pass along chosen information to the next cell at

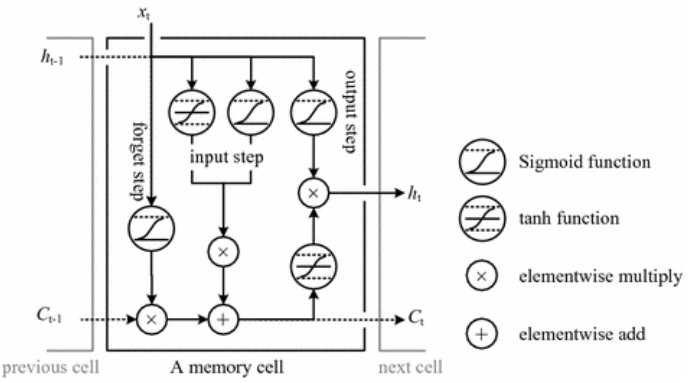

Figure 2.2: An LSTM Cell and its components.
X_t represents input at time-step t
h_t is the output for one time step
C_t is additional dependencies remembered from previous time-steps which is added to the ordinary output
Source : [5]

the next time step, depending on the gates. Figure 2.3 depicts an example where three consecutive time-steps are used to predict the fourth. Notice the additional output C_t compared to 2.1, which only uses h_t.

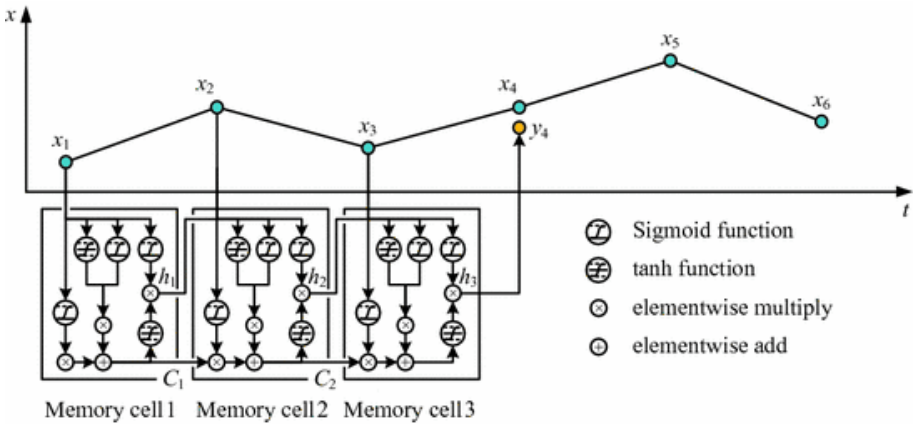

Figure 2.3: Depicts an example of how time-series forecasting might look like using LSTM-cells
Source : [5]

2.2.2 ARIMA

The Autoregressive Integrated Moving Average (ARIMA) model has been a popular approach to forecasting problems. A linear combination of past values and

errors is used to make predictions of future values [3]. The general forecasting equation is given by:

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} - \theta_1 e_{t-1} - ... - \theta_q e_{t-q} \tag{2.1}$$

Where $\hat{y}_t$ is the value, e_t is the error at time t, ϕ_i and θ_j are coefficients, μ is so called white noise and p and q are the number of autoregressive (AR) terms and moving average (MA) polynomials respectively. The AR part implies that the variable (y_k) is regressed on its prior values while the MA part indicates that the regression error is a linear combination of previous errors. The I indicates that the regular data values have been replaced with the difference between the data and the previous values [3], [21].

The model is well-known for its forecasting accuracy as well as its adaptability to different types of time-series problems [7]. However, as it is a linear model, ARIMA suffers from limitations when handling non-linear problems such as forecasting [3], [7] as it is expected to perform better when the time periods are short-term rather than long-term [22], [23].

2.3 Previous Research

This section will present some relevant previous research that has been done in the field of time-series forecasting. Several previous studies suggest that LSTM is a promising model in the field of forecasting. One study also confirms that ARIMA is used in sales forecasting with good results. The studies presented in this section has also been used as inspiration in terms of methodology.

2.3.1 LSTM-RNN Forecasting Model For Cloud Datacenters

In [11] the authors state that using hte LSTM model could address the issues cloud systems have, being fragile and expensive when dealing with issues such as dynamic resource scaling and power consumption. The authors state that if it would be possible to determine a server's accurate future workload, the resources can be adjusted according to demand and hence both retain quality of service and shorten power consumption. The study therefore aims to conclude if LSTM is a model to be used for the forecasting problem [11].

The authors mention that LSTM networks perform well in forecasting problems, requiring that information is preserved and remembered, making it a reason that LSTM is the natural choice of model to use for [11]. Using three different datasets, the authors normalize the data in the range $[0, 1]$ and use Python along with the Keras library to implement the model. The model is evaluated by comparing it to other prediction methods based on so called black hole and back propagation learning algorithms using the mean squared error [11].

Concluding that the LSTM model is promising for forecasting problems, the authors encourage further exploration of the domain. The results shows that LSTM cleary outperforms both approaches the model was compared against making it relevant not only for cloud datacenters but also for forecasting problems in general. The authors also state that the prediction model can help in smart resource scaling decisions and in reducing the number of active machines [11].

2.3.2 LSTM to Sales Forecasting in Retail — A Case Study

In the study [5] the authors test a LSTM network on real sales data. The model was built to experiment on 66 products with 45 weeks worth of data. The model predicts sales in week level, using four consecutive weeks to predict the sales of the fifth week. Also, normalization between $[0, 1]$ is used as well as iterating the experiment ten times to get the average mean square error to calculate the performance [5].

The results obtained in [5] show that only a fourth of the products have relatively low forecasting errors. However, the authors motivated that the limitation in form of lack of data, lack of long-term seasonality as only 45 weeks worth of data was accessible and using one LSTM network for all products, could have affected the results immensely. The authors argue that the network was not optimized with the respect to characteristics of the specific products as it was a generalized implementation. Nonetheless the authors conclude that the LSTM network still shows potential and is a domain that should be further looked at, considering the seasonality of products and data from a longer time period, among other things, when developing a new network [5].

2.3.3 Ensembles of LSTM for time-series Forecasting

The authors of study [8] mention that a single LSTM network that is trained with a particular dataset is very likely to perform poorly on an entirely different time-series unless rigorous parameter optimization is performed. However, the authors mention that LSTM is a very successful in the forecasting domain. Hence, the authors use a so called stacking ensemble approach where multiple LSTM networks are stacked and combined to provide a more accurate prediction, aiming to propose a more generalized model to forecasting problems [8].

By conducting the experiment on four different forecasting problems, the authors of [8] concluded that the stacked LSTM networks outperformed the regular LSTM networks as well as the ARIMA model used for comparison. They do nonetheless state that there exists single LSTM models that performs better than the stacked model in terms of the evaluation measure RMSE. The authors also conclude that the general quality of the ensemble method studied could be in-

creased through tuning the parameters for each individual LSTM. However the heavy tuning of parameters for the LSTM networks is, what the authors suggest, one reason to why the individual LSTM network might perform poorly when it is used for a different dataset than the one it was trained with. The authors hence suggest that the suggested ensemble method might be a better choice for forecasting problems, to both reduce the need to heavily optimize parameters and to increase the quality of the predictions [8].

2.3.4 Sales Forecasting with ARIMA

The authors of study [22] use ARIMA models to predict the number of newspapers sold for a real case study of a newspaper company in Surakarta. As newspapers are supplied with demand, the switch from printed newspapers to electronic news have caused that many newspapers are returned. The authors do mention that ARIMA is used for forecasting short term, making the result of the study invalid for long term modeling. The authors select an appropriate ARIMA model for the sales forecasting of newspapers. Particularly they found the best parameters to use for the ARIMA model [22].

Using evaluation measures such as RMSE and MAE as well as mean absolute percentage error, the results of [22] shows that the best ARIMA model was with the parameters $(1, 1, 0)$ without constant. Furthermore the authors conclude that the suggested ARIMA model can be used in practice for short-term forecasting of the sales of the newspapers as well as to reduce the amount of newspapers that are not sold [22].

Chapter 3

Methods

3.1 Data Formatting

Data from an Ecuadorian grocery chain was collected from a sales forecasting competition on Kaggle. To see exactly how the data is formatted, please refer to [12]. Since the data is from the grocery chain, and not only an individual store, there exists sales data from several stores between 2013 to 2017. Specific holiday data was excluded. Over four thousand products are logged in the dataset, for simplicity the ten most sold products were used for analyzing the models. Around 80 000 rows worth of data existed for each of the most occurring products. To create a time-series, all sales from each store were summed up for each date for each unique product. For the dates where no sales existed for a particular product a zero was added to create a continuous time-series. Furthermore, only the sales parameter was used as a feature in the models. The data can be expressed as $[s_0, ..., s_t]$ where s is the sales for a product and the subscript denotes which day, t is the total amount of days logged. To be able to utilize the data for training, it had to be split up into time windows as follows:

$$
\begin{gathered}
[s_0, ..., s_d, s_{d+1}, ..., s_{d+p}] \\
[s_1, ..., s_{d+1}, s_{d+2}, ..., s_{d+p+1}] \\
[s_2, ..., s_{d+2}, s_{d+3}, ..., s_{d+p+2}] \\
\vdots \\
[s_{t-d-p}, ..., s_{t-p-1}, s_{t-p}, ..., s_t]
\end{gathered}
$$

Where d is the number of previous time steps the model looks at to predict p future time steps. The first d columns in each row are used for training and the rest of the columns are the labels for each corresponding row.

3.2 Model implementations

3.2.1 Baseline model

One baseline model was implemented using a naive forecasting strategy. The model simply looks at the each value in the time-series and predicts the same value for the following time-step. The baseline model simply gauges whether the ARIMA-model and LSTM-models can predict better than a naive forecasting strategy.

3.2.2 LSTM-implementations

Two LSTM-models were implemented following the two previously mentioned scenarios. $LSTM_1$ and $LSTM_7$, which predicts 1 and 7 days ahead respectively. One which predicted only one day ahead, and one which predicted the following seven days. Both models were implemented using the Python library Keras. 80% of the data was used as training data by both models and the remaining 20% of the data was used as the test set. The data was also normalized and differenced.

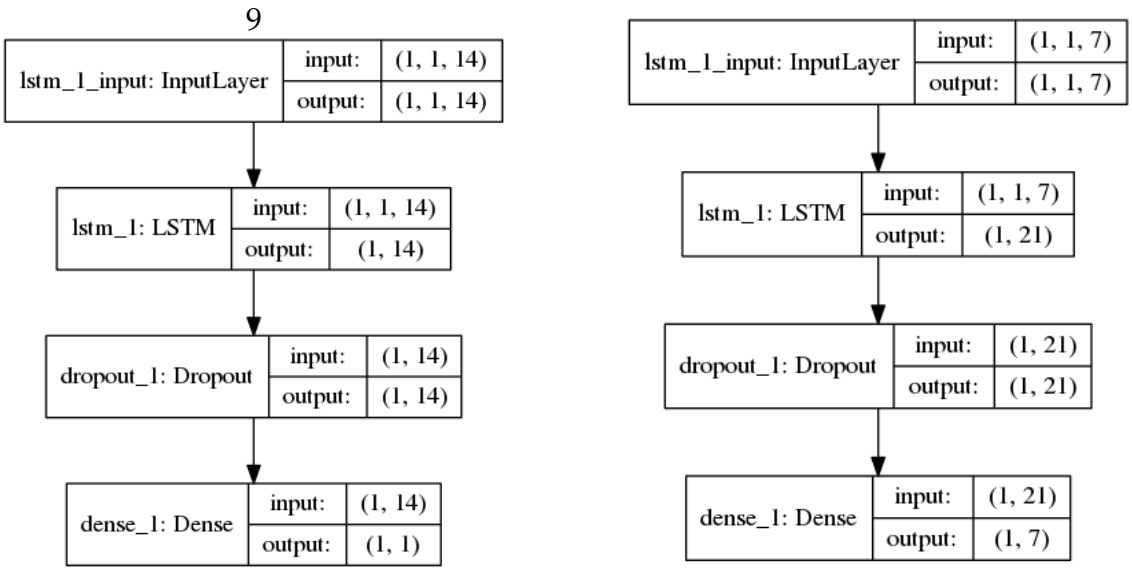

Figure 3.1: Model summaries: The figure shows how the layers and the input/output-shapes looks like in the network. Using only one LSTM-layer provided the best results. The left model shows the structure for $LSTM_1$, the right shows $LSTM_7$ as can been seen by the last output layer

3.2.3 ARIMA-implementation

For the ARIMA-implementation the statsmodel library was used within Python. Similarly to the LSTM-model 80% of the data was used for training and 20% of the data was used for testing. The lag-order, similar to the previous days looked at in the LSTM-model for each prediction, was also set to 7 days, the degree of differencing was set to 1 and the moving average was set to 0. The same model was used for both prediction scenarios.

3.2.4 Hyperparameters

To decide which values to use as hyperparameters - the values which customizes the models - a so called grid search was performed. A grid search is a commonly used technique for optimizing hyperparameters in machine learning models. [24] It is a form of exhaustive search where a large combination of hyperparameters are tested and the ones which yields the best results are used for the final model. As both the MAE- and RMSE error measures were part of this test, the analysis had to take both evaluation measures into consideration when choosing the most optimal parameters. Therefore the top 10 results for both RMSE and MAE were chosen and compared. Through the grid search the following values were chosen:

Parameters for LSTM_1

Parameters include those of sequence length, the parameter that decides for how long the LSTM method should remember information, dropout, the parameter that counteracts overfitting and parameters that controls the training, which can be seen in table 3.1

	chosen value	**grid search interval**
Sequence Length	14	[1, 7, ... , 28]
Dropout	0.2	[0.1, 0.2, ... , 1.0]
Epochs	15	[1, 5, ... , 50]
Neurons output layer	14	[1, 7, ... , 28]
Activation Function	tanh	[relu, tanh, sigmoid]
Optimization Function	adam	[adagrad, adam, RMSprop]

Table 3.1: Values used for the respective parameters in the LSTM_1 model as well as the grid search domain for each parameter is presented.

Parameters for LSTM$_7$

The values used for the LSTM$_7$ model can be seen in table 3.2.

	chosen value	**grid search interval**
Sequence Length	7	[1, 7, ... , 28]
Dropout	0.3	[0.1, 0.2, ... , 1.0]
Epochs	25	[1, 5, ... , 50]
Neurons output layer	21	[1, 7, ... , 28]
Activation Function	tanh	[relu, tanh, sigmoid]
Optimization Function	adam	[adagrad, adam, RMSprop]

Table 3.2: Values used for the respective parameters in the LSTM$_7$ model as well as the grid search domain for each parameter is presented.

Parameters for ARIMA

The values used for the ARIMA model can be seen in table 3.3.

	chosen value	**grid search interval**
P	7	[1, 7, ... , 28]
D	1	[1, 2, 3]
Q	0	[0, 1, 2]

Table 3.3: Values used for the respective parameters in the ARIMA model as well as the grid search domain for each parameter is presented.

3.3 Evaluation Measures

To compare the performance of LSTM and ARIMA, two different evaluation measures were used: Mean absolute error (MAE) and root mean square error (RMSE). A lower value for both measures implies better accuracy. Using F_t as the forecast value (the prediction), A_t as the actual value and n as the number of time steps, RMSE and MAE can be defined as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}} \tag{3.1}$$

$$\text{MAE} = \frac{\sum_{t=1}^{n} |A_t - F_t|}{n} \tag{3.2}$$

The evaluation measures have been used in previous research ([25], [26]) to measure the prediction performance of different models, including LSTM and ARIMA [10], [20], indicating that they are reliable.

3.4 T-Test

To validate the results, four t-tests were performed. The t-test is used to compare the means of two populations [23], in this case the LSTM model and the ARIMA model. One t-test was performed for each prediction model - both for the 1-day-ahead and 7-day-ahead prediction scenario. H_0 was that the LSTM model has a smaller prediction error than the ARIMA model, making the alternative hypothesis (H_a) that the LSTM model has a prediction error as small as or greater than the ARIMA model. For each t-test the same 10 products were used to generate data. The null hypothesis for each of the four t-tests performed are presented in table 3.4.

$\mathbf{H}_0$	**Evaluation Measure**
$\text{LSTM}_1 < \text{ARIMA}_1$	RMSE
$\text{LSTM}_7 < \text{ARIMA}_7$	RMSE
$\text{LSTM}_1 < \text{ARIMA}_1$	MAE
$\text{LSTM}_7 < \text{ARIMA}_7$	MAE

Table 3.4: The null hypothesis for each t-test performed on the different scenarios and the evaluation measurement used in the particular t-test

Chapter 4

Results

The following results present how the LSTM and ARIMA model perform when predicting sales in two scenarios. In the first scenario the models predict the sales for one day ahead, while they predict everyday for a whole week, 7 days ahead, in the second scenario. For the models predicting seven days ahead for each time-step (the ones with a subscripted seven) the difference $(p_1 + ... + p_7) - (s_1 + ... + s_7)$, where p_n is the prediction for n time-steps ahead and s the actual sale for n time-steps ahead, was used in the error measurements formulas, presented in section 3.3, for each time-step.

Each model predicted the sales for the ten products used as datasets, resulting in a error value in both RMSE and MAE. To compare the models and their error-values, a t-test was conducted, comparing if there exists a significant difference between the models. The difference between the models for each individual product, the overall difference in error as well as the p-value from the t-test is presented in table 4.1 and 4.2 for the two respective scenarios. For both the individual products as well as for the total difference in error, a negative value in the tables indicates that the LSTM model has a lower error, while a positive value indicates that the ARIMA model has a lower error.

In table 4.1 the results for the one-day-ahead prediction scenario is presented. As seen by the total difference in error, both RMSE and MAE indicate that the LSTM model has a lower error overall. The total difference in error for MAE is however affected by an anomaly. Nonetheless, the p-value implies that the difference is not of statistical significance for both evaluation measures, indicating that the LSTM model is not necessarily better than the ARIMA model in the scenario of one-day-ahead prediction.

	$\mathbf{LSTM}_1 - \mathbf{ARIMA}_1$	
Product	**RMSE**	**MAE**
1	5	12
2	-4	1
3	-30	-18
4	11	-239
5	-22	-14
6	-3	15
7	-7	-4
8	-12	-5
9	-9	-24
10	18	9
Total	-53	-267
p-value	0.1375	0.1468

Table 4.1: The table displays the results for the 1-day-ahead prediction scenario where the first 10 rows present the difference between the LSTM and ARIMA model for each respective product. The total error, over all products, is presented for both evaluation measures as well as the p-value generated by the t-test. The t-test hypothesis H0: LSTM has a lower error than ARIMA.

In table 4.2 the results for the seven-days-ahead prediction scenario is presented. As seen by the total difference in error, both RMSE and MAE indicate that the LSTM model has a lower error overall in this case as well. It is also observed that both evaluation measures give a similar total difference in error. As opposed to the one-day-ahead prediction scenario however, the p-value in this case implies that the difference between the models is of statistical significance for both evaluation measures. This indicates that the LSTM model in the seven-day-ahead prediction scenario does have higher accuracy than the ARIMA model.

Figure 4.1 shows the average result for the models in both the one-day and seven-days ahead prediction scenario for both error measures. There is an observed tendency when considering all three models, where the baseline model seems to perform worse than both the ARIMA model and the LSTM model.

	$\mathbf{LSTM_7 - ARIMA_7}$	
Product	**RMSE**	**MAE**
1	7	5
2	6	-3
3	-35	-40
4	-20	-27
5	-39	-28
6	-14	-1
7	-18	-14
8	-18	-14
9	-19	-27
10	-11	-16
Total	-161	-165
p-value	0.0037	0.0025

Table 4.2: The table displays the results for the 7-days-ahead prediction scenario. The first 10 rows present the difference between the LSTM and ARIMA model for each respective product. The total error is presented for both evaluation measures as well as the p-value generated by the t-test. The t-test hypothesis H0: LSTM has a lower error than ARIMA.

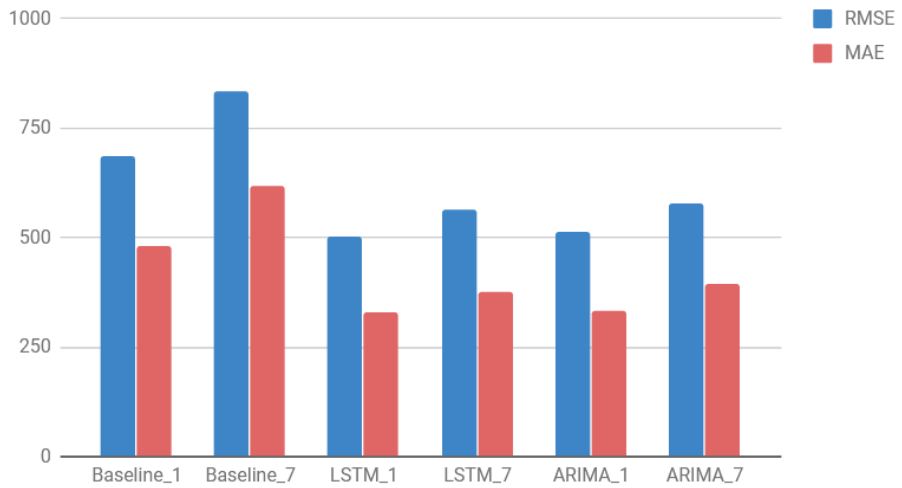

Figure 4.1: The figure displays the average error, estimated over products 1 to 5, for each prediction model. Both 1-day ahead and 7-days-ahead prediction scenarios are included.

Chapter 5

Discussion

5.1 Key findings

In this study two different prediction models' accuracy in sales prediction were compared. Investigating two scenarios, one-day-ahead prediction and seven-days-ahead prediction, the results of the overall difference in error between the models indicate that the LSTM model could have higher accuracy than the ARIMA model. However only the seven-days-ahead prediction scenario was proven statistically significant according to the conducted t-test. Nonetheless, the one-day-ahead prediction came close to the p-value threshold which indicates that with further research a different or more conclusive result can be found for the one-day-ahead prediction scenario.

5.2 Result Discussion

Due to the complexity of the problem we assumed that the non-linear LSTM model would outperform the linear ARIMA model in both scenarios as non-linear models are expected to render a higher accuracy for complex problems [6], [7]. However, even though the results shows that LSTM could be better, it is far from being superior as the one-day-ahead prediction scenario results were not statistically significant according to the t-test. By increasing the prediction length and complexity in future work, it would probably be possible to rule out if LSTM is better at handling complexity, which previous research suggests [8], [11].

The number of previous values looked at, used as input for the LSTM, did not seem to matter. We tried a sequence length ranging between 7-28 days hoping that the LSTM would pick up on the concept of weeks, however the prediction accuracy did not seem to increase. In our grid search of parameters, we found that models containing a sequence length of 7, 14, and 21 were on the top ten

list. However, in hindsight, seeing as we do not label each weekday (added as a feature, for example), it is probably difficult for the model to learn the concept of weeks, especially since the datasets were limited [5].

The findings in our study, at least for the seven-days ahead prediction scenario, are also in agreement with previous research also comparing LSTM and ARIMA for time-series forecasting problems. In [8] they compared ARIMA and LSTM on four datasets and their RMSE values favored LSTM in three out of the four datasets.

5.3 Limitations and Relevance

Having a proper amount of data is essential to properly train and evaluate the network [5]. As the accessible data was limited to only one grocery chain during a period of four years, the results might have been affected. It would have been desirable to have access to sales data from multiple chains and stores during a longer period of time to enable a more varied training dataset for the network as well as a more diverse test dataset. Additionally, more data would have made it possible to conduct further experimenting, generating results that could increase the credibility of the study or that could highlight problems of the method.

Each unique problem in machine learning use different hyper parameters for optimal prediction accuracy, even if the underlying model is the same. Seeing as this is the case, it might prove difficult to find the optimal parameters. One possibility is to take inspiration from other models in similar settings, in our case time-series forecasting. However, seeing as we have created a relatively fundamental model, it has been hard to take inspiration from the more advanced models in the research papers we studied. Instead we manually implemented a grid search, which proved rather inefficient in terms of time efficiency. There exists libraries such as Hyperas or Sklearn which can do this much more efficiently, but unfortunately we discovered this late into our work. The grid search did yield results however, but it was rather coarse, and if time allowed could have been more fine-tuned.

As our knowledge of the implementation of LSTM was limited we opted for a higher abstraction library - Keras - to simplify the implementation of the LSTM network. However, as Keras has a higher level of abstraction, it becomes challenging to have full control of the network. The lack of control, even though it benefits us with a lack of necessary knowledge, can cause uncertainty in the results considering that it is unclear exactly what kind of network is built in the background. The models predicted better than the baseline model however, which at least is a little bit assuring.

The majority of the time have been spent on the LSTM-model, and not so

much the ARIMA-model. Both in terms of learning how it works, and the implementation. However, the LSTM-model provides considerably more options for fine-tuning compared to ARIMA. At least using the libraries we ended up with. This might discredit the ARIMA results. But yet again, both models are implemented using high-abstraction libraries and could probably be improved quite a bit. One thing that probably could have improved the ARIMA results might have been to preprocess the data differently. For example summing the sales for weeks instead of days.

It also seems necessary to discuss the actual usability of the model. Our ultimate goal was a model that could help a store order food more accurately to help reduce food waste/empty shelves. However, this model lack actual practical applications. Seeing as the model only predict sales one to seven days ahead, given a number of days worth of sales beforehand it could never be used in an actual store given delivery times. However, this could be expanded in future work. In addition to this, the models only covers one product, and is not generalized to all products which can affect the usability as it can be necessary to create more complex system for a more generalized prediction model [8]. Ultimately we would like to achieve a model where one could input a product ID to get the sales prediction of the desired product. Even though the model lacks practical usage, the conclusion that the LSTM-model predicts better than the ARIMA-model in the seven-day-ahead prediction scenario, motivates further research into the area.

5.4 Ethics and Sustainability

By studying the ecological lifecycle of food it's easy to see the huge impact that it has on the environment [27]. The food has to be grown/raised and after that be shipped and then potentially cooled/frozen in the stores. If the food then is thrown, the waste has to be taken care of, which also has negative impact on the environment. 100 000 tons worth of food is thrown each year by stores in Sweden [2]. Food which has been produced, sent, stored, cooled and thrown all in vain. Our aim for the study was to find efficient prediction models to reduce this waste and in turn benefit the sustainable development.

Chapter 6

Conclusion

The results of total difference in error show that the LSTM-model seemed to have higher prediction accuracy than the ARIMA-model in terms of both RMSE and MAE. However the t-test shows that the difference between the models in the one-day-ahead prediction scenario is not of statistical significance. Only in the seven-days-ahead prediction scenario does the t-test indicate that the LSTM model does have higher accuracy than the ARIMA model. Given that ARIMA is a state-of-the-art widely used model, the LSTM-network shows promising results for sales prediction in the scenario of seven-days-ahead prediction and is therefore concluded to be a model that can compete against the ARIMA-model

6.1 Further Research

Using the discussion above as a basis for this section we can draw some conclusions of things to be improved upon. First and foremost, a more practical application of the model seems reasonable to give the actual research an underlying motivation. Building a model which predicts sales for say two weeks/a month ahead could be a good start as we believe that grocery stores do not order food on a daily basis. This would also, as mentioned, provide a better discussion whether LSTM is superior to ARIMA for more complex models. We are not certain that the results obtained for our fundamental models will be consistent if expanded, even if that is our own prediction. Moreover, the fine tuning of hyperparameters can be improved. For more control an additional fine-tuning, building the LSTM within Tensorflow should be explored as well as implementing the ARIMA-model in a more detailed fashion.

Bibliography

[1] A. Galatsidas, *The guardian; sustainable development goals: Changing the world in 17 steps – interactive*, 2015. (visited on 01/01/2018).

[2] ICA.se, *Delat ansvar för matsvinnet*, 2017. (visited on 01/03/2018).

[3] P.-F. Pai and C.-S. Lin, "A hybrid arima and support vector machines model in stock price forecasting", *Omega*, vol. 33, no. 6, pp. 497–505, 2005, ISSN: 0305-0483.

[4] J. W. Taylor, P. E. McSharry, R. Buizza, *et al.*, "Wind power density forecasting using ensemble predictions and time series models", *IEEE Transactions on Energy Conversion*, vol. 24, no. 3, p. 775, 2009.

[5] Q. Yu, K. Wang, J. O. Strandhagen, and Y. Wang, "Application of long short-term memory neural network to sales forecasting in retail—a case study", in *Advanced Manufacturing and Automation VII*, K. Wang, Y. Wang, J. O. Strandhagen, and T. Yu, Eds., Singapore: Springer Singapore, 2018, pp. 11–17, ISBN: 978-981-10-5768-7.

[6] P. Doganis, A. Alexandridis, P. Patrinos, and H. Sarimveis, "Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing", *Journal of Food Engineering*, vol. 75, no. 2, pp. 196–204, 2006, ISSN: 0260-8774.

[7] I. Khandelwal, R. Adhikari, and G. Verma, "Time series forecasting using hybrid arima and ann models based on dwt decomposition", *Procedia Computer Science*, vol. 48, pp. 173–179, 2015, International Conference on Computer, Communication and Convergence (ICCC 2015), ISSN: 1877-0509.

[8] S. Krstanovic and H. Paulheim, "Ensembles of recurrent neural networks for robust time series forecasting", in *Artificial Intelligence XXXIV*, M. Bramer and M. Petridis, Eds., Cham: Springer International Publishing, 2017, pp. 34–46, ISBN: 978-3-319-71078-5.

[9] L. Wang, Y. Zeng, and T. Chen, "Back propagation neural network with adaptive differential evolution algorithm for time series forecasting", *Expert Systems with Applications*, vol. 42, no. 2, pp. 855–863, 2015, ISSN: 0957-4174.

[10] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Chi, "Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation", *Environmental Pollution*, vol. 231, pp. 997–1004, 2017, ISSN: 0269-7491.

[11] J. Kumar, R. Goomer, and A. K. Singh, "Long short term memory recurrent neural network (lstm-rnn) based workload forecasting model for cloud datacenters", *Procedia Computer Science*, vol. 125, pp. 676–682, 2018, The 6th International Conference on Smart Computing and Communications, ISSN: 1877-0509.

[12] kaggle.com, *Corporación favorita grocery sales forecasting*, 2017. (visited on 03/12/2018).

[13] C. Chatfield, *Time-series forecasting*. CRC Press, 2000.

[14] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, "Introduction", in *Time Series Analysis*. Wiley-Blackwell, 2013, ch. 1, pp. 7–18, ISBN: 9781118619193. DOI: 10.1002/9781118619193.ch1.

[15] C. Chatfield, *The analysis of time series: an introduction*. CRC press, 2016.

[16] G. Zhang, "Time series forecasting using a hybrid arima and neural network model", *Neurocomputing*, vol. 50, pp. 159–175, 2003, ISSN: 0925-2312.

[17] I. A. Gheyas and L. S. Smith, "A novel neural network ensemble architecture for time series forecasting", *Neurocomputing*, vol. 74, no. 18, pp. 3855–3864, 2011, ISSN: 0925-2312.

[18] Z.-L. Sun, T.-M. Choi, K.-F. Au, and Y. Yu, "Sales forecasting using extreme learning machine with applications in fashion retailing", *Decision Support Systems*, vol. 46, no. 1, pp. 411–419, 2008, ISSN: 0167-9236.

[19] R. Rojas, *Neural networks: a systematic introduction*. Springer Science & Business Media, 2013.

[20] B. Cortez, B. Carrera, Y.-J. Kim, and J.-Y. Jung, "An architecture for emergency event prediction using lstm recurrent neural networks", *Expert Systems with Applications*, vol. 97, pp. 315–324, 2018, ISSN: 0957-4174.

[21] S. Ji, H. Yu, Y. Guo, and Z. Zhang, "Research on sales forecasting based on arima and bp neural network combined model", in *Proceedings of the 2016 International Conference on Intelligent Information Processing*, ser. ICIIP '16, Wuhan, China: ACM, 2016, 41:1–41:6, ISBN: 978-1-4503-4799-0.

[22] C. I. Permatasari, W. Sutopo, and M. Hisjam, "Sales forecasting newspaper with arima: A case study", *AIP Conference Proceedings*, vol. 1931, no. 1, p. 030 017, 2018. DOI: 10.1063/1.5024076.

[23] V. Ş. Ediger and S. Akar, "Arima forecasting of primary energy demand by fuel in turkey", *Energy Policy*, vol. 35, no. 3, pp. 1701–1708, 2007, ISSN: 0301-4215.

[24] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization", *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

[25] M. Khashei and M. Bijari, "An artificial neural network (p,d,q) model for timeseries forecasting", *Expert Systems with Applications*, vol. 37, no. 1, pp. 479–489, 2010, ISSN: 0957-4174.

[26] N. S. Arunraj and D. Ahrens, "A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting", *International Journal of Production Economics*, vol. 170, pp. 321–335, 2015, ISSN: 0925-5273.

[27] J. Larsson, *Hållbara konsumtionsmönster: Analyser av maten, flyget och den totala konsumtionens klimatpåverkan idag och 2050*. Naturvårdsverket, 2015.