university of groningen

faculty of science and engineering

Mining sales data to identify customer profiles, and predict sales

Bachelor's Project Computing Science

June 2024

Author: Lukáš Lovás
Student Number: S4715179
First supervisor: Dr. Dilek Düştegör
Second supervisor: Huy C. Truong, MSc

Abstract

In the current world of marketing, the importance of using the data that companies collect to their advantage is rising. Data mining makes the process of analysing the data and providing insights more accessible and much more efficient.

We will adapt existing models, a sailfish optimization algorithm with random disturbance strategy - extreme learning machine (SFOR-ELM) and Long Short-Term Mem- ory network (LSTM), for data prediction to our cause and data, and we will try to predict the sales per month of products the company offers, as well as the total amount of sales a week. These predictions are going to help the company with planning and management processes, and thus save them money as well as time. Later we will compare these models for the prediction based on the accuracy and analyze which of them is more suitable for the data we were given.

Secondly, we will use statistical tools, to provide customer segmentation of the company customers, providing an overview of the customers, helping the company to understand them better, and selecting groups of customers that the marketing should be focused on.

The paper demonstrates how to perform customer profiling using the RFM model and predict customers' churn with almost 96% accuracy. Furthermore, the predictions of the total amount of orders in a week with mean squared error of 0,156 and mean average error of 0,307 and the predictions of parts sold in a month with mean squared error of 0,148 and mean average error of 0,108.

Contents

List of Figures

List of Tables

1 Introduction

In the current world, companies are trying to collect big amounts of data about their customers and sales. Even though, this data holds enormous potential, complex analytical techniques are required to gain useful information. Data mining and customer profiling are tools that can ease this process and make it more efficient and precise at the same time [10]. Results bring valuable insights that help businesses optimize sales performance, improve customer relationships, or decrease storage costs. This will lead to improved and more precisely tailored marketing strategies, or more efficient ways of acquiring raw materials needed for production, as the company will have approximate information about coming sales. [21].

The process of data mining can be divided into three separate steps that are performed, this division makes the whole process simpler and easier to understand. We can say that data mining consists of preprocessing the data, mining the information, and interpreting the result for the end user, this can be seen in the Figure 1 [8].

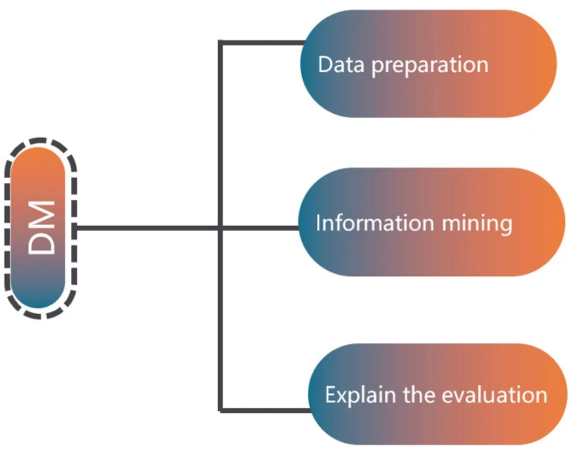

Figure 1: Simplified data mining scheme, adapted from [8]

In recent research different techniques were used for future sales prediction and customer profiling, mostly techniques from the fields of statistics, machine learning, and neural networks. However, the application of them to real-world problems or more complex datasets was done very rarely. We plan to contribute to the academic field by employing various tools, adapted to work with our data, to the actual business scenario we were given, and producing sales prediction, prediction of the sales that the company will make in next time frame, and customer segmentation, dividing customers into groups based on specified metrics, for an existing company. If the data and tools we plan to use are compatible, the results are expected to have a direct impact on the sales performance of the company that provided the data set. The impact will be achieved by using our results to optimize marketing strategies, improve new customer acquisition, and help to retain current customers using customer segmentation. Also, storage efficiency is expected to be increased and costs reduced, as the amounts of raw material that will be acquired and will be needed to be stored can be enhanced with the use of predictive data. Secondly, the company will have the predicted approximate numbers of sales of their products so they can produce these products in advance, and so improve their efficiency as when some unexpected order arrives, the machines will be free for production. The staff that works with the production can also be reduced in times when the lines are not that busy, as only minimal staff will be needed to handle unexpected orders. Predicting total amounts sold in a week will help the company plan its investments and potential expansion.

We chose to use one algorithm to divide customers into clusters and one algorithm to divide customers based on their churn(whether they make another purchase in the future or not). For the sales prediction, we will use a Recurrent Neural Network, specifically LSTM and a Feed-Forward Neural Network, specifically SFOR-ELM[16]. The research questions that emerge are :
“Is it possible to predict sales data accurately using LSTM and SFOR-ELM?”
“How can the insights gained from sales predictions and customer profiling optimize the company’s operations?”

Related Work summatizes related work in the fields of data preprocessing, customer segmentation and data prediction. Dataset and preprocessing introduces the dataset we are working with and describes preprocessing methods we use for the experiments. Models and Algorithms explains why we chose given algorithms and how they work. Results of Profiling models and Results of Predicting models present the findings and compare the results. Discussion Discuses the implications of the results for the company and Conclusion and Future Works summarizes the whole research and provides suggestions for future work. In Appendix results of further experiments are presented.

2 Related Work

We will analyze the current state of the art in first two steps of data mining, as seen in Figure 1, data preprocessing and information mining, choose algorithms to use and provide reasoning for the selection.

2.1 Preprocessing

The techniques in preprocessing of the data consist mainly of data cleaning, integration, and reduction. Data cleaning, a technique that is normally performed at the start of the data mining, is used to solve the problem of so-called "dirty data", which would negatively affect the process of data mining later on. Data integration is used when the data is stored across different data sets, a common problem is when the representation of the same data features has different names inside of datasets. Correlation analysis helps to transform these data into a suitable form for data mining. Mining a large dataset might take an unrealistic amount of time, we can use data reduction tools to reduce this time, by compressing the size of the input but keeping the integrity of data, using models like the principal component analysis [8].
What to do with missing values is another issue that preprocessing is trying to solve. Missing data can harm the integrity of data but also can deviate from the mining process and lead to different results. A missing value imputation algorithm based on the evidence chain (MIAEC) was built to solve this issue by building an evidence chain to estimate the missing values.[22]

2.2 Customer profiling

One way to use the preprocessed sales data is to segment the customers to provide a better overview of this sector for the company. Not all customers are going to stay with the company for a long time, some will be more satisfied while others less, same with amount of money they bring to the company by purchasing their products. The main reason to do customer profiling for the company is to target customers that bring the most value to the company rather than those that bring little value.[18] When deciding which tool to use for this process we need to take into consideration what data or data features we want to extract, what type of data we have in the data set we work with, and also how we want to interpret the results and what is our goal. One option for customer profiling is to use the Incremental recency, frequency, monetary value(RFM) Model to profile customers into different groups [4][18]. This model builds on the general RFM model, which evaluates customer value, loyalty, and tendency based on three indexes, these three indexes are real-time values so the results need to be recalculated with the income of new data, incremental RFM model makes this process more efficient by using historical/older data. Then this can be used in this segmentation of customers to compute the Target group index(TGI) of most sold products, this index indicates preferences for different products of various customer groups.
Another approach to customer segmentation is to use supervised machine learning techniques

and their combination on the data set to try to predict the churn of the customers. A combination of Logistic regression, eXtreme gradient boosting, and Random Forest algorithms as an ensemble approach had the best results in experiments done on multiple datasets from different business sectors[9]. Random Forest algorithm constructs deep decision trees and provides prediction by majority vote. eXtreme gradient boosting algorithm combines the results of several classifiers while trying to minimize the error of the previous model.
More ways to do customer segmentation include K-mean clustering or BIRCH (Balanced iterative reducing and clustering using hierarchies) clustering method. To improve the process of K-means clustering, hybrid models can be used to process unstructured data, as with typical clustering methods this process can be unrealistic for some datasets [14]. BIRCH is an unsupervised data mining scalable algorithm that can profile customers quickly and is used for large datasets.[12]

2.3 Sales prediction

Predicting future sales is another approach to make use of the sales data, for this machine learning algorithms, neural networks, or more concretely deep learning models can be used. SFOR-ELM model was constructed and tested on various datasets, and proven to be efficient in sales forecasting in comparison to models previously used in this field. A sailfish optimization algorithm with random disturbance strategy(SFOR) is used to solve the random parameter problem of extreme learning machine(ELM)[24]. ELM is an algorithm that can learn significantly faster than other learning algorithms for neural networks, which is why it is the optimal choice for big datasets[13]. The SFO algorithm is an optimization model inspired by sailfish hunting. "This method consists of two tips of populations, sailfish population for intensification of the search around the best so far and sardines population for diversification of the search space."[19]. SFOR-ELM is a model that uses the SFOR optimization for ELM algorithm.
Another way, to do sales prediction is using the Deep learning model, this model consists of interconnected layers, with input and output layers and multiple hidden layers, trained through gradient descent optimization. These layers perform complex computations and add weight to the output that is passed to the next layer.[23].
Long Short-Term Memory (LSTM) and Autoregressive Integrated Moving Average (ARIMA) are two more algorithms that are being used for sale forecasting, they were compared in predicting retail sales, and even though on short-term prediction(predicting one day ahead) they had almost similar results, on medium-term prediction(predicting one week ahead) LSTM had better accuracy scores.[7]

3 Dataset and preprocessing

3.1 Dataset

The dataset we are working with was provided by a company that produces parts for fans and sells these parts to companies all over the world. The dataset has 113024 rows and consists of 14 features, some of them represent the same thing and some of them are irrelevant to our study, as this dataset was not created only for this study. Data reduction is needed, as the time needed for training, when working with machine learning algorithms, hugely depends on several features in the dataset, so we will remove irrelevant and repeated features. The repeated features are easy to find as they have the same name or the same values for each data entry, irrelevant features are also quite easy to find as they do not have anything to do with sales, such as features that have information about lines, since we predict sales and not when the line is working, or features that represent the same thing, such as Part and Part ID, where both have one specific value for each part we sell, but only need one of them for the models. We will introduce each feature of our dataset and explain what this feature means and what can we use it for in our study.

- Order
  This, as the name suggests, is the ID of the order, every order has a unique order, and this helps us with identifying orders. Each Order ID is featured only once in the dataset.

- Customer
  Similarly with Order, this is a unique Customer ID. Each customer is assigned an ID with their first order, the IDs are featured once or more in the dataset as customers can make more than one purchase.

- Order Date
  This indicates the date of the order, this will be very useful for predicting sales, as well as customer segmentation, as we know the specific dates of each customer's purchases.

- Order Need By
  This is the date by which the product needs to be sent, this is irrelevant in our study, but it would be relevant in studies that try to make the production more efficient by dividing the production between different production lines or factories.

- Order Ship By
  This feature indicates when the order was shipped. It is irrelevant in our study the same way as the Order Need By feature, as it indicates when the product was produced/shipped and it does not help us with predicting the sales.

- Order Amount
  This feature indicates the amount of products that the customer has purchased, this is an important feature for our study, as it shows how much value the customer brings to the company, and the amount he has spent.

- Order
  This is the same Order ID as before, the company tried to merge two datasets and most probably forgot to remove it.

- Line
  This is ID of the line that will produce the part.

- Line Type
  Feature showing the type of the production line. Line and Line Type are irrelevant to our study, the same way as the Order Ship and Need By.

- PartID
  ID of the Part, we will be using this to identify parts sold/produced, this will be used mainly in the sales prediction.

- Part
  This is a string representing the part, we will be using PartID since it is less complicated to work with numbers than with strings, and we just need to differentiate parts from each other.

- Line Qty
  This represents how long or rather how many cycles of the line the order needs to be fully ready for shipping, also a feature that would be used for study on making the production line more efficient.

- Line Ship By & Line Need By
  The last two features are the same as Order Ship/Need By, and so are irrelevant to us.

3.2 Data preprocessing

Preprocessing is the data mining step, that eases the implementation of the process of mining the data and makes it more efficient, as we will not have to focus on problems with the data itself in later phases. Taking into consideration, that well-known and widely used methods of Machine learning are often involved in data mining, the importance of data preprocessing can be easily recognized, as preprocessing always affects the performance of already mentioned machine learning methods [2]. “In industrial applications, more than 80% of the effort is devoted to data preprocessing”[22], this statement just shows how important preprocessing is in machine learning.

3.2.1 Data cleaning

As we can see in Figure 2, some of the data inputs are missing features of when the order shipped and when is it needed, it is caused only in very recent data(at the time of collecting the data) by the order not being handled yet. As this data occurs only in less than 1% of data we can easily solve this problem by removing this data input as it is not a significant data loss. [3]

330517	3847	2023-12-07			104.9	330517	1	PART	T10937	KIS-M25-650x287x240/7-F7-112-T0
330517	3847	2023-12-07			104.9	330517	2	PART	T10942	KIS-M25-650x287x240/8-M5-103-T0
330517	3847	2023-12-07			104.9	330517	3	PART	T10943	KIS-M25-650x287x240/8-M6-75-T0
330518	2647	2023-12-07	2023-12-29	2023-12-29	4209.59	330518	1	PART	T00410	Dazymo sienele
330518	2647	2023-12-07	2023-12-29	2023-12-29	4209.59	330518	2	PART	T27081	Ortakiai ir fasinines dalys
330518	2647	2023-12-07	2023-12-29	2023-12-29	4209.59	330518	3	PART	T01238	Filtro kasete FKAK-315
330518	2647	2023-12-07	2023-12-29	2023-12-29	4209.59	330518	4	PART	T28910	Reguliavimo-matavimo sklende IRIS-315
330518	2647	2023-12-07	2023-12-29	2023-12-29	4209.59	330518	5	PART	T17099	Montavimo medziagos
330518	2647	2023-12-07	2023-12-29	2023-12-29	4209.59	330518	6	PART	T17097	Montavimo darbai (Servisas)
330520	2571	2023-12-07	2023-12-14	2023-12-12	99.95	330520	1	PART	T20030	MPL-K-858x287x46-M5-43-R0-L1-T0
330521	3969	2023-12-07	2023-12-19	2023-12-15	13.24	330521	1	PART	T21642	MPL-PP-330x203x10-G4-8-R0-L1-T0
330521	3969	2023-12-07	2023-12-19	2023-12-15	13.24	330521	3	PART	T17488	MPL-K1-400x190x24-F7-22-R0.0-L1-T0.0
330522	6461	2023-12-07			442.26	330522	1	PART	T09259	KIS-M25-592x392x520/10-F7-85-T0
330522	6461	2023-12-07			442.26	330522	2	PART	T06971	KIS-M25-490x392x520/8-F7-85-T0
330522	6461	2023-12-07			442.26	330522	3	PART	T07444	KIS-M25-490x592x520/8-F7-85-T0.0

Figure 2: Missing data features

3.2.2 Data reduction

We will remove all the features connected to the production line from the dataset as we will not be using those for our study. We will also remove features order need by and order ship by. Other than that we will reduce the dataset more for each algorithm, depending on which features we will need for prediction or profiling, we decide which features the algorithm based on what it does, for example RFM algorithm only need features to compute RFM values so we remove the other features, this is explained in the Models And Algorithms. There are also some data inputs with an order amount of 0, as can be seen in Figure 3 which we will remove, as they are a typo.

112869	330476	3935	2023-12-0	2024-01-1	2024-01-1	7752.17	T29498	V3-KOM-HEPA-ZN-305x305x292-H13S/25-R5.1-L1-T5.3
112870	330476	3935	2023-12-0	2024-01-1	2024-01-1	7752.17	T29502	V3-KOM-HEPA-ZN-610x305x292-H13S/25-R5.2-L1-T5.3
112871	330476	3935	2023-12-0	2024-01-1	2024-01-1	7752.17	T29503	V3-KOM-HEPA-ZN-610x305x292-H13S/25-R5.2-L1-T5.6
112872	330476	3935	2023-12-0	2024-01-1	2024-01-1	7752.17	T29643	V5-KOM-HEPA-ZN-592x592x292-E11S/25-T5.3
112873	330476	3935	2023-12-0	2024-01-1	2024-01-1	7752.17	T29653	V5-KOM-HEPA-ZN-610x610x292-H13S/25-R5.2-L1-T5.6
112874	330477	3387	2023-12-0	2024-01-1	2023-12-2	483	T24722	MPLZ-K-435x190x25-G4-22-6310/180-R1.12-L1-T0
112875	330478	3387	2023-12-0	2023-12-1	2023-12-1	299.2	T27691	PF-2180x770-Lubinis-P1.3
112876	330479	2121	2023-12-0	2023-12-1	2023-12-1	0	T26314	MPLZ-S-150x150x25-G4-22-6310-180-R0-L1-T0-U1
112877	330479	2121	2023-12-0	2023-12-1	2023-12-1	0	T20188	MPL-M-250x250x48-F7-45-R0-L1-T0
112878	330479	2121	2023-12-0	2023-12-1	2023-12-1	0	T21325	MPL-P-250x250x48-F7-43-L1
112879	330479	2121	2023-12-0	2023-12-1	2023-12-1	0	T23810	MPLZC-287x287x25-G4-22-6310/180-R0-L1-T0
112880	330479	2121	2023-12-0	2023-12-1	2023-12-1	0	T18100	MPL-K-287x287x25-F7-22-R0-L1-T0
112881	330479	2121	2023-12-0	2023-12-1	2023-12-1	0	T17145	MPL rėmelis K 46mm (19mm briauna) L-1450
112885	330482	2561	2023-12-0	2024-02-2	2024-02-2	4670.64	T28781	PFMPL-900x835-M5-92-P1.3
112886	330482	2561	2023-12-0	2024-02-2	2024-02-2	4670.64	T28632	PFMPL-1200x610-M5-92-P1.3

Figure 3: Orders with amount equal to 0

3.2.3 Data analysis and further reduction

We analyzed our data to find out the number of orders that were made for each part/product the company sells. This is important for the study as we need to remove the data that does not have enough data entries to properly train the neural networks we are going to use for the sales prediction of each part, this is done only in the experiments where predict sales of parts/products in a month, in the other experiments we use this data, as in the other experiments are not limited by parts, that were sold only few times. From the results in Figure 4 we can see that most of the parts were sold less than five times, we will remove these data entries from the dataset for prediction, as the information we get for those parts is not enough to make accurate predictions and it would skew our results. To do this we will group the data by part IDs and then remove those that do not have enough orders(in our case have less than 5 data entries). This data reduction will be done only for the prediction part of our study, for Customer profiling we will be using the whole dataset, as we are interested in data grouped by Customer, not Part.

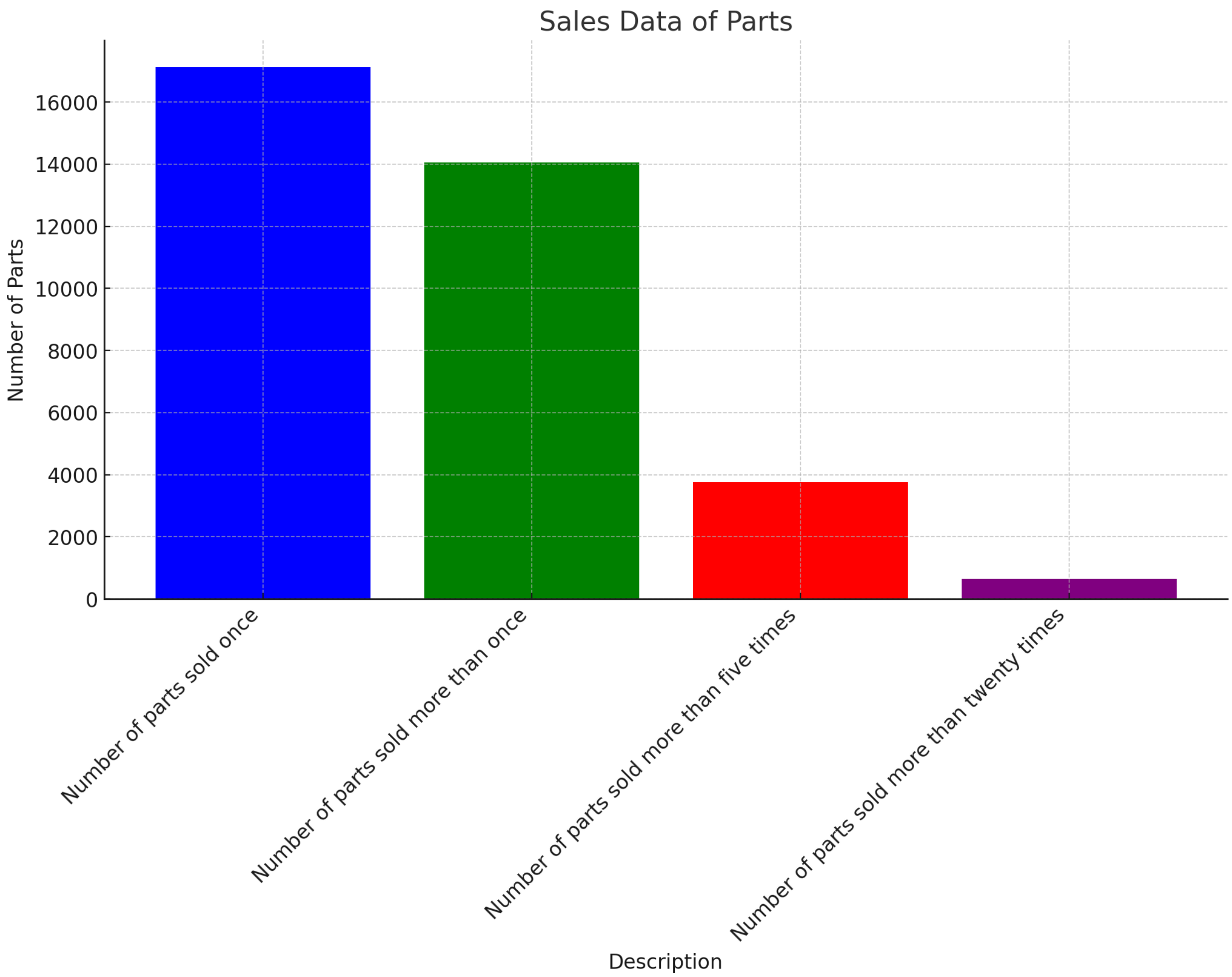

Figure 4: Analysis of Parts that were sold

4 Models and Algorithms

4.1 Customer profiling Algorithms

We are doing two different types of customer profiling. RFM-based clustering and churn prediction using machine learning algorithms. We firstly do RFM algorithm, as we will be using the RFM values as input for the ML algorithms for churn prediction as mentioned in Churn prediction using Logistic Regression, eXtreme Gradient Boosting model, Random Forest and ensemble model. There is also a possibility of using the churn and RFM values for the sales data prediction models, this is mentioned in Conclusion and Future Works and can be seen in Figure 5

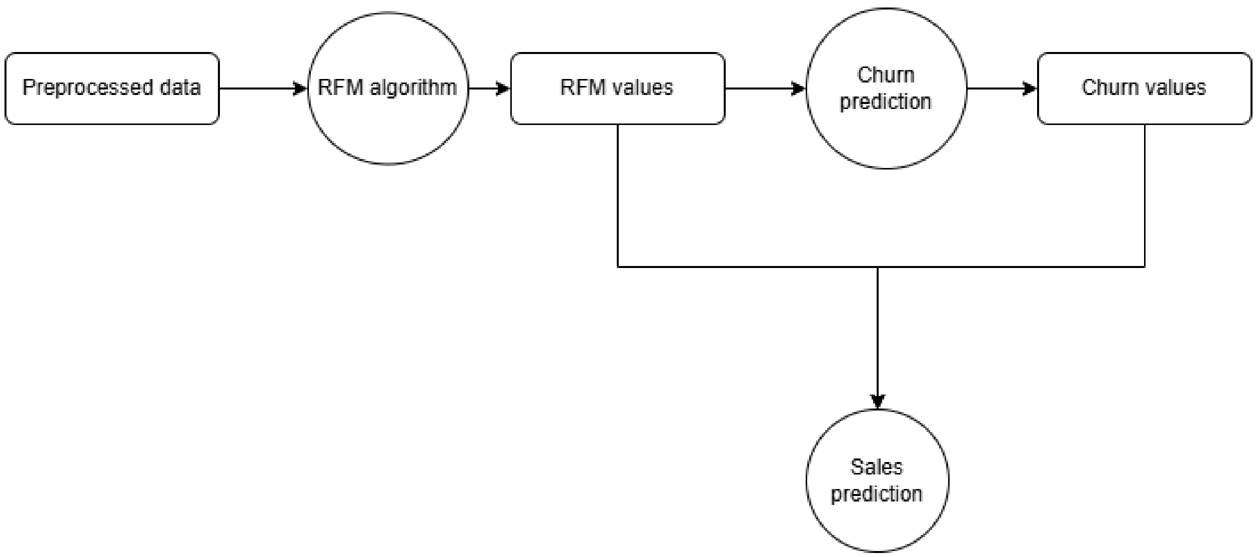

Figure 5: Leveraging outcomes of task to use for another task

4.1.1 Incremental RFM

"Incremental RFM model is designed to classify the value of historical consumers and support the dynamic update of the model"[4]. The three indicators of the RFM model are real-time; that is, the recency, frequency, and monetary. These indicators will change with the passage of time and the generation of orders. We will use the incremental RFM model, as with new data we would need to recompute the RFM values from the whole data set, while with incremental RFM, we just recompute for the new values. As we are working with only a small part of the dataset the company has, approximately 2 years of sales data, the efficiency difference between normal RFM and the incremental version might not be seen as big enough, but with larger datasets the difference rises.
The RFM model compares the customers based on the three indicators mentioned:

- recency - when was the last time the customer made an order? It helps us to divide users to the ones that did not order for a long time and those that were more recently active. We get this value by computing the gap between today and order date of the last order of the customer.

- frequency - how many orders the customer has made together. We can split up the customers into those with the most orders made and those with less. We get this value simply by counting the number of orders the customer has made.

- monetary - how much money has the customer spent in the time frame we are using. This makes it possible to split the customer into those that bring the most money to our company and those that bring less. Our dataset does not contain the prices of all the products, but it does involve how many parts were in each order, so in this study, we assume that each part has the same price (in reality these prices will be different, but the algorithm can be easily adapted to consider this).

The RFM algorithm is quite simple, we compute the average of each of these 3 indicators for all customers, and then we compare each customer's values of indicators with the average, marking the customer with 1 if he has given indicator value over average, otherwise we mark him with 0. This splits the customers into 8 groups:

- 111 - Important value consumers, make orders frequently, are active, and valuable. The company needs to focus on maintaining these customers as they provide the biggest value to the company.

- 110 - General value consumers, active and frequent customers, their value to the company is not as big, so the company does not need to focus on them

- 101 - Important development consumers, these are the active customers that have big value to the company but they do not order as frequently as others. The company should find ways to try to convince them to make more frequent orders

- 011 - Important maintain consumers, these are customers that order frequently and are valuable, but in recent times their activity decreased, the company should try to regain them.

- 100 - General development consumers, have shown recent interest but do not order frequently and do not have big value for the company. Develop strategies to increase both their order frequency and monetary value.

- 010 - General maintain consumers, these customers order frequently, but not in recent times or of great value, it would be nice to keep them, but focusing on them too much is not necessary.

- 001 - Important retain consumers, these customers have significant value to the company but are not active or frequent customers, most probably because they purchased a lot in the past but are not active customers anymore. Focus on retention strategies to bring them back.

- 000 - General retain consumers, customers we do not mind keeping, but since they do not bring much of value to the company they do not need to be focused on.

The order of indicators is Recency, Frequency, Monetary. To implement Incremental RFM, we need to update the RFM values(not the marks) first. There are 3 cases:

- New customer - the customer has no historical RFM values, we compute for him the values as in the normal RFM
- Old customer - the customer has historical RFM values and has made new orders in the meantime. The customer RFM values need to be updated, for frequency and monetary value we just add the old values with the values from the new orders, recency needs to be recomputed from the newest order.
- Old customer with no new orders - this is a special case of the previous case, but since there is no new order recency would not be updated, therefore we need to either recompute recency with today's date or if we know the difference between last update of the RFM values we just add this difference.

After each customer has updated RFM values, we need to compute the averages for the indicators to mark the customers again. The averages can be computed using the old and new RFM values using the formula, where N_2^{new} and N^{old} represent number of new customers and old customers with and without new orders respectively. After computing the new averages we continue the same way as with the RFM algorithm and mark the customers with 0s and 1s the same way, so we compare the values for each customer with the new averages and assign new indicators for RFM. [4]

$$F_{\text{avg}} = \frac{(F_{\text{avg}}^{\text{old}} \times N^{\text{old}} + \sum_{j \in U^{\text{new}}} F_{\text{j}}^{\text{new}})}{N^{\text{old}} + N_2^{\text{new}}} \tag{1}$$

$$M_{\text{avg}} = \frac{(M_{\text{avg}}^{\text{old}} \times N^{\text{old}} + \sum_{j \in U^{\text{new}}} M_j^{\text{new}})}{N^{\text{old}} + N_2^{\text{new}}} \tag{2}$$

$$R_{\text{avg}} = \frac{(\sum_{i \in U'} (R_i^{\text{old}} + 30) + \sum_{j=1}^{N^{\text{new}}} R_j^{\text{new}})}{N^{\text{old}} + N_2^{\text{new}}} \tag{3}$$

4.1.2 Churn prediction using Logistic Regression, eXtreme Gradient Boosting model, Random Forest and ensemble model

Churn is a feature in data that shows whether the customer is still active, or has purchased a product in the specified time frame. This information helps us to divide customers into those that are still likely to be involved with our business and those that are not. The importance of the churn lies in the statement "Companies with loyal, long-time customers can financially outperform competitors with lower unit costs and high market share but high customer churn."[17] So the company needs to know which customers are active and focus on maintaining them while trying to regain those that are not active.
We need to process the data to perform churn prediction, first, we have to calculate churn for each customer in our data set, we do this by setting a value of how many days need to have passed since their last order to be considered inactive, in our case we need to set this value to be quite high since the data has more than 3 years, but with current data, the value

could and should be set to lower numbers. Another change is that we need our data in the format of having one entry in the dataset per customer, we already did this with the RFM algorithm we added churn values for each customer and removed the recency value, as from the recency churn value is calculated. We are going to use Logistic Regression(LR), eXtreme gradient boosting(XGB), and Random Forest(RF) algorithms first separately, and then their combination as an ensemble model. This ensemble approach had the best results in experiments done on multiple datasets from different business sectors. [9]. "Logistic regression is a well-known classification technique for predicting a dichotomous dependent variable"[6] It models a relationship between binary dependant variable and independent variables by using a logistic function. Logistic regression is often used in finance or marketing problems for classification problems, because of simplicity and robustness. XGBoost is a state-of-the-art machine learning model that can solve problems using minimal resources. This model ensemble decision trees sequentially so they correct errors of the previos decision trees. It uses gradient descent optimization to minimize the objective function and update model parameters[5]. Random Forest is a machine learning algorithm usually used for classification problems, it creates multiple decision trees, but each tree is constructed only using randomly selected subset of features. The output is produced by taking majority vote of all trees [1]. For the ensemble model, we will average the scores of each of the models and that is going to be the resulting score of the ensemble model, we are trying to see whether it is worth using the ensemble model with our dataset. We will compare accuracy scores and if the ensemble model has better scores, we will also compare the efficiency of the models, to evaluate if it is worth using.

4.2 Sales prediction

We will be predicting two different scenarios in our experiment, we will predict the total order amount of all customers in a week, and the amount of each part that will be sold in a month. We chose to predict the amount of parts for months for a few reasons. We have more than 30 thousand different parts and 110 thousand orders, which means that one part has on average 3 inputs in our data set, if we split our dataset into weeks, this is simply just too little data, when we split the dataset into months, the predictions become easier and more precise. For the prediction of the total amount ordered, we can predict for weeks, as this does not depend on each part, but on the sum of their order amounts, so the data is sufficient. The model we will be using will be very similar for both of our approaches of predicting, as can be seen in the Figure 6. Data preprocessing involves, as already mentioned removing the parts that were sold less than 5 times, feature engineering where we create new features to increase the accuracy of the model, the new features that we create are mean of order amounts for given part and previous 2 order amounts for given part, as this gives the model better idea of what is the trend for ordering given part, encoding of the part and customer IDs, this is done only for the parts in a month prediction experiment. Then we divided the dataset into the training and testing set, we used split 90:10, 90% of the data for training and 10% for testing. Then we train the model with the parameters that we set, first we set the parameters to values that are typically used for neural network model, such as learning rate of 0,001, number of epoch to 100 or number of hidden layers to 100. After initial experiment, based on the results, we increase or decrease the values of the parameters,

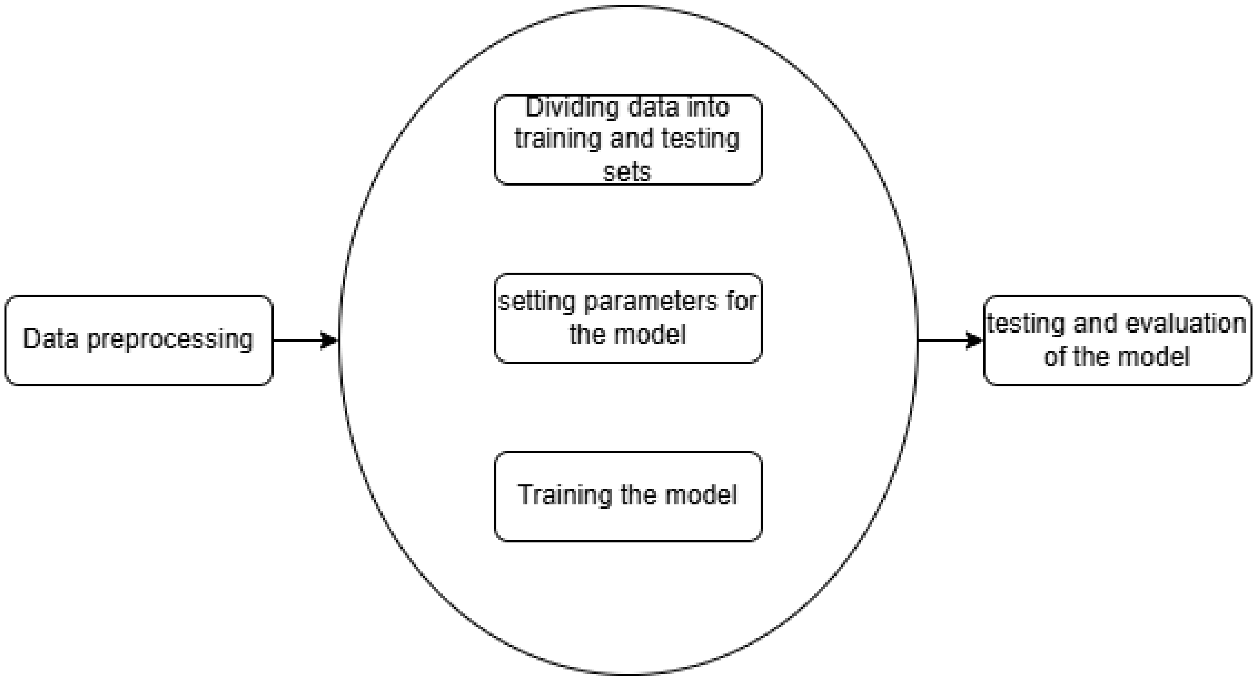

Figure 6: Model diagram

and we test the models and evaluate them using mean absolute error and mean squared error as the functions for evaluation. The outputs of the models are predicted order amount for each part in a month and total amount of orders in week, respectively.

4.2.1 Recurrent Neural Network

We are going to use recurrent neural network(RNN), more precisely Long Short-Term Memory network(LSTM) as one of the algorithms for predicting sales of the company[20]. Since the introduction of the LSTM, many experiments have been done on sequential data with astonishing results. RNN has been designed to handle sequential data, as sales data are sequential, since data entries depend on previous entries, RNNs are ideal for the experiment we doing. "LSTM has the capability of the earlier stages value remembering. So, the past values can be used as the future"[15]. By retaining information over long sequences, LSTM can leverage historical trends and patterns to forecast future sales.

Our model consists of an input layer, taking inputs of dates of orders and other features, an LSTM layer to process the sequential data, one dropout layer, one more LSTM layer to output the final state, one Dense layer to learn complex patterns, and output layer to produce outputs. In a dropout layer neurons are dropped, to prevent overfitting, and the network learning too well and fast[11]. Before we feed the data to this algorithm we use a random forest classifier to predict which parts will have zero sales in a given month. Classifiers are much more accurate in classification problem as predicting zeros than the LSTM algorithm, in our case, we achieved 0.99 accuracies with the random forest algorithm, we only have 4 zeros in the data set that we are using, but the random forest still improves the results of our model, as can be seen in Figure 16. The RF then can be taken as another preprocessing step, and after the data is fed to LSTM, we predict the sales, without the parts/products that we predicted to have zero sales with RF.

4.2.2 SFOR-ELM

"A sailfish optimization algorithm (SFO) algorithm with random disturbance strategy (SFOR) was proposed based on the SFO to improve the prediction effect of clothing sales. The benchmark function test results showed that the SFOR algorithm effectively avoided local extreme points. The SFOR algorithm was used to solve the extreme learning machine (ELM) random parameter problem, and the SFOR-ELM-based online sales prediction model of clothing products suitable for multiple scenarios was constructed."[24] ELM was proposed as a solution to efficiency problems of feedforward neural networks.

ELM randomly chooses hidden nodes and determines output weights of single-hidden layer feedforward neural networks (SLFNs).[13] as can be seen in the Figure 7. ELM is very useful with bigger datasets, as it can learn thousands of times faster than learning algorithms for recurrent neural networks. We will compare accuracy with the LSTM model and also the efficiency, as this plays an important part, especially when the company will do this with the whole dataset of their sales not only the time frame bounded dataset they provided to us. We use a random forest classifier that firstly predicts which parts will have zero sales a month, this is another step of preprocessing, then we feed the data with non-zero parts to the SFOR-ELM model, where we change the number of hidden layers to achieve maximum accuracy, to predict the order amount of each part.

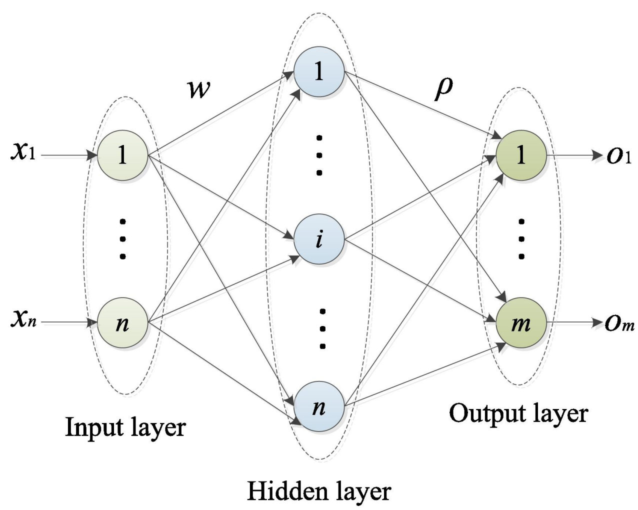

Figure 7: ELM topology adapted from [24]

5 Results of Profiling models

5.1 PROFILING USING RFM METHOD

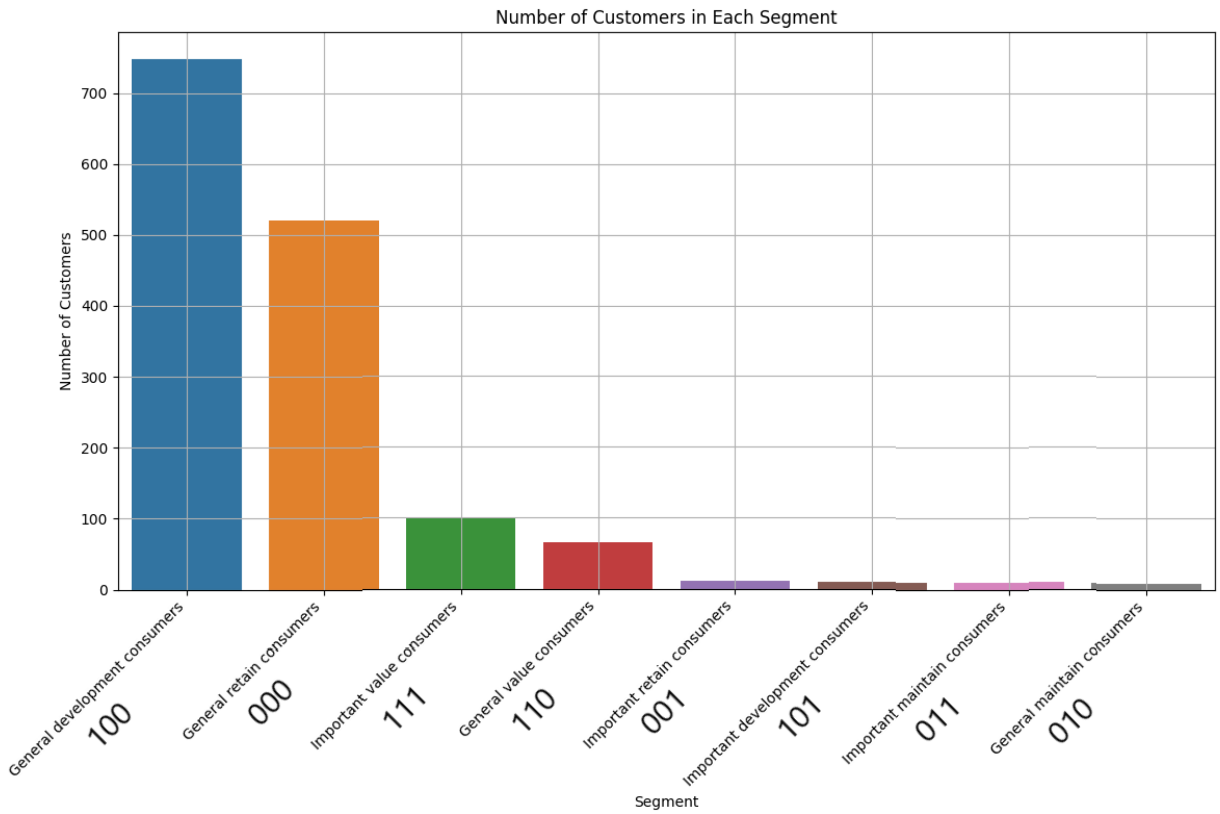

Figure 8: Customers segmentation using RFM model

Figure 8 shows the segmentation of customers based on the RFM model that compares the customer on the recency, frequency, and monetary value for the company. The most represented groups are general development and retain consumers, this suggests that the company has many customers that order only a few times(or at least in the time frame of our dataset). The recommendation would be for the company to find ways to make these customers order more frequently. These two groups will always include a big part of the customers of the company though, since there are always people/companies who need the type of product that the company sells only once in a long time. The other groups with quite significant representation are important value consumers and General value consumers, these are the customers that order frequently and are still active. The only difference is in the amount of money they bring to the company, but this is just a matter of time, as someone who is an active consumer and was recently active, will increase his monetary value over time. The other four groups are represented sporadically, this indicates the fact that these groups represent customers that behave somewhat unconventional, either they are not active anymore, or they make a big order once at a time, but not frequently. Focusing on these customers is quite hard, as their behavior is kind of unpredictable.

5.2 PROFILING BY PREDICTING CHURN OF CUSTOMERS USING DIFFERENT MODELS

Model	**Accuracy**
Logistic Regression (lr)	0.9424
Random Forest (rf)	0.9593
XGBoost (xgb)	0.9492
Ensemble Accuracy	0.9593

Table 1: Comparison of models based on accuracy

Table 1 shows the accuracy values of three different models and an ensemble model, that were used to predict the churn of customers. From the table we can see that, all of the models have really good and similar accuracy, while random forest and ensemble models have the same accuracy, this implies that using the ensemble model is not worth it, since it needs to run all three models and then do voting, which needs more resources than just running the random forest model while getting the same result(with our data, but since all 4 models had similar accuracy results, we can say that on data with similar features as ours, the ensemble model will not be worth to use). The small differences in accuracy are caused mainly by good predictability of the data for churn.

6 Results of Predicting models

In this section we display results of models we chose after doing experiments, results from some other models that showed reasonable results can be found in Appendix. The experiments are done by changing parameter in the models, in the LSTM we change learning rate and number of epochs and in SFOR-ELM we change number of epoch and number of hidden layers. We will evaluate the models using two performance metrics, Mean Squared Error (MSE) and Mean Absolute Error (MAE). MSE computes the squares of the errors and then measures the average of the squares. It gives a higher weight to larger errors, making it sensitive to outliers. A lower MSE indicates a better fit of the model to the data. MAE calculates the average of the absolute errors. MAE treats all errors equally, providing a measure of prediction accuracy without giving extra weight to larger errors. After choosing our model for predicting parts sales, we came with an idea of running the model with sales data only from Important value and General value customers, as data from these customers have more continuity and thus could produce better results. The result of this experiment can be found in Figure 13.

6.1 LSTM

6.1.1 Prediction of number of all orders a week

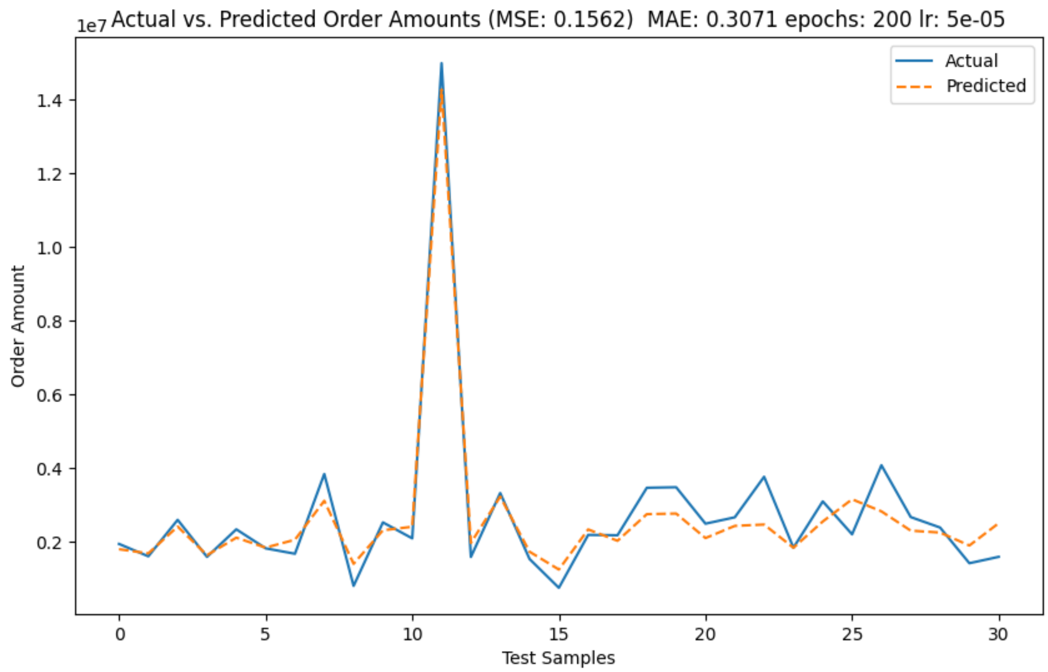

Figure 9: Prediction of Total amount of orders using LSTM

Epochs	**Learning Rate**	**MAE**	**MSE**
200	0,00005	0,3071	0.1562

Table 2: Empirical metrics

In this test, we set the number of epochs to 200 and the learning rate to 0,00005, MSE and also MAE have both decreased compared to the model with a learning rate of 0,0001 indicating increased performance. Validation loss starts to plateau at epoch 180, so the amount of epoch are enough. This model's performance improves as the learning rate is reduced. However, when the learning rate is decreased too much, the model struggles to escape local minima, leading to a significant drop in performance. We chose this model, as it had the best performance when comparing the MSE and MAE values to other experiments we did. In the results we can see that we predicted most of the months quite precisely, even the outlier value, some of the predictions are a bit off, and this might be caused by some unexpected purchases.

6.1.2 Predictions of number of parts sold in a month

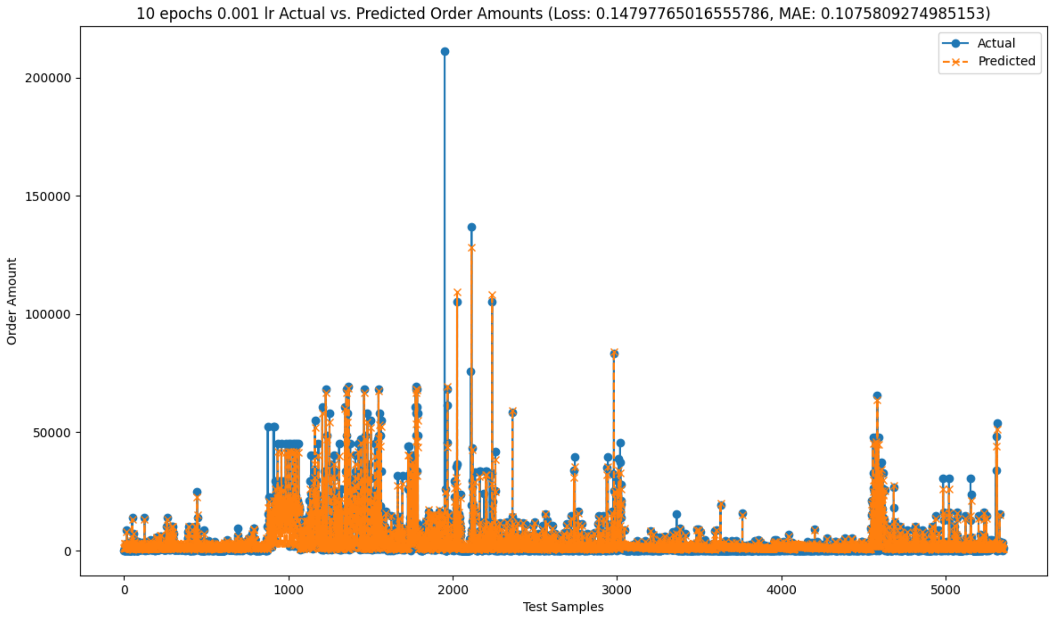

Figure 10: Prediction of amount of parts sold using LSTM

Epochs	Learning Rate	MAE	MSE
10	0,001	0,1076	0.1480

Table 3: Empirical metrics

In this test we set the number of epochs to 10 and and learning rate to 0,001, this model performs similarly to the one with a learning rate of 0,0001. There are slight differences in MSE and MAE, where this model's MSE is lower, while MAE is higher than the other model. In the next test, we will set the learning rate to 0,005 to see whether the slight increase in learning rate increases or decreases performance.
In the end, we chose to this model, as it had the best performance when comparing the MSE and MAE values. The results show that the prediction for each part is quite accurate, almost never overshooting(which is good as we do not want to have too much extra material in warehouses) and sporadically not predicting some outliers, which might be caused by unexpected purchases.

6.2 SFOR-ELM

6.2.1 Prediction of number of all orders a week

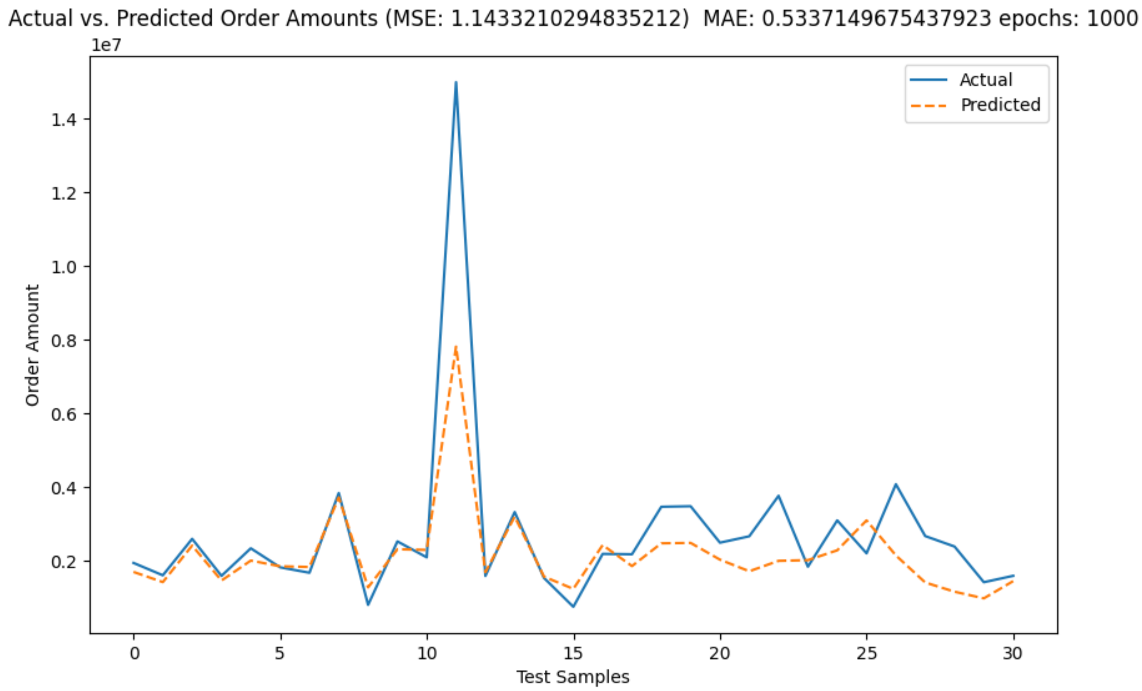

Figure 11: Prediction of Total amount of orders using SFOR-ELM

Epochs	**Hidden layers**	**MAE**	**MSE**
1000	10	0.5337	1.1433

Table 4: Empirical metrics

In this test we set values of hidden layers to 10 and epochs to 1000, we tried many different values for a number of hidden layers and epochs, but this is the closes we could get to the predictions using the LSTM model, this shows that the SFOR-ELM is probably not the best choice for this scenario.

6.2.2 Predictions of number of parts sold in a month

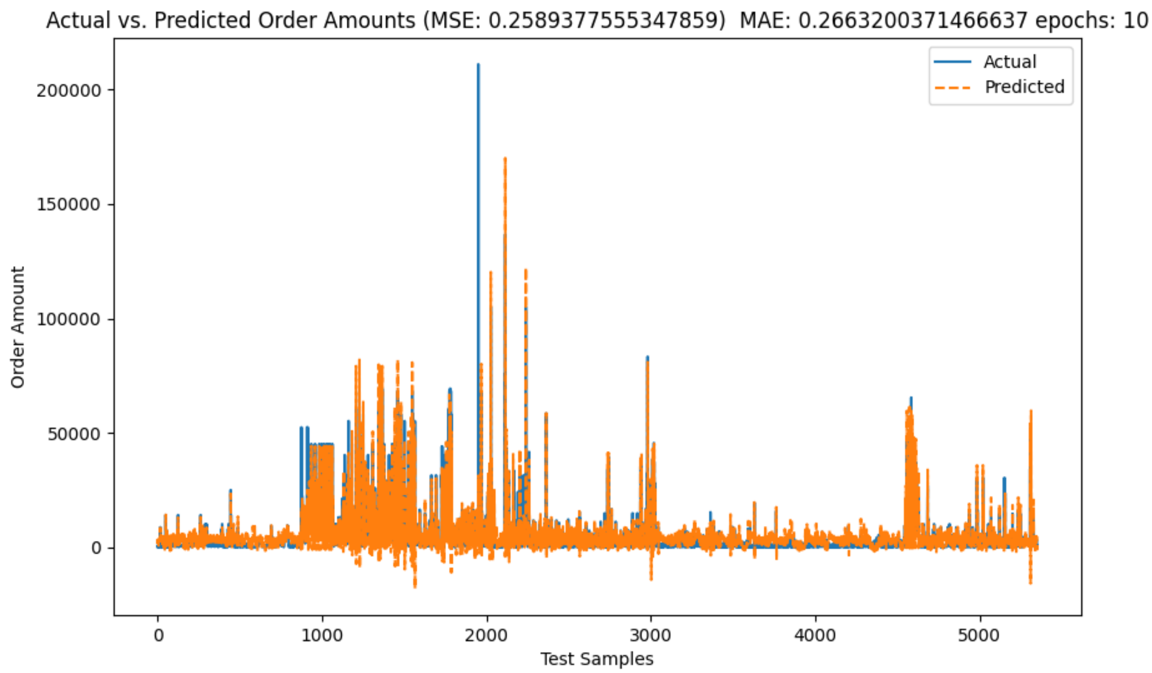

Figure 12: Prediction of amount of parts sold using SFOR-ELM

Epochs	**Hidden layers**	**MAE**	**MSE**
10	50	0.2663	0.2589

Table 5: Empirical metrics

In this test we set values of hidden layers to 50 and epochs to 10, the results significantly improved with the increased number of hidden layers, we will increase the number of layers even more, but each added layer significantly slows the training process, so we need to be careful with improving number of hidden layers.

6.3 Predicting parts sales with General and Important value customers data only

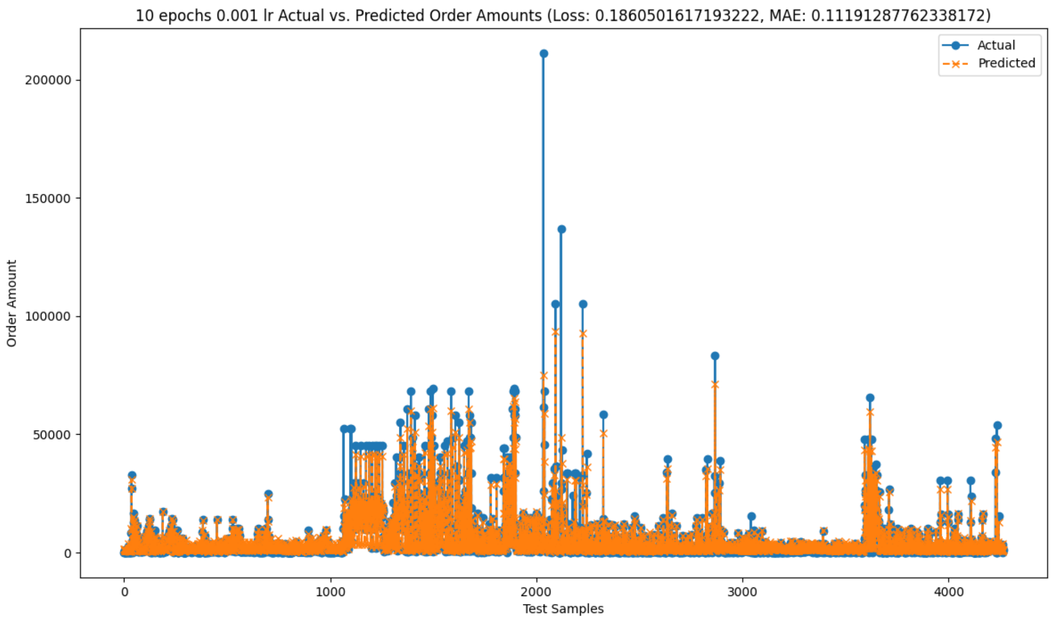

Figure 13: Prediction of amount of parts sold using LSTM with high value customer sales data only

Epochs	Learning Rate	MAE	MSE
10	0.001	0,1119	0.1861

Table 6: Empirical metrics

Here we take only the High-value customers's data, to see whether this improves the results of the LSTM model, we expected that this should improve the result, as the model learns better from data that is more frequent and continuous. The model results did not improve, this is probably caused by the fact that these two types of customers have made almost three-quarters of all orders (91717 out of 122668). Hence, the dataset was not that much different from the previous one, and so the results were not that different, another reason might be that the customers we removed from the data set, might have been easy to predict (they make orders not that frequently, once or twice a year but in the same months with similar order amount).

7 Discussion

Main goal of this thesis was to find a use for the data that company acquired from their sales. In this paper we have proposed to, first use the data for customer profiling to help the company understand the customers better and provide more information on what type of customers the company has. Secondly we proposed to use the data for predicting future sales of the company in two different ways, predicting total amount of orders in a week and predicting amount of orders for each part in a month.

We accomplished the customer profiling using RFM algorithm that groups customer based on their recency, frequency and monetary values. This provides split of customers into 8 groups for the company, where each group has different characteristic, and according to this characteristics different marketing strategies can be used to either maintain or regain customers. A few suggestions to keep important value consumers and general value consumers would be, to seek their feedback to understand their needs and implement their suggestions or to provide premium services for them (discounts for bundles, priority or free shipping or premium customer service). We also predicted churn of customers using ensemble model, this model worked with almost 0.96 accuracy score, which makes it more than usable in practice. Churn of customers is very valuable for the company as it shows customers, who are not going to make more purchases, become passive, and company can either ask them for feedback and find out the reason for their passivity or try to convince them to stay with the company either by providing some benefits for them or by other means.

We provided two solutions for the sales prediction using LSTM model and SFOR-ELM model. The LSTM was able to predict amount of part sold in a month with mean squared error of 0,15 and mean average error of 0,11, while SFOR-ELM with mean squared error of 0,26 and mean average error of 0,27. The prediction using LSTM model are more than usable for the company, even though the model were not able to predict some of the unexpected orders, considering the amount of data we had in our disposal, the results are good and the company can use them for improving their efficiency in storing and acquiring of raw materials as well as for production costs improvement. Secondly, we were able to make the predictions of total amounts ordered in a week using LSTM with mean squared error of 0,16 and mean average error of 0,30 and with SFOR-ELM with mean squared error of 1,14 and mean average error of 0,53. The company will be able to use the predictions from LSTM model to plan their investments better. In the end we can conclude that the choice of SFOR-ELM model was not the best one, or at least not as good as the choosing of LSTM model. The results can also be seen in Table 7 and Table 8

Model	**MAE**	**MSE**
LSTM	0,11	0,15
SFOR-ELM	0,27	0,26

Table 7: Comparision of best performing sales prediction models in predicting amounts of parts sold in a month

Model	**MAE**	**MSE**
LSTM	0,30	0,16
SFOR-ELM	1,14	0,53

Table 8: Comparision of best performing sales prediction models in predicting total amount sold in a week

8 Conclusion and Future Works

We can conclude that we provided a robust solution to the problem, helping the company identify and understand their customers and predict future sales. This, in turn, provides the company with critical information to enhance management efficiency.

However, there are several ways to improve these results and unlock new potential applications for the data the company collects. One significant enhancement would be for the company to gather more comprehensive data about its customers. For example, additional demographic information, purchasing behaviors, and customer feedback could provide deeper insights and allow for more tailored marketing strategies and personalized customer experiences.

In this thesis, we focused primarily on predicting sales. Another valuable application of the data could be optimizing the production process. The dataset includes features containing information about production lines and the duration of production processes. By analyzing this data, the company could improve production schedules and overall efficiency.

Furthermore, the company could explore the use of predictive maintenance. By collecting and analyzing data on machine performance and maintenance history, the company could predict when machines are likely to fail and schedule maintenance proactively, reducing downtime and saving costs.

Lastly, using the results of the profiling models, the RFM model, and Churn prediction, as input into the prediction models could improve the prediction accuracy, and thus the company would have more precise data to work with.

References

[1] Rebecca Abraham, Mahmoud El Samad, Amer M. Bakhach, Hani El-Chaarani, Ahmad Sardouk, Sam El Nemar, and Dalia Jaber. Forecasting a stock trend using genetic algorithm and random forest. *Journal of Risk and Financial Management*, 15(5), 2022.

[2] Stamatios-Aggelos N Alexandropoulos, Sotiris B Kotsiantis, and Michael N Vrahatis. Data preprocessing in predictive data mining. *The Knowledge Engineering Review*, 34:e1, 2019.

[3] Ilham Battas, Ridouane Oulhiq, Hicham Behja, and Laurent Deshayes. A proposed data preprocessing method for an industrial prediction process. In *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, pages 98–103, 2020.

[4] Huihua Chen, Huahong Zuo, Sike Yang, Hailong Wu, Wei Guo, Lina Wang, Xiao Chen, and Yingqiang Su. A data-driven customer profiling method for offline retailers. *Computational Intelligence and Neuroscience*, 2022:8069007, 2022.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[6] Kristof Coussement and Dirk Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, 34(1):313–327, 2008.

[7] Ajla Elmasdotter and Carl Nyströmer. A comparative study between lstm and arima for sales forecasting in retail, 2018.

[8] Zhihua Gan. Research on sales forecasting method based on data mining. In Mohammed Atiquzzaman, Neil Yen, and Zheng Xu, editors, *2021 International Conference on Big Data Analytics for Cyber-Physical System in Smart City*, pages 659–666, Singapore, 2022. Springer Singapore.

[9] Louis Geiler, Séverine Affeldt, and Mohamed Nadif. An effective strategy for churn prediction and customer profiling. *Data & Knowledge Engineering*, 142, 2022.

[10] Dimitris Gkikas and Prokopis Theodoridis. *AI in Consumer Behavior*, pages 147–176. 01 2022.

[11] Asmaa Halbouni, Teddy Surya Gunawan, Mohamed Hadi Habaebi, Murad Halbouni, Mira Kartiwi, and Robiah Ahmad. Cnn-lstm: Hybrid deep neural network for network intrusion detection system. *IEEE Access*, 10:99837–99849, 2022.

[12] MMTM Hassan and M Tabasum. Customer profiling and segmentation in retail banks using data mining techniques. *International journal of advanced research in computer science*, 9(4):24–29, 2018.

[13] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1):489–501, 2006. Neural Networks.

[14] Mahmoud Kasem, Mohamed Hamada, and Islam Taj-Eddin. Customer profiling, segmentation, and sales prediction using ai in direct marketing. 02 2023.

[15] Ashutosh Kumar Dubey, Abhishek Kumar, Vicente García-Díaz, Arpit Kumar Sharma, and Kishan Kanhaiya. Study and analysis of sarima and lstm in forecasting time series data. *Sustainable Energy Technologies and Assessments*, 47:101474, 2021.

[16] Muhammad Aamir Haseeb Ali Muhammad Zulqarnain Muhammad Faheem Mushtaq, Urooj Akram. Neural network techniques for time series prediction: A review. *JOIV: International Journal on Informatics Visualization*, 3(3):314–320, 2019.

[17] F.F. Reichheld and W.E. Sasser Jr. Zero defections: quality comes to services. *Harvard business review*, 68(5):105 – 111, 1990. Cited by: 3314.

[18] İbrahim SABUNCU, Edanur TÜRKAN, and Hilal POLAT. Customer segmentation and profiling with rfm analysis. *Turkish Journal of Marketing*, 5(1):22, 2020.

[19] S. Shadravan, H.R. Naji, and V.K. Bardsiri. The sailfish optimizer: A novel nature-inspired metaheuristic algorithm for solving constrained engineering optimization problems. *Engineering Applications of Artificial Intelligence*, 80:20–34, 2019.

[20] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[21] Treehouse techonology group. How to use data mining for business analytics. *Treehouse techonology group*, n.d.

[22] Xiaolong Xu, Weizhi Chong, Shancang Li, Abdullahi Arabo, and Jianyu Xiao. Miaec: Missing data imputation based on the evidence chain. *IEEE Access*, 6:12983–12992, 2018.

[23] Fong-Ching Yuan and Chao-Hui Lee. Intelligent sales volume forecasting using google search engine data. *Soft Computing*, 24, 02 2020.

[24] Bo Zhang, Ming-Lang Tseng, Lili Qi, Yuehong Guo, and Ching-Hsin Wang. A comparative online sales forecasting analysis: Data mining techniques. *Computers Industrial Engineering*, 176:108935, 2023.

9 Appendix

9.1 LSTM

9.1.1 Predictions of number of parts sold in a month

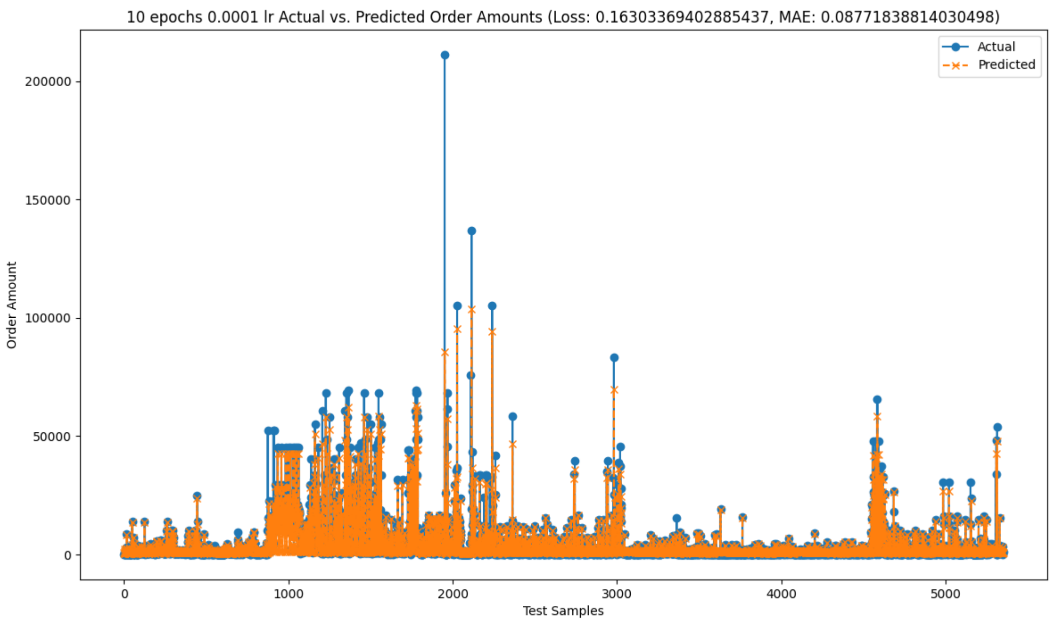

Figure 14: Prediction of amount of parts sold using LSTM

Epochs	Learning Rate	MAE	MSE
10	0,0001	0,0877	0.1630

Table 9: Empirical metrics

In this test we set a number of epochs to 10 and and learning rate to 0,0001, the training Loss decreases consistently over epochs, indicating the model is learning from the training data. Validation Loss also decreases over epochs initially but plateaus around epoch 9, suggesting the model's performance on unseen data is improving and that we do not need more epochs for our training. Test Loss and MAE indicate that the model performs reasonably well on the test data, with a relatively low error margin.

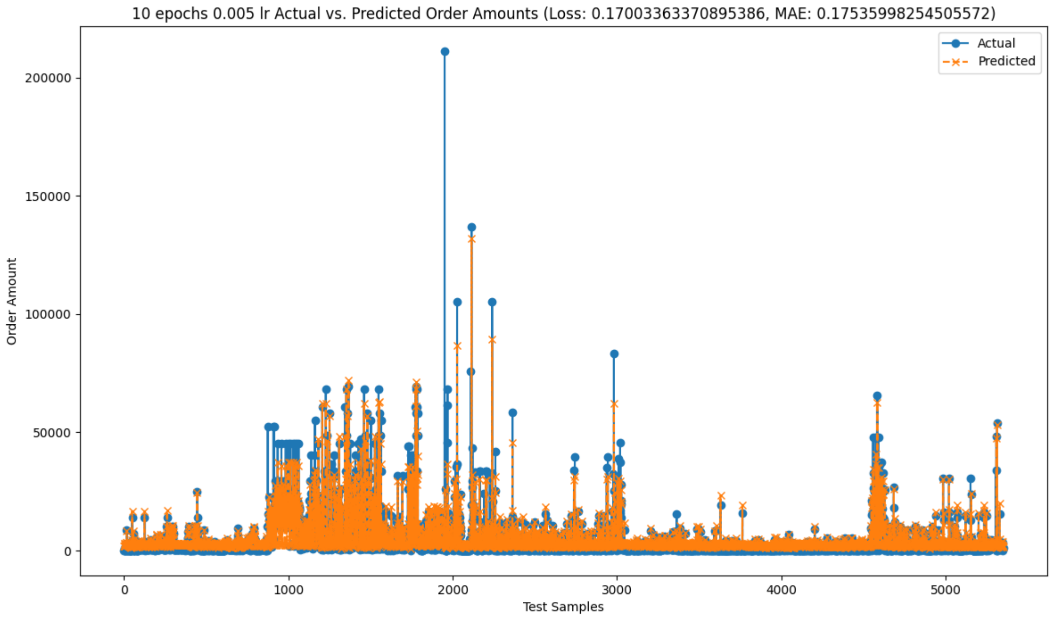

Figure 15: Prediction of amount of parts sold using LSTM

Epochs	**Learning Rate**	**MAE**	**MSE**
10	0,005	0,1754	0.1700

Table 10: Empirical metrics

In this test we set number of epochs to 10 and and learning rate to 0,005, this model performs worse than both model with 0,001 and 0,0001 learning rate, this indicates that setting learning rate to 0,001 makes the model converge more controlled, which improved the performance, while learning rate of 0,0001 makes the model's updates smaller and thus fine tune the predictions at loss of effectiveness of minimizing the overall loss, The reason why learning rate of 0,005 performs worse is that it is not low enough for fine tuning and it is too high and makes the updates too big.

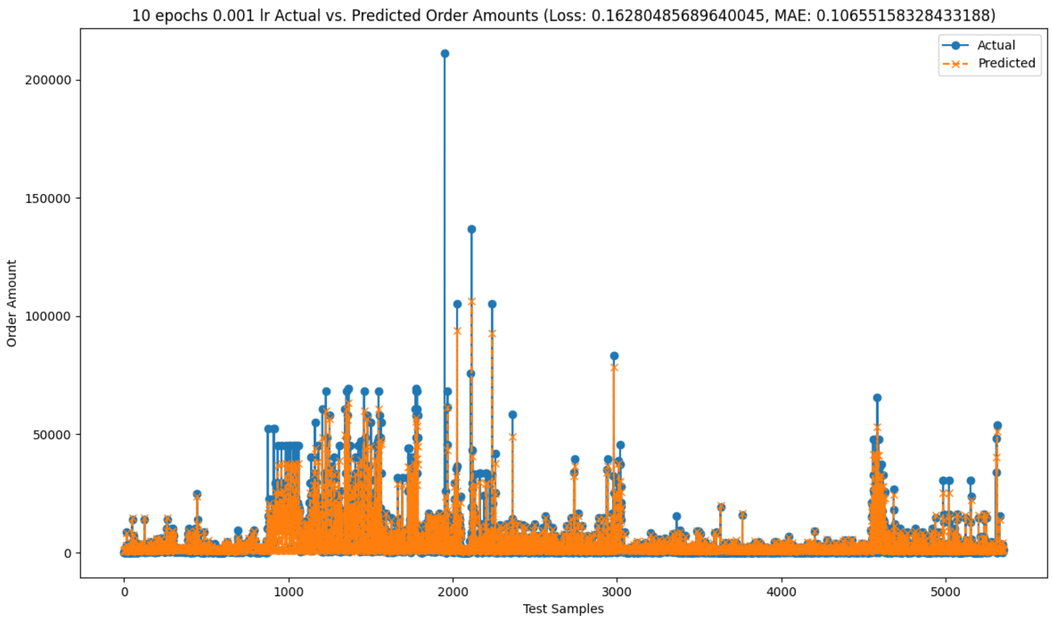

Figure 16: Prediction of amount of parts sold using LSTM without Random Forest

Epochs	**Learning Rate**	**MAE**	**MSE**
10	0,001	0,1066	0.1628

Table 11: Empirical metrics

In this test we removed the Random Forest classifier for predicting zeros in our data from the model, the MAE and MSE have increased, so the performance of the model is not better than with the Random Forest predicting zeros in the model.

9.1.2 Prediction of number of all orders a week

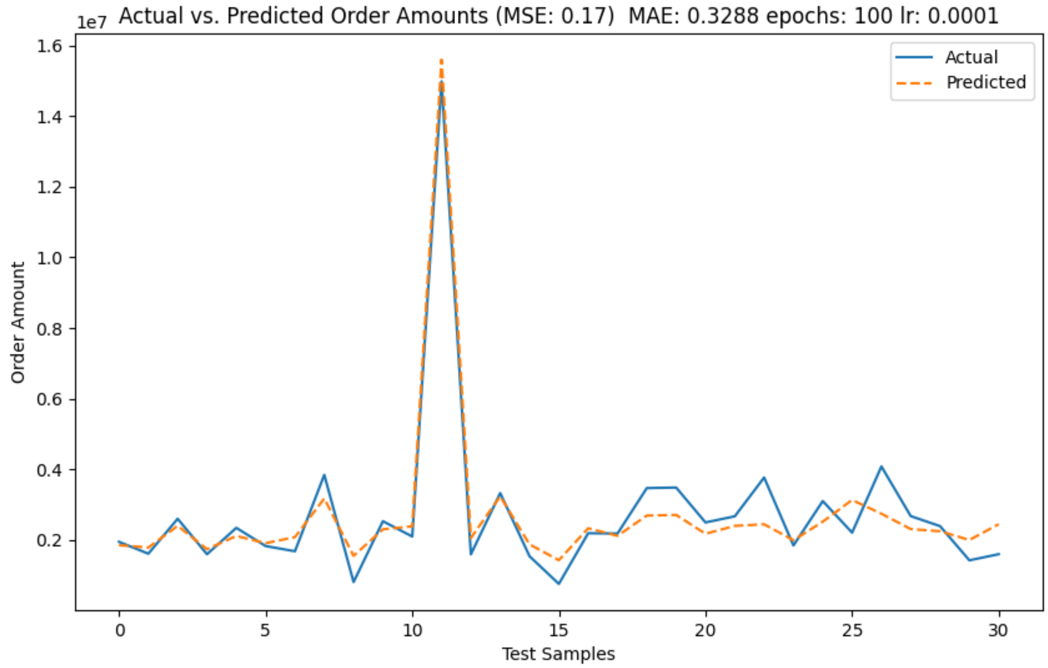

Figure 17: Prediction of Total amount of orders using LSTM

Epochs	**Learning Rate**	**MAE**	**MSE**
100	0,0001	0,3288	0.17

Table 12: Empirical metrics

In this test we set number of epochs to 100 and and learning rate to 0,0001, the training Loss decreases consistently over epochs until epoch 80 where it starts to plateau, indicating the model is learning from the training data until then. Validation Loss also decreases over epochs initially but plateaus around epoch 80, suggesting the model's performance on unseen data is improving and that we do not need more epochs for our training. Test Loss and MAE indicate that the model performs reasonably well on the test data, with a relatively low error margin. We will test the model with learning rate of 0,001 and 0,00005 to see whether either of these changes makes the model more accurate.

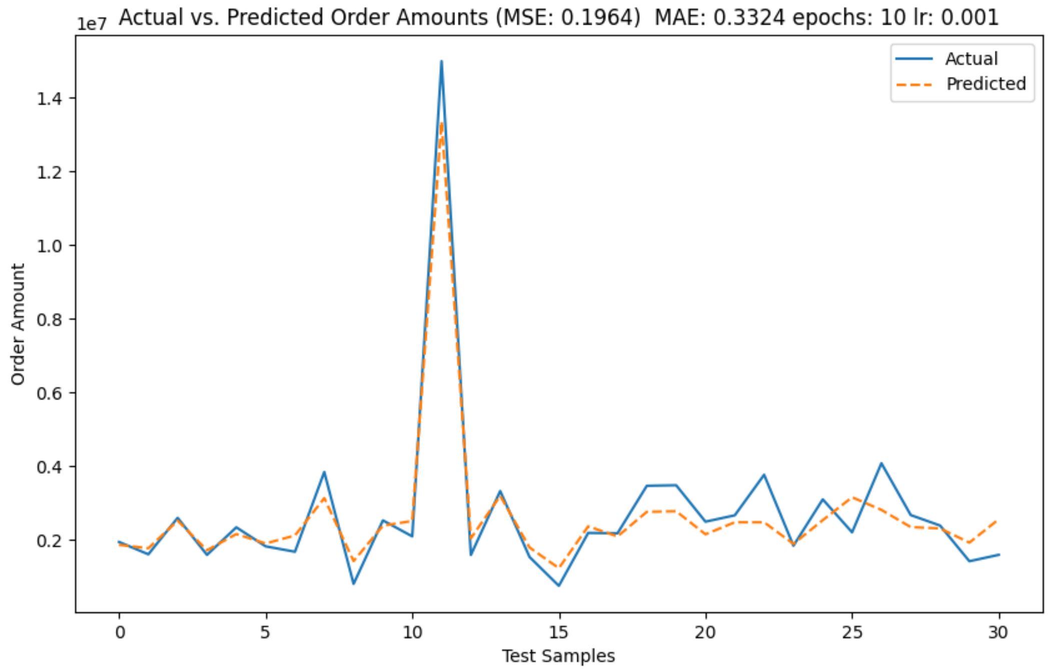

Figure 18: Prediction of Total amount of orders using LSTM

Epochs	Learning Rate	MAE	MSE
10	0,001	0,3324	0.1964

Table 13: Empirical metrics

In this test we set number of epochs to 100 and and learning rate to 0,001, MSE and also MAE have both increased indicating decreased performance. Validation loss plateaus at epoch 9, so the amount of epoch are enough.

9.2 SFOR-ELM

9.2.1 Predictions of number of parts sold in a month

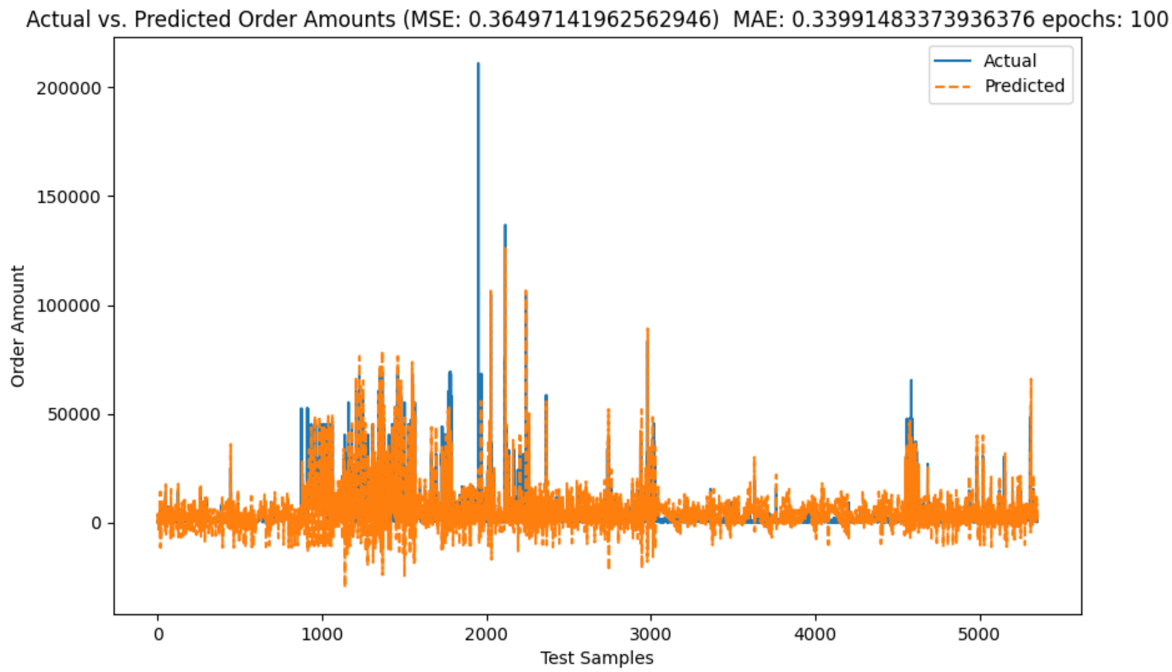

Figure 19: Prediction of amount of parts sold using SFOR-ELM

Epochs	**Hidden layers**	**MAE**	**MSE**
100	10	0,3399	0.3650

Table 14: Empirical metrics

In this test we set values of hidden layers to 10 and epochs to 100, the results are clearly worse than the ones using LSTM model, so we will change the value of hidden layers to increase the model's capacity and performance.

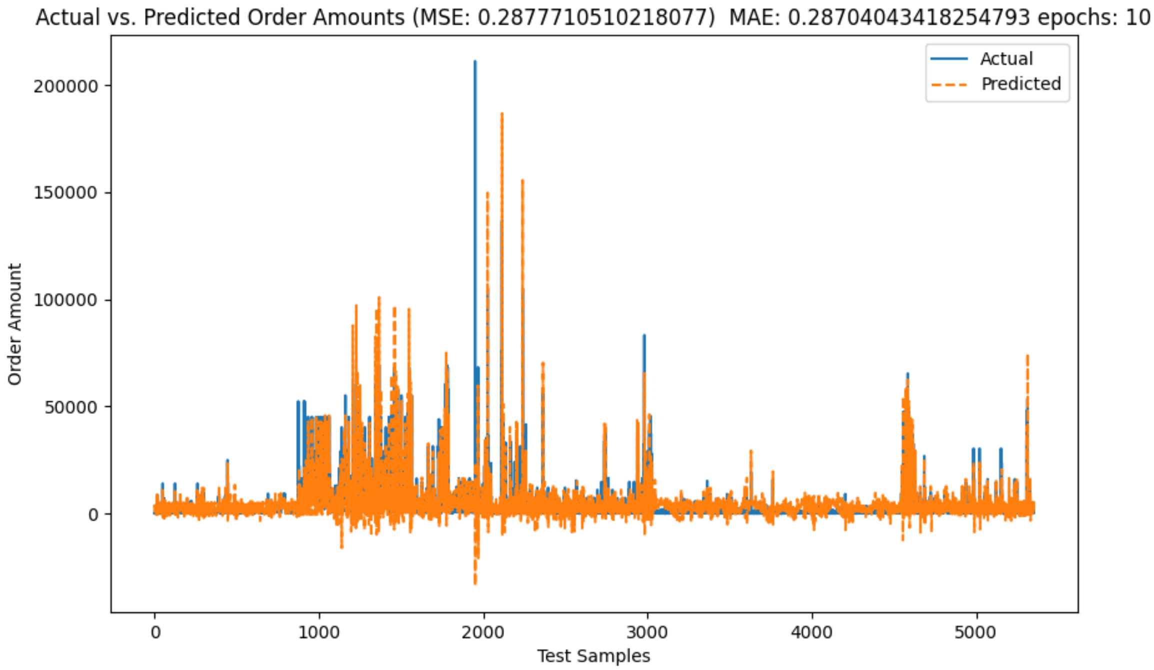

Figure 20: Prediction of amount of parts sold using SFOR-ELM

Epochs	Hidden layers	MAE	MSE
10	75	0,2870	0.2878

Table 15: Empirical metrics

In this test we set values of hidden layers to 75 and epochs to 10, the result did not improve which indicates either overfitting or numerical instability due to calculation of output matrix.