Sales Forecasting Using Different Forecasting Methods

Marta Rut Ólafsdóttir

Thesis of 30 ECTS credits
Master of Science (M.Sc.) in Engineering Management

June 2019

Sales Forecasting Using Different Forecasting Methods

by

Marta Rut Ólafsdóttir

Thesis of 30 ECTS credits submitted to the School of Science and Engineering
at Reykjavík University in partial fulfillment
of the requirements for the degree of
Master of Science (M.Sc.) in Engineering Management

June 2019

Supervisor:

Eyjólfur Ingi Ásgerisson, Supervisor
Associate Professor, Reykjavík University, Iceland

Examiner:

Michal Borsky, Examiner
Postdoctoral Researcher, Reykjavík University, Iceland

Copyright
Marta Rut Ólafsdóttir
June 2019

Sales Forecasting Using Different Forecasting Methods

Marta Rut Ólafsdóttir

June 2019

Abstract

The advantages to accurately forecasting sales are significant. For any company it is important to have foresight knowledge of financial outcomes and to have confidence in the forecasting process to be able to trust its results. This knowledge is the basis for all operations planning and makes the company better equipped to deal with situations that may arise. Recently machine learning has become the new buzzword in business and business leaders are keen to find out if these methods are applicable in today's business environment. One of the main challenges for businesses, when striving to adapt machine learning in their processes, is lack of appropriate data. Performance of machine learning models has been associated with access to large amounts of data that enable the models to learn. This thesis examines if published external data can be used to generate accurate sales forecasts for the medical devices company Össur. Two different approaches were used; traditional time series methods using only historical quarterly sales data from 2001 to 2018 and machine learning methods using the historical sales data as well as exogenous variables believed to influence sales. The traditional time series models that were applied were simple moving average, decomposition using least squares regression and Holt-Winters. The machine learning models applied were multiple linear regression, random forest and neural networks. The accuracy of the models was then compared using the RMSE value for the testing data set. The machine learning methods all yielded lower RMSE values then the traditional time series methods. The model the yielded the lowest RMSE value was the random forest model.

Notkun spálíkana við gerð söluspár

Marta Rut Ólafsdóttir

júní 2019

Útdráttur

Mikilvægi söluspáa fyrirtækja er óumdeilanlegt. Það er mikilvægt að stjórnendur fyrirtækja hafi yfirsýn yfir mögulega fjárhagslega afkomu og treysti sínum spálíkönum. Þessi yfirsýn gerir stjórnendum kleift að skipuleggja reksturinn á sem hagkvæmastann hátt og vera undirbúin fyrir þær aðstæður sem að mögulega gæti komið upp. Undanfarið hefur vitvélafræði verið að ryðja sér til rúms og eru stjórnendur fyrirtækja almennt áhugasamir um að komast að því hvort að þessum aðferðum megi beita innan fyrirtækja. Eitt meginvandamálið við þessar aðferðir hefur verið það að aðgangur að viðeigandi gögnum hefur verið ábótavant. Frammistaða líkanna af þessu tagi veltur á gæði og magni gagna sem til eru. Í þessu meistaraverkefni eru tvær mismunandi nálganir á söluspár skoðaðar. Annarsvegar hefðbundnar tölfræðilegar aðferðir og hinsvegar vitvéla aðferðir. Einnig er skoðað hvort að birt, opinber gögn séu nýtanleg til að þess að búa til söluspár með þessum aðferðum. Skoðaðar voru sölutölur frá árinu 2001-2018 frá stoðtækjafyritækinu Össur. ÖSsur er skráð fyrirtæki á markaði og eru þessar upplýsingar því opinberar í ársskýrslum. Þreumst hefðbundnum tölfræðilegum aðferðum var beitt til að búa til söluspárlíkön og þremur vitvélaaðferðum. Frammistaða líkananna var svo metin með því að bera saman RMSE gildi líkananna. Öll vitvéla líkönin skiluðu lægra RMSE gildi en hefðbundnu tölfræðilegu líkönin. Það líkan sem skilaði lægsta RMSE gildinu var slembiskógar líkan.

Sales Forecasting Using Different Forecasting Methods

Marta Rut Ólafsdóttir

Thesis of 30 ECTS credits submitted to the School of Science and Engineering
at Reykjavík University in partial fulfillment of
the requirements for the degree of
Master of Science (M.Sc.) in Engineering Management

June 2019

Student:

.....

Marta Rut Ólafsdóttir

Supervisor:

.....

Eyjólfur Ingi Ásgerisson

Examiner:

.....

Michal Borsky

The undersigned hereby grants permission to the Reykjavík University Library to reproduce single copies of this Thesis entitled **Sales Forecasting Using Different Forecasting Methods** and to lend or sell such copies for private, scholarly or scientific research purposes only.
The author reserves all other publication and other rights in association with the copyright in the Thesis, and except as herein before provided, neither the Thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

.....
date

.....
Marta Rut Ólafsdóttir
Master of Science

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Amputation is defined by the Oxford English Dictionary as the action of cutting of a limb [1]. There are nearly 2 million people living with limb loss in the United States alone and among those living with limb loss 54% is caused by vascular disease, such as diabetes, 45% is caused by trauma and cancer is the cause of less then 2% of amputations [2]. Since 1980 the number of adults with diabetes worldwide has quadrupled [3]. According to the World Health Organization there are currently 422 million people living with diabetes worldwide [4]. The prevalence of diabetes is expected to double in the United States by the year 2030 solely because of changes in the demographic composition of the population [2]. Similarly the number of people with diabetes who are living with limb loss is estimated to nearly triple by the year 2050 [2]. For individuals living with limb loss there are solutions other then being confined to a wheelchair. Most amputees should be able to maintain an active lifestyle with the help of prosthesis. There have been significant technological enhancements in recent years making it possible for amputees to walk, run and be active. Companies such as Össur ® produce prosthetic limbs making it possible for amputees to live a life without limitations.

1.1 Össur ®

Össur® is a multinational medical devices company producing non invasive orthopedics. Founded in Iceland by prosthesist Össur Kristinsson in the year 1971, Össur has grown immensely through a series of strategic acquisitions and their operations expanded from production of silicon sockets to producing prosthetic solutions, braces and supporting products with the aim to improve peoples mobility. Össur's headquarters are located in Iceland and a large part of the prosthetic limb production takes place in Iceland. There are over 3000 employees at Össur, located in 25 countries. Össur has extensive operations in America, Europe and Asia, with numerous distributors in other markets [5].

The company divides its products into bracing and support solutions and prosthetic solutions, each segment making up about 50% of sales. The prosthetic solutions include artificial limbs and related products for amputees, all designed to improve quality of life with an emphasis on clinical outcomes. Össur offers a full spectrum of lower-limb prosthetic products, including mechanical knees, feet and silicone liners. In addition to this Össur's bionic technology combines mechanics and electronics to effectively mimic amputee's natural sensory and motor control functions. Össur's prosthetic solutions product selection provides options for people of all ages and activity levels. They are divided in to 7 categories based on the need of the customer [6].

- **Balance solutions** are designed to support less active people who may struggle to maintain the ideal balance of safety, comfort and mobility.
- **Dynamic solutions** are designed to encourage increasingly active people who want to enhance their mobility and return to regular activities.
- **Impact solutions** are designed to enable especially active people to engage in high impact endeavors.
- **Junior solutions** are designed to champion children and their aspirations, from playing with friends, to competitive sports.
- **Sport solutions** are designed to empower athletes to fulfill their potential on sporting events and activities.
- **Post-op solutions** are designed to assist patients during post operative healing, with a view to regaining mobility as soon as possible.
- **Touch solutions** are designed to ensure the best possible outcomes for people with upper limb deficiencies.

Össur's braces and supporting solutions include product solutions designed to enhance the quality of life for people living with osteoarthritis by reducing pain and improving mobility. Össur also produces injury solutions designed to enhance the healing process of bone and soft tissue injuries. In addition to this Össur has a comprehensive compression therapy product portfolio. Compression therapy is a preferred treatment for venous ulcers and edema [7]. Products such as compression socks, tights and bandages are used to apply pressure to the vascular system in order to improve circulation and minimize swelling. The compression therapy products are sold under the french brand name Gibaud. Össur acquired Gibaud in 2006. Össur's growth has largely been through a series of strategic acquisitions, as previously stated. The first acquisitions were made in the year 2000 and since then companies have been acquired on a regular basis [6].

Even though Össur's customer segments are quite diverse, most amputees are people over the age of 65 since most amputations are caused by vascular problems, such as diabetes. Therefore these people make up a vast majority of customer segmentation. Össur contributes to changing the image of amputees and partners with elite disabled athletes that form **Team Össur**. In 2018 Team Össur athletes won 10 medals and set 2 world records in Berlin at the IPC Athletics European Championship and Markus Rehm, a long jumper with a prosthetic leg from Össur, set a new world record at the Japan Para Championships.

Össur is a market leader in non-invasive orthopedics and, according to the 2018 annual report, it has maintained its market position as the second largest player in both prosthetic and bracing and support solutions in the year 2018 [8]. The company experienced 5% organic growth, 38% net profit growth and had a 19% EBITA margin. 52% of sales came from prosthetics and 48% was from bracing and supports. Net sales grew form 569 million USD in 2017 to 613 million USD in 2018 [8].

1.2 Research Aim and Objective

This thesis aims to find an effective way to forecast net sales at a consolidated level at Össur ® using published historical data and quantitative methods. Another focus is to determine if

the data available is sufficient to provide reliable forecasts. Six different models were built and it was evaluated which one provides the most accurate results. Two different fields of data processing was examined; traditional time series analysis and machine learning. The first three models are based on traditional time series analysis and the last three are based on machine learning. The first model utilizes the method of simple moving average. The second model will utilize decomposition using least squares regression and the third is a variation of exponential smoothing called Holt-Winters. For the second part exogenous variables will be introduced. The machine learning techniques examined include multiple linear regression, random forest and artificial neural networks. The exogenous variables utilized in making the machine learning models will be covered in detail in Chapter 4.

The advantages to accurately forecasting net sales are significant. It is also of interest to examine if reliable forecasts can be developed using only published data. For any company it is important to have foresight knowledge of financial outcomes and to have confidence in the forecasting process to be able to trust its results. This knowledge is the basis for all operations planning and makes the company better equipped to deal with situations that may arise.

The following research questions were applied:

- Can published and external data be used to generate forecasting models for a publicly traded company like Össur?
- Is the data available sufficient to generate forecasts using machine learning methods?
- Do machine learning algorithms provide more accurate forecasts by returning smaller forecasting errors for Össur's net sales forecast compared to traditional time series analysis?

1.3 Thesis Outline

The literature review can be found in Chapter 3 and in Chapter 4 the data is introduced and analyzed. The methods examined are covered in Chapter 2 and in Chapter 5 results will be introduced. Chapter 6 contains discussions and thoughts on next steps.

Chapter 2

Methods

2.1 Description of Methodology

When forecasting using historical data there are many methods to choose from. In this thesis it is attempted to demonstrate which methodology is best suited for the data being used.

2.2 Performance criteria

In order to compare the methods it is important to define a performance criteria. These metrics are used to compare the difference between the actual data and the forecasted data. The following model performance metrics were used to evaluate the performance of the models: mean absolute deviation (MAD) and root mean square error (RMSE). The following equations demonstrate how these metrics are found:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i}^2)} \tag{2.1}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\mid y_i - \hat{y_1}^2 \mid}{y_i} * 100 \tag{2.2}$$

where y_i is the actual output while the $\hat{y}_i$ is the predicted output. The RMSE for the testing data is compared for all the models. The RMSE of the training data is only applied to the machine learning models and compared to the RMSE of the testing data to help identify overfitting.

2.3 Traditional Time Series Forecasting

A time series is a sequence of data points measured at successive points in time spaced at even time intervals. Time series forecasting produces forecasts solely based on historical data [9]. It is important that sufficient historical data is available and that the data is reliable [9]. In this thesis the traditional time series forecasting methods are used to create a baseline to compare the machine learning methods to. The time series methods used are simple moving average, linear least squares regression and Holt-Winters. The traditional time series methods were chosen based on what methods are frequently used within companies. When evaluating this the author relied on her personal experience, as well as Gartner, financial leadership council fro mid sized companies, according to surveys they have conducted, these methods are widely used within companies [10].

2.3.1 Simple moving average

Moving average forecasting is concerned with averaging past demand to project a forecast. This implies that the underlying pattern is constant with random fluctuations around the average [11]. Simple moving average is a widely used approach to forecasting. It is a simple mathematical way of converting past information into forecasts. Equation 2.3 demonstrates the mathematical expression of a moving average.

$$MAF_t = \sum_{i=t-n+1}^{t} Actual_i/n \tag{2.3}$$

where:

- MAFt = Moving average forecast at the end of period t
- i = period number
- t = current period (the period for which the most recent actual is known
- n = the number of periods in the moving average

As can be seen from Equation 2.3 the moving average is only examining the n most recent periods. Recent data may reveal current conditions better than older data. Once Equation 2.3 has been applied to all periods containing actual historical data, the most recent forecasted data point will be utilized to forecast the subsequent data point. So, for example, the the forecast for Q2 2018 is calculated using actual data from Q2, Q3 and Q4 of 2017 and the forcasted data point from Q1 2018.

2.3.2 Decomposition Using Least Squares Regression

Regression can be defined as a functional relationship between two or more correlated variables [11]. Linear regression is used for both time series forecasting and for causal relationship forecasting. The least squares method tries to fit a line to the data that minimizes the sum of squares of the vertical distance between each data point and its corresponding point on the line [11]. If a straight line is drawn through the general area of the points, the difference between the point and the line is y - Y [11]. The sum of squares of the difference between the plotted data points and the line points is as stated in Equation 2.4

$$Sum of squares = (y_1 - Y_1)^2 + (y_2 - Y_2)^2 + \ldots + (y_n - Y_n)^2 \tag{2.4}$$

The equation for a straight line is as demonstrated in Equation 2.5

$$Y = a + bx \tag{2.5}$$

where

- Y = dependent variable computed by the equation
- y = the actual dependant variable data point
- a = Y-intercept
- b = slope of the line

- x = time period

In the least squares method, the equations for a and b are the following:

$$a = \overline{y} - b\overline{x} \tag{2.6}$$

$$b = \frac{\sum xy - n\overline{x} * \overline{Y}}{\sum x^2 - n\overline{x}^2} \tag{2.7}$$

where

- $\overline{y}$ = average of all y
- $\overline{x}$ = average of all x
- x = x value at each data point
- y = y value at each data point
- n = number of data points

The line that optimizes the results is the one that minimizes the total represented in Equation 2.7. The method of decomposition using least squares regression was applied. Decomposition of a time series refers to identifying and separating the time series data in to the components of demand, those being: trend, seasonal, cyclical, autocorrelated and random [11]. When demand contains both seasonal and trend effects at the same time it is important to identify how they relate to each other. In this case the multiplicative seasonal variation was applied. In that approach the trend is multiplied by the seasonal factor. This method can be interpreted as a way of stating that the larger the basic amount projected, the larger the variation around this can be expected. The seasonal factor is the amount of correction needed in a time series to adjust for the season of the year. Indexes are calculated for seasons and cycles. The forecasting procedure then reverses the process by projecting the trend and adjusting it by the seasonal and cyclical indexes, which were determined in the decomposition process [11]. The process is as follows:

1. Decompose the time series into its components
    - Find seasonal component
    - Deseasonalize demand
    - Find trend component
2. Forecast future values of each component
    - Project trend component into the future
    - Multiply trend component by seasonal component

2.3.3 Holt-Winters

The Holt-Winter's method is a variation of exponential smoothing. The term "exponential smoothing" refers to the fact that the weights decrease exponentially as the observation gets older. The following equations make up the specific formula for Holt-Winters method, triple exponential smoothing:

$$L_t = \alpha(Y_t/S_{t-M}) + (1-\alpha)(L_{t-1} + T_{t-1}) \tag{2.8}$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \tag{2.9}$$

$$S_t = \gamma(Y_t/L_t) + (1-\gamma)S_{t-M} \tag{2.10}$$

$$F_(t+k) = (L_t + k * T_t) * S_{t-M+k} \tag{2.11}$$

Where

- L_t denotes the level at time t
- T_t denotes the trend at time t
- S_t denotes the seasonality at time t

The values of the α, β and γ weights are optimized for, by minimizing the RMSE error for the periods preceding the periods meant to be forecasted. This was done by using Excel Solver. The objective function is to minimize the RMSE subject to the following constraints:

$$0 \leq \alpha \leq 1 \tag{2.12}$$

$$0 \leq \beta \leq 1 \tag{2.13}$$

$$0 \leq \gamma \leq 1 \tag{2.14}$$

2.4 Machine Learning

In 1969 Donald Michie wrote an article about the usefulness of machines being able to learn from experience [12]. Michie wrote:

> It would be useful if computers could learn from experience and thus automatically improve the efficiency of their own programs during execution. A simple but effective route-learning facility can be provided within the framework of a suitable programming language [12].

Machine learning is a method of data analysis that is based on the notion that systems can identify patterns and learn from data without being explicitly told to do so. It is a branch of artificial intelligence that automates the process of analytic model building. There are several applications of machine learning that are being used prominently today, one of which is targeted ads from Google and Netflix.

2.4.1 Multiple linear regression

As stated in section 2.3.2, regression is the functional relationship between two or more correlated variable. The goal of multiple linear regression is to predict the value of one or more continuous target variables t given the value of a D-dimensional vector, x, of input variables [13]. The simplest linear regression model involves a linear combination of input variables

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + ... + w_D x_D \tag{2.15}$$

where

$$\mathbf{x} = (x_1, ..., x_D)^T \tag{2.16}$$

This is simply called linear regression. The key property of this model is the linear function of the parameters w_0,...,w_D. What imposes significant limitations to the model is the fact that it is also a linear function of the input variables x_i. The class of models can therefore be extended by considering linear combinations of fixed nonlinear functions of the input variables. Extending linear models with basis functions is known as polynomial regression. The form can be seen in the following equation:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) \tag{2.17}$$

Where $\varphi_j(\mathbf{x})$ is called the basis function and w_0 is called the bias parameter. The basis function represents some preprocessing or feature extraction. The basis function used can be a polynomial, represented by the following equation,

$$\phi_j(\mathbf{x}) = x^j \tag{2.18}$$

The total number of parameters in the model will be M by denoting the maximum value of the index j by M-1. The parameter w_0 allows for an offset in the data.

2.4.2 Random Forest

The feature space can be split into regions, or hypercubes, using decision trees. A visual representation of the feature space split into hypercubes can be seen in Figure 2.1. A tree is made where each node represents a split in the data according to a criteria, see Figure 2.2 for a visual representation of a decision tree. There are many types of trees, collectively called *Classification and Regression Trees (CART)*. Within each region there is a separate model to predict the target variable. To determine the structure and the parameter Θ_j it is common to use a greedy strategy. The greedy strategy starts with all the data and tries to find the split $x_i \leq \Theta_1$. This minimizes the error. It then iterates through the child nodes.

Random forest is a specific technique that applies bootstrap-and-aggregation, or *bagging* using *CART* as base classifiers. It builds a large collection of decorrelated trees and then aggregates them. Random forest is a very popular technique to use and is implemented in many packages. Its performance is on par with boosting techniques and they are very easy to tune.

For m=1 to M, representing the number of trees in the forest, a bootstrap sample of size N_m is drawn from the training data $\{\mathbf{x}_n, t_n\}_{n=1}^{N}$. Then a random forest tree T_m is grown to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree, until a minimum node size n_{min} is reached:

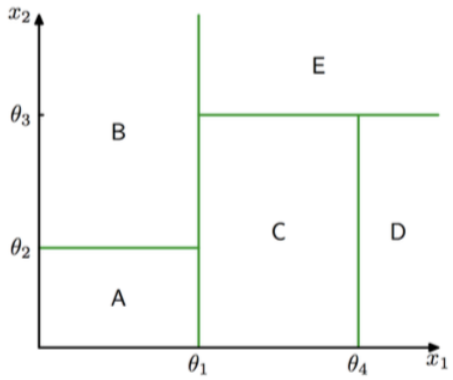

Figure 2.1: The feature space split into hypercubes [13]

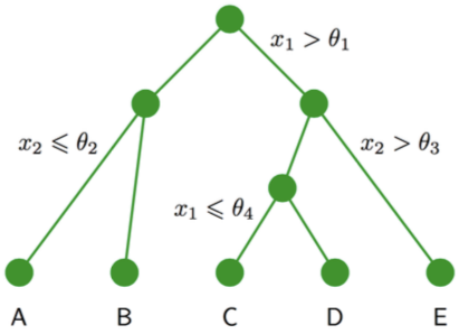

Figure 2.2: A decision tree [13]

1. Select d variables from the D dimensions
2. Pick the best variable, or split point, among the d
3. Split the node into two child nodes

The ensemble of trees is given $\{ \mathrm{T_m} \}_1^{\mathrm{B}}$. The prediction at the new point $\mathbf{x}$ is made using the following equation

$$y_B(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} T_m(\mathbf{x}) \tag{2.19}$$

2.4.3 Neural Networks

The term *neural network* stems from attempts to mathematically represent the information processing of biological systems [14]. The multilayer perceptron has proven to be of greatest practical value for statistical pattern recognition [13]. A visual representation can be seen in Figure 2.3. Each output is

$$y_j(\mathbf{x}, \mathbf{w}_j) = h(a_j) \tag{2.20}$$

and

$$a_j = \sum_{i=0}^{D} w_{ji} x_i = \mathbf{w}_j^T \mathbf{x} \tag{2.21}$$

A feed forward neural network is defined with L layers. The output of each layer forms the input to the next layer. The input of the overall network is still $x_\mathrm{i} = {z_i}^0$ while the output

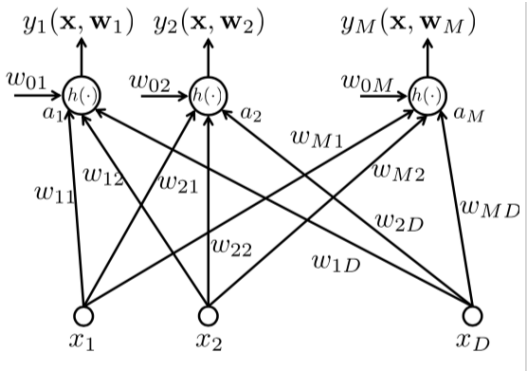

Figure 2.3: The perceptron extended to many outputs [13]

of the overall network is defined as

$$y_k(\mathbf{x}, \mathbf{w}) = z_k^{(L)} \tag{2.22}$$

the weights in the network should be considered as either a vector $\mathbf{w}$ or a sequence of matrices $\mathbf{W}^{(\mathrm{L})}$. To calculate the output from the input one has to successively apply

$$a_j^{(l)} = \sum_{i=0}^{M_{l-1}} w_{ji} z_i^{(l-1)} \tag{2.23}$$

and

$$z_j^{(l)} = h(a_j^{(l)}) \tag{2.24}$$

This thesis utilizes feed-forward neural networks. Therefore, the network is acyclical and information in later layers can not affect earlier layers. One property of the feed-forward neural network architecture are the weight space symmetries. If the signs of all weights leading into node k are flipped and compensated by flipping signs of all weights leading out of node k, that generates 2M versions of the same mapping. If all weight leading into and out of hidden unit k are interchanged with hidden unit j, that generates M! versions of the same mapping. Supervised learning methods, like neural networks, rely on training data to determine the value of the weights. The assumptions is made that the training data is of the form $\{ \mathbf{x}_\mathrm{n}, \mathrm{t}_\mathrm{n} \}_{\mathrm{n}=1}{}^{\mathrm{N}}$ where $\mathbf{x}_\mathrm{n}$ is an input vector and t_n is the desired output, or target value, for that input. The assumptions that the network is performing regression on a single value, so $t_\mathrm{n} \in R$ and that there is only a single neuron in the output layer and the activation function is linear

$$h_{M_L}^{(L)}(a_{M_L}^{(L)}) = h^{(L)}(a^{(L)}) = 1 \tag{2.25}$$

as $\mathrm{M_L}$=1. This is a natural choice for a single value regression. Multi-value regression and classification work similarly, but with different error functions and output activation functions. The input and target values are now constant so the output only depends on the weights. For a single value regression a natural choice of error function is the sum-of-squares

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (y(\mathbf{x}_n, \mathbf{w}) - t_n)^2 \tag{2.26}$$

the aim is to minimize the error by changing the weights. Because it is difficult to solve $\nabla E(\mathbf{w})$=0 directly, the numerical procedure called gradient decent is applied,

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E(\mathbf{w}^{(\tau)}) \tag{2.27}$$

where η is the learning rate and τ is the learning iteration. The error function is defined with respect to the entire training data, this is called batch learning. The error function can be decomposed into terms specific to individual data points $\mathbf{x}_\mathrm{n}$

$$E(\mathbf{w}) = \sum_{n=1}^{N} E_n(\mathbf{w}) \tag{2.28}$$

leading to online learning using either sequential or stochastic gradient decent.

Chapter 3

Literature Review

Forecasting is defined as the process of making predictions for the future based on historical data [9]. Business forecasting is an estimate of future development in business, such as sales. It is not easy to determine exactly how the future will unfold but various forecasting methods facilitate the making of forecasts and allow for a better prediction. For businesses it is extremely important to have confidence in the forecasting process and use the forecasts generated to react to upcoming economic fluctuations.

The development of business forecasting goes back to the seventeenth century and over the last three hundred years significant advances have been made in data-based forecasting [9]. Forecasting can be categorized in to qualitative and quantitative methods. The most common approach being a combination of the two. The role of judgment and judgmental approaches to forecasting, along with the development of data-based methods, has grown significantly in the past 25 years [9]. If no data is available human judgment is the only way of forecasting. In cases where data is available, human input should be used to review forecasts generated by data-based models and quantitative methods. When forecasting methods are chosen it is important to examine the different types of data patterns. There are normally four different types of patterns: horizontal, trend, seasonal and cyclical [9]. When data collected over time fluctuates around a constant level or mean, a horizontal pattern exists. This type of series is said to be stationary in its mean. The trend is the long-term component that represents the growth or decline in the time series over an extended period of time. The cyclical component is the wavelike fluctuation around the trend. The seasonal component is a pattern of change that repeats itself year after year [9].

A simple moving average (SMA) is used to generate forecasts based on an average of the most recent past observations. As each new observation becomes available, a new mean is computed by adding the newest value and dropping the oldest [9]. How quickly the model reacts to changes in the underlying pattern is determined by the number of periods included in the moving average. The smaller the number of periods used for the average the greater weight is given to recent periods. Even though it does not handle trend or seasonality well, a simple moving average approach is widely used for forecasting in practice. In a survey conducted by Gartner, financial leadership council, it found that out of 77 mid sized companies asked, 10% use a simple moving average method in their forecasting process [15].

Decomposition of time series means finding the series' basic components of trend, seasonal and cyclical. Indexes are calculated for seasons and cycles. The forecasting procedure then reverses the process by projecting the trend and adjusting it by the seasonal and cyclical

indexes, which were determined in the decomposition process [11]. The trend is the component that represents the underlying growth (or decline) in a time series. The trend may be produced, for example, by consistent population change, inflation, technological change, and productivity increases. The cyclical component is a series of wavelike fluctuations or cycles of more than one year's duration. Changing economic conditions generally produce cycles. Seasonal fluctuations are typically found in quarterly, monthly, or weekly data. Seasonal variation refers to a more or less stable pattern of change that appears annually and repeats itself year after year. Seasonal patterns occur because of the influence of the weather or because of calendar-related events such as school vacations and national holidays [9]. The least squares method is one of the oldest statistical methods and its' first modern precursor is likely to have been Galileo [16]. Least squares methods are very commonly used in modern day statistics.

While a moving average takes into account only the most recent data points, exponential smoothing provides an exponentially weighted moving average of all previously observed data. In 2004 Charles Holt developed a variation of exponential smoothing that allows for evolving local linear trends in a time series that can be used to generate forecasts. His original work was unpublished but widely cited. Holt's work on additive and multiplicative seasonal exponential smoothing became well known through a paper by one of his students, Peter Winters, who provided empirical tests for Holt's method. As a result the Holt-Winters method was developed [17]. The Holt-Winters method is widely used in practice and it is fundamentally triple exponential smoothing [18] [19]. It comprises forecast equation and three smoothing equations - one for level, one for the trend and one for the seasonal component [11]. In their forecasting method selection guide, CEB Financial Planning & Analysis Leadership Council recommends using the Holt-Winters method for forecasting if the underlying model is seasonal [15].

Traditional time series methods all have in common that they examine historical data in their attempt to foresee the future. As previously mentioned a combination of qualitative and quantitative methods is likely to yield the most accurate forecasting results, so human judgment is an important tool in forecasting. In recent years artificial intelligence has been gaining attention as a tool that mimics the human mind. Machine learning is a branch of artificial intelligence that bases its assumptions on the statement that systems can learn from data, identify patterns and make decisions without specifically being told to do so. When choosing machine learning methods to apply, the scale of intepretability versus accuracy was applied. Figure 3.1 shows a graph representing those metrics. The methods chosen were linear regression, random forests and neural networks. These methods seem to be representative of the spectrum of methods, shown in Figure 3.1. Interpretibility refers to how easily the methods can be explained and interpreted to individuals not extremely educated in the field. The accuracy index refers to how accurate the methods can be, if the applied to the appropriate data.

As previously stated, regression is a very commonly used tool in business forecasting [11] [9] [16]. Although linear models have significant limitations as practical techniques for pattern recognition, partially for problems involving input spaces of high dimensionality, they have nice analytical properties and form the foundation for more sophisticated models [13]. In a comparative study of the performance of four machine learning based algorithms in business forecasting, linear regression performed the best, in terms of mean absolute error and mean square error, compared to Gaussian process, multilayer perceptron and SMOreg

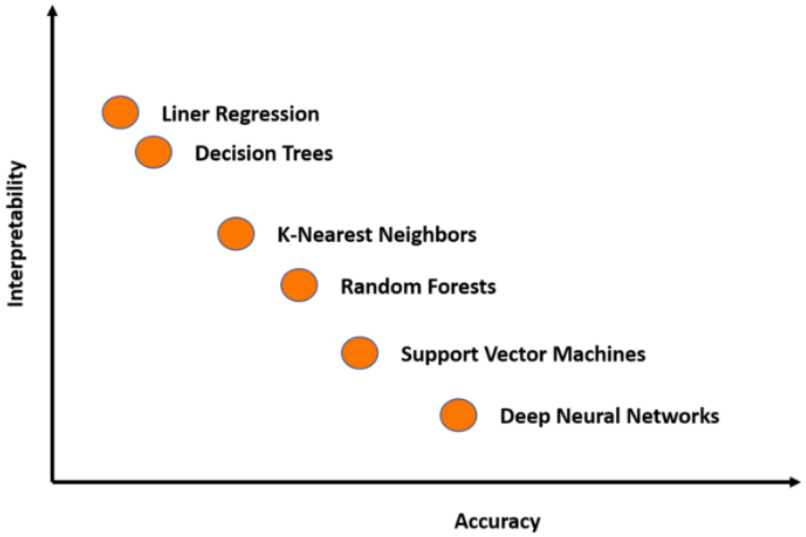

Figure 3.1: Interpretability versus accuracy, a criteria for model selection [20]

[21].

As described in Section 2.4.2, random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [22]. Random forest is known for its low tendency for overtraining and its high accuracy [22]. Because of these tendencies random forest is frequently used in practice and as a regression based method for forecasting [23].

In 1957 Frank Rosenblatt simulated something he called the *perceptron* on a 704 IBM computer at Cornell and by the early 1960s he had built a special-purpose hardware that provided a direct, parallel implementation of perceptron learning [13]. The objective of the perceptron was to be a model able of "...perceiving its environment, and learning to recognize those objects or events which it has perceived in the past" [14]. In 1962 Rosenblatt published "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms". The book was widely misinterpreted at the time as showing that neural networks were fatally flawed and could only learn solutions for linearly separable problems. However, the method only provided such limitations in the case of single-layer neural networks such as the perceptron and incorrectly conjectured that they applied to more general network models [13]. As a result the book contributed to the decline of research funding for neural computing. It was not until the mid 1980s that the situation was reversed. In present time there are probably thousands of applications of neural networks in widespread use, including in business forecasting [24]. A derivative analysis on how to best model and forecast aggregate retail sales with strong seasonal and trend patterns showed that a neural network model is able to capture the dynamic nonlinear trend and seasonal patterns, as well as the interaction between them [25].

Chapter 4

Data

4.1 The data set

In an effort to determine which of the previously stated methodologies is best suited for sales forecasting at Össur, first the tradition time series methods were applied to the historical sales data retrieved from Össur's published reports. Using quarterly data from the year 2001 to the year 2017, the models were used to generate a forecast for the year 2018. The results were then compared to the actual sales figures for the year 2018 and the method that generated the smallest deviation from the actual sales figures was chosen to act as a baseline. The machine learning models were then compared to the baseline to determine if they performed better. The data set used includes 72 data points on consolidated net sales retrieved from published quarterly and annual reports from the years 2001-2018. From the annual reports some exogenous variables used in the machine learning models were identified. The exogenous variables used are the following:

- Acquisitions as reported
- Number of employees
- Diabetes rates worldwide
- Percentage of the population over 65
- Working days per quarter
- Dow Jones industrial average
- Health expenditure as a percentage of GDP for major markets

Acquisitions as reported in the annual reports was identified as an exogenous variable. The presence of an acquisition in annual reports is identified using a binary variable, 1 if an acquisition occurred during the quarter, and 0 if it did not. The average number of employees at Össur per year is also identified as an exogenous variable. As the average number of employees for the full year is only reported once a year, it was decided to let that number be representative of the number of employees at the end of the second quarter of each year, at the middle of the year. It was estimated that the number of employees grows linearly over the year so to estimate the number of employees in other quarters the average growth in the number of employees per quarter was determined by calculating the difference in number of employees in year N and year N+1 and dividing it by 4. This resulted in an estimated average number of employees at Össur per quarter. Exogenous variables related to the demographic

are diabetes rate worldwide and the percentage of the population over 65 years old [26]. This dataset was processed in the same way as the number of employees, as it is only reported once a year. The number reported yearly as the percentage of the population over 65 is estimated to be representative of the middle of the year, end of the second quarter. So the same method was applied as for the number of employees and the quarterly value estimated. The data on diabetes rates was found by extracting numerical data from a graph representing the total number of adults with diabetes found in the IDF Diabetes Atlas [27], the graph is shown in Figure 4.1. The data was then preprocessed by extracting data points representative of each year from the line and averaging yearly data down to quarters. After interviewing employees of the financial planning and analysis department at Össur it became evident that the number of working days per quarter does, in some way, affect sales numbers. Therefore, the number of working days per quarter is identified as an exogenous variable.

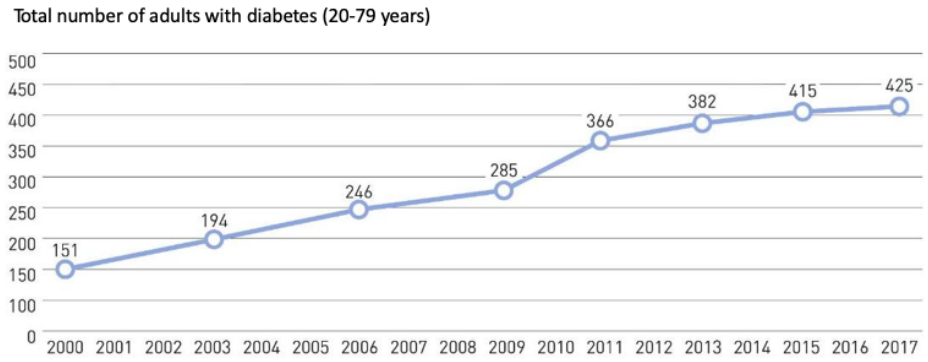

Figure 4.1: Total number of adults with diabetes, in millions, represented by a graph from the Diabetes Atlas by the International Diabetes Federation [27]

Economic exogenous variables chosen are the Dow Jones industrial average, the monthly average was downloaded and the quarterly average found [28]. Health expenditures as a percentage of gross domestic product for major markets was also identified as an economic exogenous variable. The countries identified as representative of major markets are the following:

- The United States
- Canada
- Australia
- Germany
- France
- Sweden
- Iceland

The health expenditure as a percentage of GDP is reported yearly. As it is likely to be predetermined on a yearly basis as a part of an annual governmental budget, the same number is applied to all four quarters in the year.

It was concluded that the number of exogenous variables should not exceed 1/5 of the total number of data points of the target variable. Given that the number of datapoints on historical sales data at Össur is 72, the number of exogenous variables should not exceed 14. The exogenous variables chosen are 7 in total, but since health expenditures as a percentage of GDP for major markets includes 7 individual data sets the total number of exogenous data

set variables totals to 13.

The data set was preprocessed to be usable to upload to Python and checked for muli-collinearity. A 15 dimensional correlation plot can be seen in Figure 4.2.

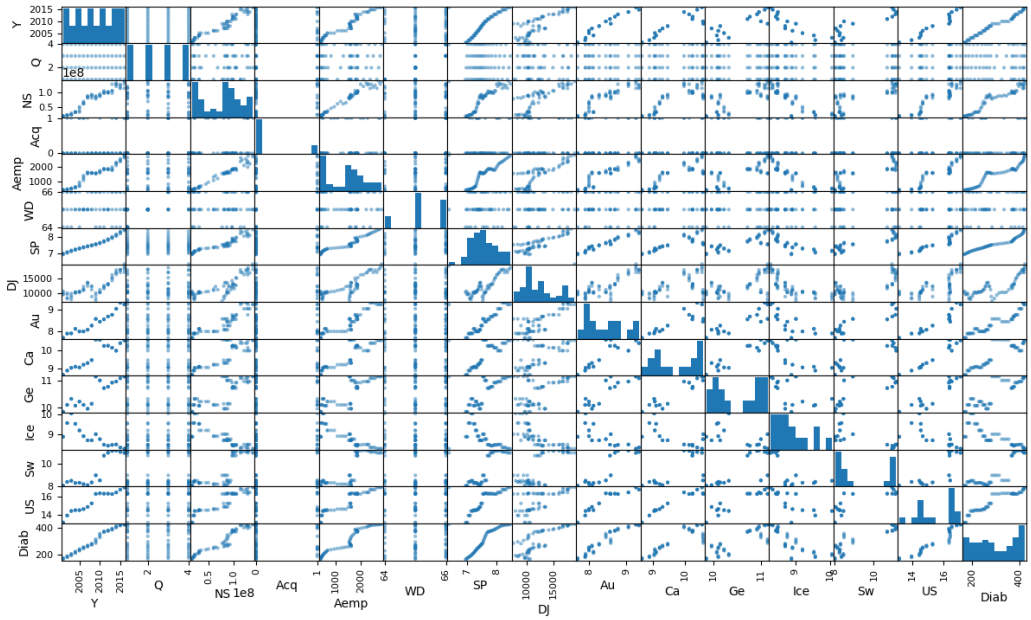

Figure 4.2: A correlation plot for all variables in the following order: year, quarter, Net sales, acquisitions, average employees, working days, senior population, Dow Jones, current health expenditures of the following markets in the following order; Australia, Canada, Germany, Iceland, Sweden, The United States of America. Diabetes is the variable to the far right.

The correlation plot visually portrays the correlation of each variable to another. If the plot shows a linear relationship between two variables there is correlation. For regression modesl, in particular, multicollinearity can be a problem. If present, multicollinearity can cause regression models to be very sensitive and decrease the precision of the model [29]. As can be seen in Figure 4.2 there is evidence of correlation between the net sales and average number of employees, sales increase as the company grows. There is also evidence of correlation e.g. between the percentage of the population over 65 and diabetes. Other variables seem to be have limited correlation. A case could be made to remove the variables that have high correlation but none will be removed to maximize the amount of data used for the prediction.

The data was compiled in an Excel file containing a column for each data set. That file was then converted to a CSV file to be suitable for upload to Python

4.2 Limitations and Assumptions

The biggest limiting factor of this thesis was access to data and data quality. Össur publishes reports both quarterly and annually and those reports were analyzed in detail to extract relevant data. The published reports are available going back to 2000. In 2001 Össur started reporting in USD so for consistency in currency calculations reports from 2001-2018 are used. The net sales data is extracted from the quarterly and annual reports as well as

other exogenous variables used for the machine learning models. Those variables include acquisitions as reported in the annual reports and average number of employees at Össur per year. The magnitude of data available on exogenous variables is also limiting when using machine learning methods, resulting in input data being limited by the variable that has the shortest available history [13]. To be able to predict sales going forward one would have to make a prediction of the variables that have not been reported since 2016 and use that prediction and apply it the machine learning models to generate the sales forecast going forward. Accuracy and reliability of the models is largely dependant on the reliability of the data set. If there are inconsistencies in the data the performance of the models will be unreliable. The size of the data set can also affect the reliability of the results.

Chapter 5

Results

After the quarterly sales data had been compiled from the quarterly reports a graph, plotting of the sales as a function of time, was created in Excel. The graph can be seen in Figure 5.1

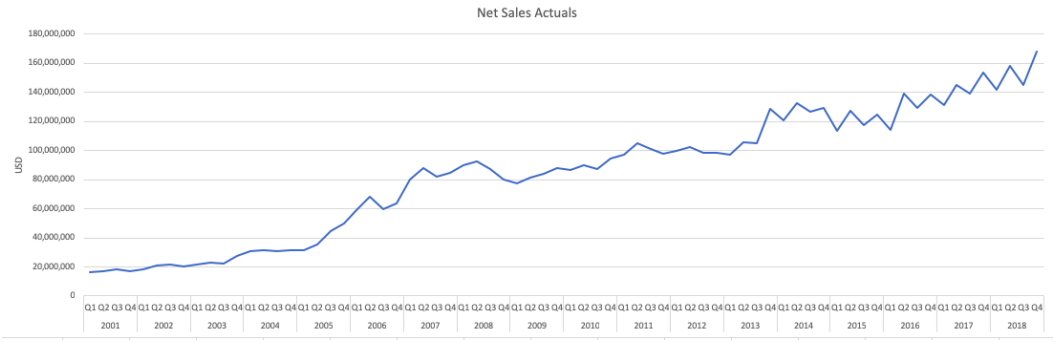

Figure 5.1: Sales as a function of time: actuals

As can be seen in Figure 5.1 the sales are growing with linear tendencies since 2001. There are fluctuations and seasonality is evident, especially in the latest quarters.

The performance criteria was applied by calculating an RMSE, MAD and MAPE indexes for the test data set, the test data set being the 2018 sales figures compared the generated forecast. The MAD and the MAPE were only calculated for the traditional time series methods, not the machine learning methods. The machine learning models are limited by the amount of available data. The sales data is available for the years 2001-2018 but other independent variables are only available through the year 2016. This limits the data available for application in the machine learning model to data from 2001-2016.

5.1 Traditional time series methods

5.1.1 Simple moving average

The results for the simple moving average method were obtained by applying Equation 2.3 to the historical sales data. As can be seen in Figure 5.2 the simple moving average approach smooths out the plot, averaging out spikes and valleys in the plot. Since this method is forecasting the next quarter by finding the average of the previous four periods it is constantly lagging in anticipating spikes and drops in sales, as could be expected. The RMSE and the MAD are both relatively high, as can be seen in Table 5.1. The RMSE is just over 11 million

USD and the MAD is just under 9 million USD, meaning that the average deviation per quarter of the forecasted period is 8.7 million USD. The MAPE, mean absolute percentage error, is 5,41%.

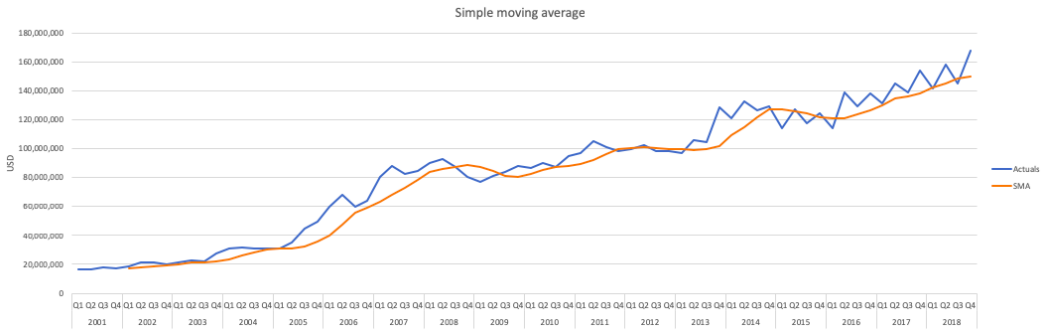

Figure 5.2: A plot of the results of the simple moving average compared to the actuals

RMSE	11.366.591 USD
MAD	8.744.500 USD

Table 5.1: Results of performance criteria tested on sales data from 2018: Simple moving average

5.1.2 Linear least squares regression

The results for the linear least squares regression were obtained by applying Equations 2.4, 2.5, 2.6 and 2.7 to the historical sales data. The graph shown in Figure 5.3 shows the evolution of the actual sales numbers against the forecast generated by the linear least squares regression method. Even though this method is supposed to minimize the sum of squares distance between the data points and their corresponding point the plot generated by the least squares method applied also used decomposition of the time series. Since the multiplicative seasonal variation was applied it is not surprising to find that the larger the basic amount projected, the larger the variation around this can be expected. This could explain the growing fluctuations in the predicted values. Even though the complexity of this method is substantially higher then the simple moving average method this one generates higher error values in the performance criteria. The RMSE is just over 12 million USD and the MAD 9,7 million USD, meaning that the deviation of the forecast from the actuals is an average of 9,7 million USD per forecasted quarter. The MAPE is 6,43 %, meaning that the forecast is off by 6.43% on average per quarter of the forecasted period.

RMSE	12.062.494 USD
MAD	9.680.584 USD

Table 5.2: Results of performance criteria tested on sales data from 2018: Linear least squares regression

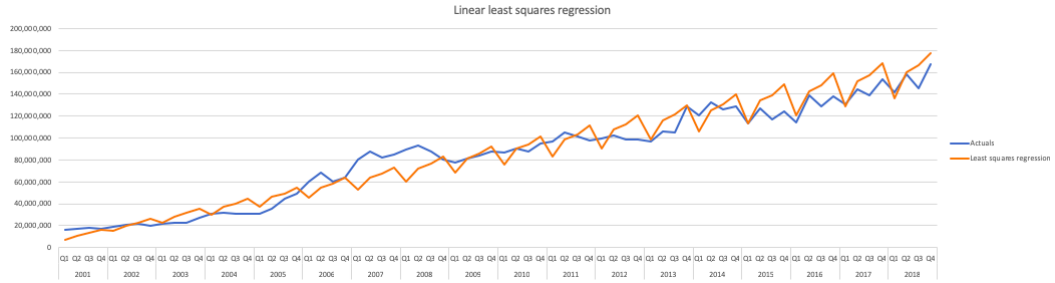

Figure 5.3: A plot of the results of the linear least squares regression compared to the actuals

5.1.3 Holt-Winters

The results for the Holt-Winters model were obtained by applying Equations 2.8, 2.9, 2.10 and 2.11 to the historical sales data. This method generated the best results out of the three traditional time series methods. The Holt-Winters method was optimized for the level, trend and seasonal indexes by minimizing the RMSE. This resulted in fairly accurate results. As can be seen in Figure 5.4, the forecasted points replicate the corresponding actual data point quite accurately and the forecasted graph seems to capture the dives and spikes of the sales data in most cases. The RMSE is significantly lower for the Holt-Winters method then the others, or 6,4 million USD and the MAD is 5 million USD. The MAPE for the test period is 3.19 %. The Holt-Winters method yields the best results of the traditional time series methods and is therefore used as a baseline to compare the machine learning models to.

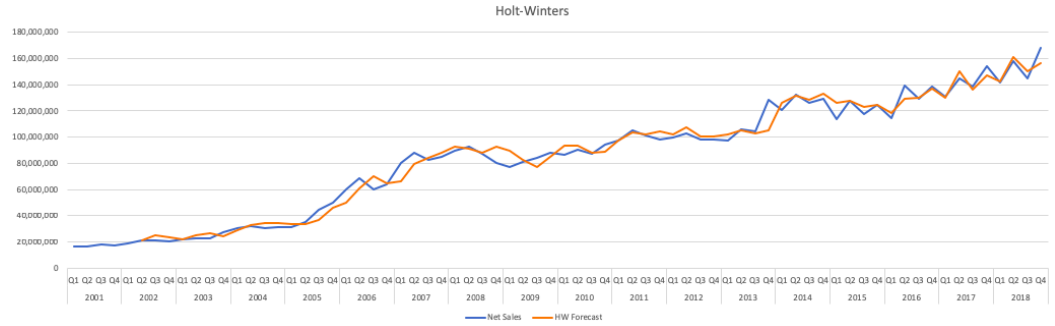

Figure 5.4: A plot of the results of the Holt-Winters method compared to the actuals

Table 5.3 shows the results of the performance criteria

RMSE	6.426.276 USD
MAD	5.071.807 USD

Table 5.3: Results of performance criteria tested on sales data from 2018: Holt-Winters method

5.2 Machine learning

The machine learning methods were applied to the data using the scikit-learn package in Python. As stated in Chapter 4 the amount of data usable for the machine learning methods is limited by the variable with the least amount of data. Even though data on net sales at Össur is available for the years 2001-2018 other variables, like the health expenditure as a percentage

of GDP, are only available through the year 2016. Because of this training and testing of the machine learning models can only be done on data from 2001-2016. The training was done on 90% of the data and the testing on the remaining 10%. The performance criteria applied to the traditional time series methods included the metrics MAD and MAPE. These metrics will no be applied to the machine learning methods as the do not apply. The performance criteria applied to the machine learning metrics will only include the RMSE metric for both the training and testing set. The train_test_split function from Python's Scikit-learn library was used to split the data randomly into training and testing sets to ensure that the data will not be bias. As previously stated the data was split into 90% training data and 10% testing data. The training prediction variable and the training independent variable therefore contains 57 rows and the testing prediction and independent variable contains 7 rows. This split was chosen to maximize the amount of training data so that it would be representative of the entire data set. The data set was standardized by using the StandardScaler function from the Scikit-learn library.

5.2.1 Multiple Linear Regression

The multiple linear regression method was applied to the data set by using Python's Scikit-learn library. Figure 5.5 shows a plot of all the actual values against the predicted values.

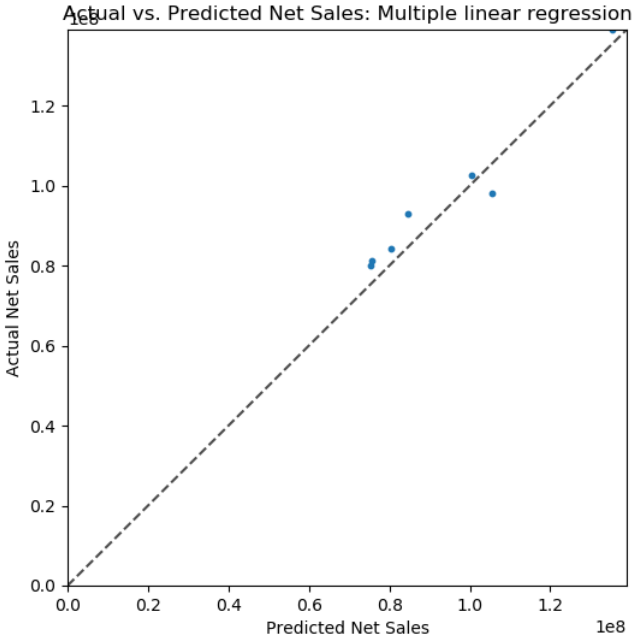

Figure 5.5: A plot of the actual net sales against the predicted net sales with the multiple linear regression method

In the graph in Figure 5.5, the closer the points are to the diagonal line the better the accuracy of the prediction. If the points are on the diagonal line that would mean the predicted values match the actual values. Points above the line indicate a prediction value lower than the actual value and vice versa. The points are quite close to the line and not all are placed on the same side. This would indicate an unbiased prediction, that is not consistently under- or overforecasting. The results of the performance criteria can be seen in Table 5.4.

RMSE training	5.342.077 USD
RMSE testing	5.535.019 USD
MAD	5.162.931 USD

Table 5.4: Multiple linear regression: Performance criteria for training and testing data

The RMSE is a good metric to evaluate regression models as it provides a clear value which represents the amount of the total error of the model [29]. As seen in 5.4, the RMSE for the training and testing set are very similar indicating limited overfitting on the training data. The RMSE of the multiple linear regression model is lower then that of all the traditional time series models.

5.2.2 Random Forest

The same steps were taken for the implementation of the random forest method on the data set as were taken for the multiple linear regression model. The same train-, test split and the same kind of standardization was applied. Various different parameters were tested before determining to use 20 trees in the random forest. The model yielded slightly better results for 20 trees than for 200 trees. Figure 5.6 shows the plot of all the actual values against the predicted values. The points are close to the diagonal line and are placed both above and below it. This also indicates an unbiased prediction.

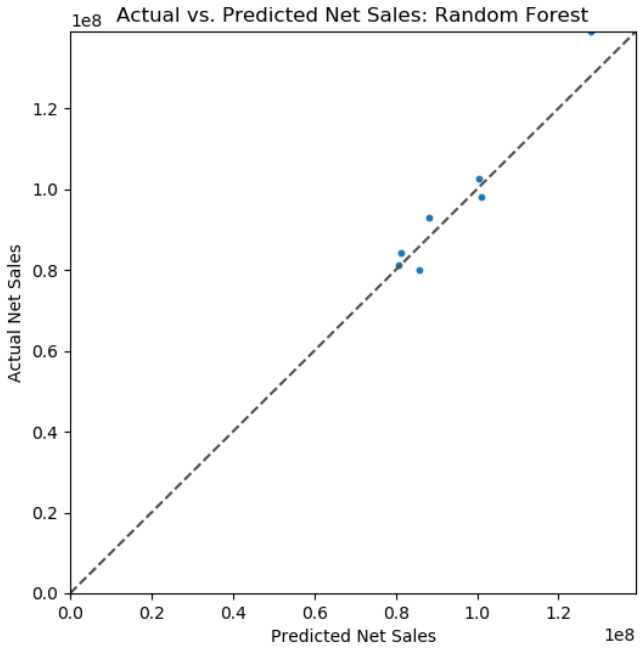

Figure 5.6: A plot of the actual net sales against the predicted net sales with the random forest method

The results of the performance criteria can be seen in Table 5.5.

As can be seen in Table 5.5, the RMSE of the training set is significantly lower then the RMSE of the testing set. This could indicate overfitting of the training data set. However, the RMSE of the testing set is still more favorable then that of the multiple linear regression.

RMSE of training set	2.706.965 USD
RMSE of testing set	5.318.036 USD
MAD	4.539.550 USD

Table 5.5: Random forest: Performance criteria for training and testing data

5.2.3 Neural networks

The neural networks method was implemented by using the MLPRegressor function from the Python Scikit-learn library. The MLPRegressor implements a multi-layer perceptron that trains using backpropagation with no activation function in the output layer, which can also be seen using the identity function as an activation function [30]. The parameters of the function include the hidden layer size, the activation function for the hidden layer and the solver for weigh optimization. The default for the activation function for the hidden layer is 'relu', the rectified linear unit function, returns:

$$f(x) = max(0, x) \tag{5.1}$$

The default function was therefore applied. When choosing the hidden layer size the following guidance was applied [31]:

- The size of the hidden layer should be 2/3 of the size of the input layer + the output layer
- Input layer size > Hidden layer size > Output layer size

when this rule of thumb was applied it resulted in the size of the hidden layer being 11 neurons. The default solver for weight optimization is 'adam', it refers to a stochastic gradient-based optimizer proposed by Kingma, Diederik and Jimmy Ba [32]. When these parameters were applied it yielded the results seen in Figure 5.7 and Table 5.6.

RMSE of training set	83.951.261
RMSE of testing set	98.719.141

Table 5.6: Neural networks model with hidden layer size of 11 neurons, default solver and default hidden layer activation function: Performance criteria for training and testing data

As can be seen by both Figure 5.7 and Table 5.6 this model does not capture the attributes of the data set. The points on the graph are very far off the diagonal line, on the far left side of the graph, indicating prediction values lower then the actual values and appear to be highly biased towards under prediction. As can be seen in Table 5.6 the RMSE of both the training and the testing sets are very high, a lot higher then those of all the other models. All of this is likely to be because of the size of the data set. Neural networks are known to work very well on large data sets. That gives them the opportunity to 'learn' the attributes of the training data set and apply them to the testing data. In this case the data set is very small so this approach does not yield good results.

It was then determined to change the parameters of the MLPRegressor function to determine if that would yield better results. The default solver for weight optimization is known to work well, in terms of both training time and validation score, with large data sets containing thousands or more training samples. As this is not the case for the data set being used in this case the solver called 'lbfgs' was applied. This one is known to converge faster

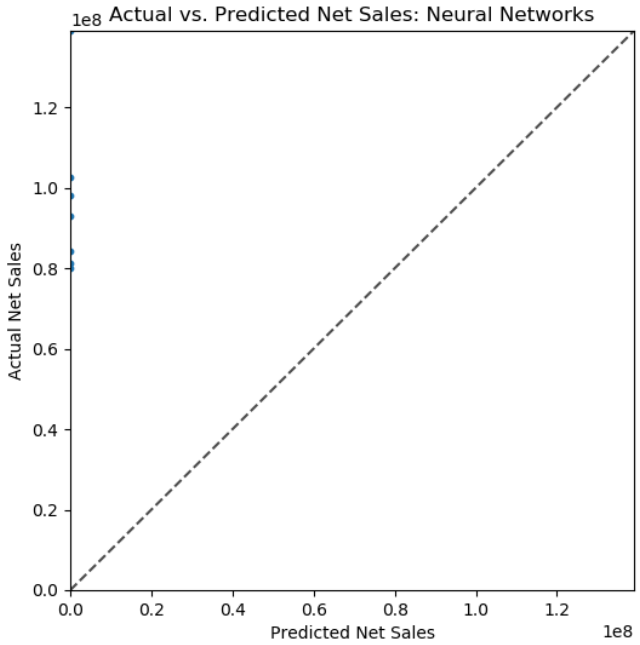

Figure 5.7: A plot of the actual net sales against the predicted net sales with the neural networks model with hidden layer size of 11 neurons and default solver and default hidden layer activation function

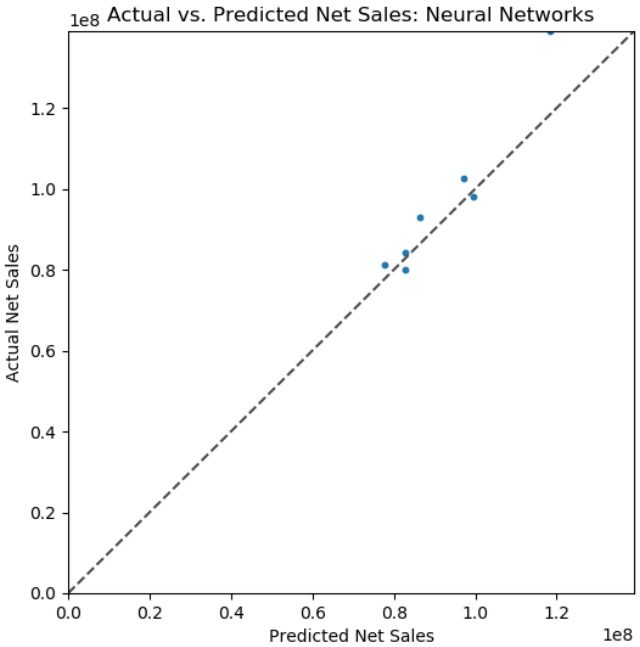

Figure 5.8: A plot of the actual net sales against the predicted net sales with the neural networks model with hidden layer size of 11 neurons and 'lbfgs' solver and default hidden layer activation function

and perform better for small data sets. The other parameters were kept the same. After this alteration the model yielded the results displayed in Figure 5.8 and Table 5.7.

RMSE of training set	2.302.711
RMSE of testing set	8.688.778

Table 5.7: Neural networks model with hidden layer size of 11 neurons, 'lbfgs' solver and default hidden layer activation function: Performance criteria for both the training and the testing data

As can be seen in Figure 5.8 and Table 5.7 this yields much better results. However, the RMSE of the training set is substantially lower then the RMSE of the testing set. This indicates overfitting. In an attempt to minimize the overfitting it was concluded to reduce the hidden layer size and test various numbers of neurons. The following results were obtained

Metric	3 neurons	2 neurons	1 neuron
RMSE of training set	2.457.760	2.865.261	5.404.047
RMSE of testing set	7.077.483	5.899.563	5.703.826

Table 5.8: Neural networks model with hidden layer size of 3, 2 and 1 neurons, 'lbfgs' solver and default hidden layer activation function: Performance criteria for both the training and the testing data

As can be seen from Table 5.8 a hidden layer size of 1 neuron does yield the best results. The RMSE of the testing set is minimized and the RMSE of the training set is in line with the RMSE of the testing set suggesting limited overfitting. A neural networks model with 1 neuron is essentially multiple linear regression. As can be seen from the results, the RMSE of the training and testing set for the neural networks model with 1 neuron and the multiple linear regression model are very similar. What could explain the slight difference is that they use different activation functions. Figure 5.9 shows the plot of the actual sales against the predicted sales with updated parameters of the neural networks model. Figure 5.9 shows that the points are much closer to the diagonal line than in Figure 5.7 and the predicted values have less bias as they are spread on both sides of the diagonal line.

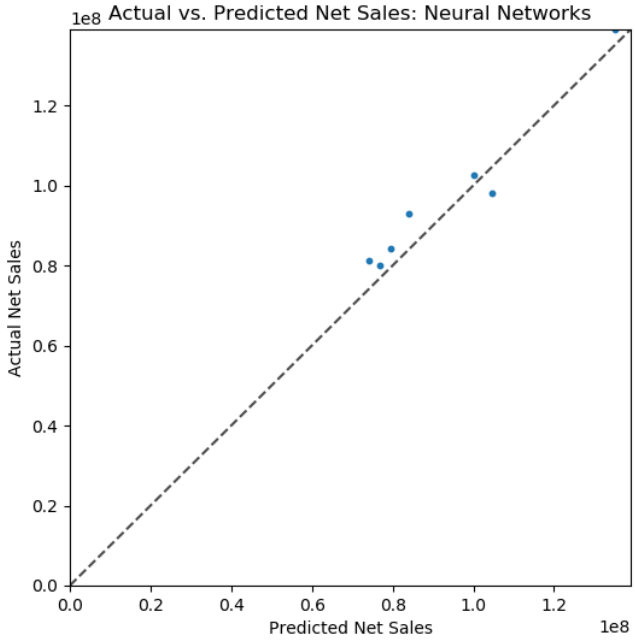

Figure 5.9: A plot of the actual net sales against the predicted net sales with the neural networks model with hidden layer size of 1 neurons and 'lbfgs' solver and default hidden layer activation function

5.2.4 Result summary

The results of the performance criteria for all the methods can be seen in Table 5.9. Table 5.9 shows that the machine learning methods all perform better than the traditional time series methods when comparing the RMSE of the testing set.

Metric	SMA	LSR	HW	MLR	RF	NN
MAD [USD]	8.744.500	9.680.584	5.071.807	5.162.931	4.539.550	5.291.446
RMSE of training set	-	-	-	5.342.077	2.288.168	5.404.047
RMSE of testing set	11.366.591	12.052.494	6.426.276	5.535.019	5.360.680	5.703.826

Table 5.9: A summary of the results of the performance criteria

Chapter 6

Conclusion

The objective of this thesis was to determine what methodology, out of the six methodologies examined, could most accurately forecast net sales for the medical devices company Össur. In addition, a part of the objective was to determine if the data available could be used to generate sales forecasts using machine learning methods. Six different methods were applied to the data set and then they were compared to one another. Table 5.9 shows a summary of the results of the performance criteria used to compare the methods. It is evident from looking at Table 5.9 that the machine learning methods perform better by yielding more favorable RMSE results for the testing data set. Out of the traditional time series methods the Holt-Winters yielded the best results with an RMSE of the testing set of 6.426.276. That was therefore used as a benchmark for the machine learning methods. The following answers to the research questions became evident in the process of writing this thesis:

- The machine learning methods generated better sales forecasts then the traditional time series methods.
- The random forest method generated the best sales forecasts when comparing the RMSE value of the testing set.
- The data, though limited, did provide adequate results when applied to the methods examined.
- The machine learning methods are likely to benefit from more data.

All the machine learning methods performed better than the benchmark, with the random forest method performing the best. The random forest method generated an RMSE of the testing set of 5.318.036. It is fairly surprising how well the machine learning methods perform given the limitations of the data set. The data set is very small and methods, such as neural networks normally perform best on very big data sets [13]. The first implementation of the neural networks model did in fact perform very poorly for this reason. However, after a new parameter was implemented in the MLPRegressor function it performed much better. The new parameter was a solver for weight optimization that is an optimizer in the family of quasi-Newton methods but quasi-Newton methods are methods used to either find zeroes or local maxima and minima of functions [33]. Once this paramater was applied the model performed much better and the RMSE metric reduced. By reducing the size of the hidden layer the performance of the neural networks model became more favorable until the size of the hidden layer was down to one neuron. Even though this is unusual, it did in fact yield the best results for the neural networks model. This might be caused by the fact that the data set is so small, it does not benefit from more neurons in the hidden layer. The trial and error approach

to finding the appropriate number of neurons in the hidden layer is well known [34]. The backward approach to trial and error was applied, starting with finding the number of neurons in the hidden layer using the rule-of-thumb of the neurons being 2/3 of the number of neurons in the input layer plus the output layer. Then the backward approach was taken, reducing the number of neurons until the best results were obtained. The changes made to the neural networks model did improve its performance but the random forest method still performed the best, generating the lowest RMSE value. There is a substantial difference of the RMSE of the training and the testing set for the random forest method. This does suggest over-fitting of the model on the training data. However, it still performs the best on the testing data.

Once results from all six models have been obtained it is evident that the machine learning methods provide a substantially better sales forecast then the traditional time series methods. Even though the performance of the Holt-Winters model is much better then that of the other traditional time series models, it is still not as good as the machine learning models. The random forest method provided the best sales forecast of the machine learning models. It would therefore be recommended that Össur seriously consider applying random forest to their sales forecasting process. If the appropriate data is available random forest, as well as the other machine learning methods, could provide more accuracy in their forecasting process then the traditional time series methods.

Bibliography

[1] *Amputation | Definition of amputation in English by Oxford Dictionaries*. [Online]. Available: https://en.oxforddictionaries.com/definition/amputation (visited on 05/06/2019).

[2] K. Ziegler-Graham, E. J. MacKenzie, P. L. Ephraim, T. G. Travison, and R. Brookmeyer, "Estimating the Prevalence of Limb Loss in the United States: 2005 to 2050", *Archives of Physical Medicine and Rehabilitation*, vol. 89, no. 3, pp. 422–429, Mar. 2008.

[3] NDC Risk Factor Collaboration, "Worldwide trends in diabetes since 1980: A pooled analysis of 751 population-based studies with 4·4 million participants", *The Lancet*, vol. 387, no. 10027, pp. 1513–1530, Apr. 2016. DOI: 10.1016/S0140-6736(16)00618-8. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140673616006188 (visited on 02/21/2019).

[4] "Global report on diabetes", en, World Health Organization, Geneva, Switzerland, Tech. Rep., 2016.

[5] *Business Forecasting | Encyclopedia.com*. [Online]. Available: https://www.encyclopedia.com/social-sciences-and-law/economics-business-and-labor/economics-terms-and-concepts/business (visited on 03/31/2019).

[6] *Össur Americas*. [Online]. Available: https://www.ossur.com/americas (visited on 02/22/2019).

[7] R. Kolluri, "Compression Therapy for Treatment of Venous Disease and Limb Swelling", *Current Treatment Options in Cardiovascular Medicine*, vol. 13, no. 2, pp. 169–178, Apr. 2011.

[8] Össur Annual Report 2018, "ÖssurAnnualReport2018.pdf", Tech. Rep. [Online]. Available: http://hugin.info/133773/R/2233435/878594.pdf (visited on 02/22/2019).

[9] J. E. Hanke and D. W. Wichern, *Business forecasting*, 9th ed. Upper Saddle River, N.J: Pearson/Prentice Hall, 2009.

[10] CEB Financial Planning & Analysis Council, *Forecasting Method Selection Guide*. [Online]. Available: https://www.cebglobal.com/member/finance-midsized/tools/17/forecasting-method-selection.html?.

[11] Jackobs, F. Robert, Berry, William L., and Whybark, D. Clay, *Manufacturing planning & control for supply chain management*, 6th ed. New York: McGraw-Hill, 2005.

[12] D. Michie, ""Memo"functions and Machine Learning", en, p. 4,

[13] C. M. Bishop, *Pattern recognition and machine learning*, en, ser. Information science and statistics. New York: Springer, 2006.

[14] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, English. Spartan Books, 1962.

[15] CEB Finance Practice, *Forecast Credibility Benchmark*. [Online]. Available: https://www.cebglobal.com/member/finance-midsized/benchmarks/17/forecast-credibility-benchmark.html?.

[16] Abdi, H., "Least squares", *Encyclopedia of Research Design*, vol. 2010, [Online]. Available: http://www.utd.edu/~herve/abdi-LeastSquares2010-pretty.pdf (visited on 04/04/2019).

[17] R. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Science & Business Media, Jun. 2008.

[18] P. S. Kalekar, "Time series Forecasting using Holt-Winters Exponential Smoothing", en, p. 13,

[19] *Robust forecasting with exponential and Holt–Winters smoothing - Gelper - 2010 - Journal of Forecasting - Wiley Online Library*. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/for.1125 (visited on 04/11/2019).

[20] J. Rodriguez, *Interpretability vs. Accuracy: The Friction that Defines Deep Learning*, Jun. 2018. [Online]. Available: https://towardsdatascience.com/interpretability-vs-accuracy-the-friction-that-defines-deep-learning-dae16c84db5c (visited on 04/22/2019).

[21] Department of Computer Science and Engineering, Anna University, BIT-Campus, Tiruchirappalli, India., D. A. A. Gnana Singh, E. J. Leavline, S. Muthukrishnan, and R. Yuvaraj, "Machine Learning based Business Forecasting", *International Journal of Information Engineering and Electronic Business*, vol. 10, no. 6, pp. 40–51, Nov. 2018. [Online]. Available: http://www.mecs-press.org/ijieeb/ijieeb-v10-n6/v10n6-5.html (visited on 03/29/2019).

[22] Breiman, L., "Random forest", *Kluwer Academic Publishers*, vol. 45, pp. 3–32, [Online]. Available: https://link.springer.com/content/pdf/10.1023%2FA%3A1010933404324.pdf (visited on 04/22/2019).

[23] F. Jiménez, G. Sánchez, J. M. García, G. Sciavicco, and L. Miralles, "Multi-objective evolutionary feature selection for online sales forecasting", *Neurocomputing*, vol. 234, pp. 75–92, Apr. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231216315612 (visited on 04/22/2019).

[24] G. P. Zhang, *Neural Networks in Business Forecasting*. Idea Group Inc (IGI), 2004.

[25] I. Alon, M. Qi, and R. J. Sadowski, "Forecasting aggregate retail sales:: A comparison of artificial neural networks and traditional methods", *Journal of Retailing and Consumer Services*, vol. 8, no. 3, pp. 147–156, May 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0969698900000114 (visited on 04/22/2019).

[26] *Population ages 65 and above (% of total) | Data*. [Online]. Available: https://data.worldbank.org/indicator/sp.pop.65up.to.zs?end=2017&start=2001 (visited on 04/13/2019).

[27] International Diabetes Federation, "IDF Diabetes Atlas", International Diabetes Federation, Tech. Rep. 8th Edition. [Online]. Available: https://diabetesatlas.org/resources/2017-atlas.html.

[28] *Dow Jones Industrial Average Historical Rates*. [Online]. Available: https://www.investing.com/indices/us-30-historical-data (visited on 04/26/2019).

[29] P. Teh, *An Introduction to Applied Machine Learning with Multiple Linear Regression and Python*, Jul. 2018. [Online]. Available: https://medium.com/@powersteh/an-introduction-to-applied-machine-learning-with-multiple-linear-regression-and-python-925c1d97a02b (visited on 05/01/2019).

[30] *1.17. Neural network models (supervised) — scikit-learn 0.20.3 documentation*. [Online]. Available: https://scikit-learn.org/stable/modules/neural_networks_supervised.html (visited on 05/02/2019).

[31] J. Heaton, *Introduction to Neural Networks with Java*. Heaton Research, Inc., 2008, Google-Books-ID: Swlcw7M4uD8C.

[32] *Sklearn.neural_network.MLPRegressor — scikit-learn 0.20.3 documentation*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html (visited on 05/02/2019).

[33] J. Nocedal and S. Wright, *Numerical Optimization*. Springer Science & Business Media, Jun. 2006.

[34] F. S. Panchal and M. Panchal, "Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network", p. 10, 2014.

School of Science and Engineering
Reykjavík University
Menntavegur 1
101 Reykjavík, Iceland
Tel. +354 599 6200
Fax +354 599 6201
www.ru.is