TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Igor Roos 204037IABM

Sales Forecasting Using Machine Learning Methods: the Case of B2B Company

Master's thesis

Supervisor: Avar Pentel
MSc

Tallinn 2022

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Igor Roos 204037IABM

Läbimüügi Prognoosimine Masinõppe Meetoditega B2B Ettevõtte näitel

Magistritöö

Juhendaja: Avar Pentel
MSc

Tallinn 2022

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Igor Roos

10.05.2022

Abstract

The topic of the master's thesis is "Sales Forecasting Using Machine Learning Methods: The Case of B2B Company". The thesis aims to find the best features and methods based on machine learning techniques for sales forecasting in short-term perspective of 31 days. This topic was formulated at the initiative of the thesis author. The data for the work was collected at Estonian private wholesale company, which is part of the global group companies. This research proposes a methodology for sales margin forecasting based on historical sales data of last three years. Using of regression models with and without time series components by such machine learning algorithms as Linear, Support Vector and Random Forest regressions. Additionally, the research present feature engineering techniques for raw data transformation process from big data to suitable for research multivariable time series dataset.

As a result of the work, it was found that the model generated by Support Vector regression algorithm made the most accurate predictions. This result was acquired by the regression model based on correlated features without time series components and by the regression model based on time series components with overlay data. The Linear regression model also has acceptable accuracy result, but only for model without time series components. In the context of data sufficiency, we assume that model's accuracy could be improved by adding new features to dataset. Which may contain, for example, geographical distribution of sales.

Keywords: sales forecasting, machine learning model, feature engineering, regression, time series data.

This thesis is written in English and is 45 pages long, including 6 chapters, 15 figures and 15 tables.

Annotatsioon

Läbimüügi Prognoosimine Masinõppe Meetoditega B2B Ettevõtte Näitel

Magistritöö teema on "Läbimüügi Prognoosimine Masinõppe Meetoditega B2B Ettevõtte Näitel". Lõputöö eesmärk on leida masinõppe tehnikatel põhinevad parimad atribuudid ja meetodid läbimüügi lühiajaliseks prognoosimiseks 31 päeva perspektiivis. See teema sõnastati autori algatusel. Andmed töö jaoks koguti Eesti hulgimüügiettevõttes, mis kuulub globaalsesse rahvusvahelisse kontserni. Selles uurimistöös pakutakse välja läbimüügi prognoosimise metoodika, mis põhineb kolme aasta ajaloolistel müügiandmetel. Kasutatud on regressioonimudelit aegreaga ja ilma ning mudelid genereeri selliste masinõppe algoritmide abil nagu lineaarne, tugivektori ja juhusliku metsa regressioon. Lisaks rakendatakse uurimistöös tunnuste hõive meetodit suurandmete teisendamiseks aegridade andmeteks.

Töö tulemusena leiti, et tugivektori regressiooni algoritm toimis kõige paremini ja andis kõige täpsemad prognoosid. Sellise tulemuse andis regressioonimudel, mis põhineb korreleeritud atribuutidel ilma aegrea komponentideta ja regressioonimudel, mis põhineb aegrea komponentidel koos ülekatteandmetega. Lineaarse regressiooni mudelil on ka vastuvõetava täpsusega tulemus, kuid ainult ilma aegrea komponentideta mudelil. Andmete piisavuse kontekstis eeldame, et mudeli täpsust saab veel tõsta andmestikku uute atribuutide lisamisega, mis võivad sisaldada, näiteks, müügi geograafilist jaotust.

Märksõnad: Müügi prognoosimine, masinõppe mudel, tunnusehõive, regressioon, aegridade andmed

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 45 leheküljel, 6 peatükki, 15 joonist, 15 tabelit.

List of abbreviations and terms

AI	Artificial Intelligence
ML	Machine Learning
B2B	Business-to-Business
CRISP-DM	Cross-Industry Standard Process for Data Mining
ERP	Enterprise Resource Planning
KAM	Key Account Manager
TS	Time Series regression model
TS+O	Time Series regression model with overlay data
LR	Linear Regression model
SVR	Support Vector Regression model
RF	Random Forest Regression model
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
STD	Standard Deviation

Table of contents

List of figures

List of tables

1 Introduction

Artificial intelligence (AI) technologies are already revolutionizing many businesses and processes. Sales business are not an exception. Especially sales are one of the areas that AI and machine learning software companies are cooperating with enterprises. One of the possibilities of applying AI technologies are implementation of various forecasts. It's hard to overstate how important it is for a company to produce an accurate sales forecast. Privately held companies gain confidence in their business when leaders can trust forecasts.

Forecast outcomes using historical data to predict future results. The accuracy of those predictions depends on the system being used and the quality of the data. As example is that, with the right inputs in the past and present, AI is capable to showing who is most likely to buy in the future [1].

Machine Learning (ML) can help to discover the factors that influence sales and estimate the number of sales in the near future. This way if the underlying trends change, the model can be retrained, and learn these changes. Also, statistical learning algorithm can discover patterns missed by business analysts [2].

Without proper forecasting, many business decisions are based on unreliable estimates or instinct – which leads to many inefficiencies and missed opportunities. Instead of applying assumptions and a complex set of rules, ML models learn patterns from the data to generate predictions. In this context, most accurate sales forecasting has great potential to generate business value.

1.1 Statement of the Problem

The difficulties of sales forecasting come into play because the future is uncertain. There are a number of mitigating factors affecting sales forecasting including consumer demand, customer preferences, fluctuations in the economy, and even global events.

One of the challenges of forecasting is finding the number of previous events that should be considered when making predictions about the future. Forecasting is based on a premise of data requirement and the application of the data in projecting future sales. A sales forecast can only be as good as the data it is based on.

According to Amazon's time series forecasting principles [3], forecasting is a hard problem for two reasons:

- Incorporating large volumes of historical data, which can lead to missing important information about the past of the target data dynamics.
- Incorporating related yet independent data (holidays/events, locations, marketing promotions)

Besides these, one of the central aspects of sales forecasting is that accuracy is key:

- If the forecast is too high, it may lead to over-investing and therefore losing money.
- If the forecast is too low, it may lead to under-investing and therefore losing opportunity.

Which sales forecasting methods are more accurate depends on business. In B2B sales business are many uncertainty and unpredictable trends which makes forecasting even more difficult. Regardless of what forecasting method are we used, all available data must be clean and as accurate as possible.

1.2 Objective of the Research

The main objective of this research is to find the best features and models for sales forecasting using Machine Learning methods. We used available historical sales data of a wholesale B2B company to forecast its sales margin up to one month in advance.

We train and test regression models with different feature datasets and compare the results. By the proposed methodology we create the next models:

- Regression models based on dataset with correlated features without time series components. In our research we refer to them as **regression models**. Those models are focused on cause-effect relationships in data.

- Regression models based on univariate data and its time series components. In our research we refer to them as **time series regression models (TS)**. Those models are observed one target variable across time. Focus on patterns based on time, i.e., trend, seasonal and cyclic components.
- Regression models based on same principles as in second case but with additional overlay data component. Such models we name as **time series regression models with overlay data (TS+O)**. Those models combine both previous approaches.

On our research core is a solution of applying Supervised Machine Learning approaches. The first approach consists in selecting features and applying machine learning algorithms to a multidimensional dataset. By this we define a pattern in historical data and those patterns help us to generate possible overlay data for future forecasting process. The second approach is a time series data transformation to regression problem solution. We use a time series components of target variable as the new features in regression models.

As a result, we evaluate all experimental cases and choosing the most accurate algorithm and the best method for B2B sales forecasting in a particular case.

1.3 Research Questions

From the above overview there are three groups of research questions this work will be seeking answers for:

1. Questions about data.

    - Find a possible pattern in historical data and evaluate data sufficiency. What patterns were found? How they affect to forecasted target variable?

    - How does the quality and availability of historical data affect to the forecasting performance?

2. Machine Learning approaches and algorithms questions.

    - Define a best approach for forecasting problem in particular case. It is regression or time series problem?

- Which machine learning algorithm gives the best forecasting performance?
- How are time series components affects to the model performance?
- How the adding of overlay data as possible input feature affects to the forecasting result?

3. Practical prospect of using machine learning approaches in a particular company.
    - What opportunities opens implementation of ML for the company?

2 Background

2.1 Prediction or Forecasting

In supervised learning, we are often concerned with prediction. However, there is also the concept of forecasting [4].

Prediction is concerned with estimating the outcomes for unseen data. For this purpose, you fit a model to a training data set, which results in an estimator $f(x)$ that can make predictions for new samples

Forecasting is a sub-discipline of prediction in which we are making predictions about the future, on the base of time series data. Thus, the only difference between prediction and forecasting is that we consider the temporal dimension. An estimator for forecasting has the form $f(x_1,\ldots,x_t)$ where $x_1,\ldots,x_t$ indicate historic measurements at time points $1,\ldots,t$, while the estimate relates to time point $t+1$ or some other time in the future. Since the model depends on previous observations, x_i, this is called an autoregressive model.

2.2 Sales Forecasting Methods

The appropriate forecasting methods depend largely on what data are available. According to R.J. Hyndman and G. Athanasopoulos [5], there are two types of forecasting methods:

- Qualitative methods
- Quantitative methods

If there are no data available, or if the data available are not relevant to the forecasts, then qualitative forecasting methods must be used. These methods are not purely guesswork, there are well-developed structured approaches to obtaining good forecasts without using historical data.

Quantitative forecasting can be applied when two conditions are satisfied:

- Numerical information about the past is available.
- It is reasonable to assume that some aspects of the past patterns will continue into the future.

2.2.1 Qualitative Methods

Qualitative methods are used when there is limited data available [6]. For example, we can use qualitative techniques when we introduce a new product in the market. There is limited data about the product to use in forecasting the future. The techniques employ human judgment and rating schemes to use the qualitative information and turn it into quantitative estimates. The objective of the method is to bring together logically and systematically all the judgments and information regarding the factors being estimated. We can use qualitative techniques where penetration rates and market acceptance of a product are uncertain.

There are several qualitative techniques which include:

- Panel consensus
- Delphi method
- Expert opinion
- Focus groups
- Market research
- Historical analogy
- End-use method

2.2.2 Quantitative Methods

Most quantitative prediction problems use either time series data (collected at regular intervals over time) or cross-sectional data (collected at a single point in time) [5].

Quantitative sales forecasting methods use data and statistical formulas or models to project future sales. The most popular quantitative methods are time series analysis and casual method.

A time series analysis model involves using historical data to forecast the future. It looks in the dataset for features such as trends, cyclical fluctuations, seasonality, and behavioural patterns. The three key general ideas that are fundamental to consider, when

dealing with a sales forecasting problem tackled from a time series perspective. Is repeating patterns, static patterns, and trends [7]. This method requires chronologically ordered data.

These methods include such techniques [5], [6] as:

- Simple Moving Average.
  By extrapolating data from a set period of time and using this to predict sales over the same period of time in the future.
- Exponential Smoothing.
  Using a forecasting process making an exponentially considered average to predict future.
- Box-Jenkins.
  Applies auto-regressive moving average (ARMA) or auto-regressive integrated moving average (ARIMA) models to find the best fit of a time series model to past values of a time series.
- Trend projection.
  Picking up on trends of past sales to predict similar fluctuations in the future.
- X-11.

The casual method looks at the historical cause and effect between different variables and sales. Causal techniques allow you to factor in multiple influences, while time series models look only at past results. With causal methods, we usually try to take account of all the possible factors that could impact your sales, so the data may include internal sales results, consumer sentiment, macroeconomic trends, third-party surveys, and more [8].

These methods include such techniques [6] as:

- Linear or multiple regression,
- Econometric,
- Leading indicators,
- Input-Output models,
- Diffusion index,
- Life-cycle analysis.

2.3 Machine Learning Approaches to Sales Forecasting

Considered that in the field of Machine Learning, sales forecasting is a time series regression problem. A regression is any task concerned with the estimation of a continuous quantity. Time series regressions are a particular case of regression, with an additional time dimension. There are two main types of time series regressions models: auto-regressive models [9] and multivariate models [10].

Auto-regressive models are based on univariate time series datasets. These are datasets where only a single variable is observed at each time. Models predict future sales solely based on past sales values. The most popular include ARIMA [5], Seasonal Auto-regressive Integrated Moving Average Exogenous (SARIMAX) [11], and Exponential Smoothing models [12].

Multivariate models are based on multivariate time series datasets. These are datasets where two or more variables are observed at each time. The most popular models include Linear Regressions [13], Neural Networks [13], Decision Tree-based methods [15], Support Vector Machines [16] and Vector Auto Regression [17].

2.4 Literature Review

The general problem of sales forecasting and B2B sales forecasting as a special case a widely discussed in the scientific community. Many papers are presented to the public over the past few years with machine learning technologies development. In the perspective of this research, the author would like to point out some of them.

M. Bohanec with colleagues in 2015 proposes a methodology for incorporating supervised machine learning in B2B sales forecasting. The goal of their paper was to investigate the possibility to develop a classification model, based on B2B sales history, which supports forecasting process and provides transparent reasoning, supported by machine learning techniques. Authors convinced that the research presented in their paper has positively confirmed viability of the novel approach [18].

N. S. Elias with S. Singh in 2018 [19]. A. Krishna with colleagues in 2018 [20] and Dr. C. Shyamala with colleagues in 2021 [21] are analysed sales forecasting for retail using

different machine learning algorithms. In all articles authors confirmed that several algorithms have very good accuracy for the best implementations.

G. Nunnari with colleague in 2017 modelling and forecasting of retail sales time series by using two neural network-based approaches. The models performance shown that there is a benefit in using these kinds of models [21].

S. Cheriyan and S. Mohanan in 2018 have concluded that an intelligent sales prediction system is required for business organizations to handle enormous volume of data. Business decisions are based on speed and accuracy of data processing techniques. Machine learning approaches highlighted in this research paper will be able to provide an effective mechanism in data tuning and decision making [23].

B. M. Pavlyshenko in 2019 was study the usage of machine-learning models for sales predictive analytics. He considered different machine learning approaches for time series forecasting. As a conclusion author points out that Sales prediction is rather a regression problem than a time series problem. The use of regression approaches for sales forecasting can often give us better results compared to time series methods. One of the main assumptions of regression methods is that the patterns in the historical data will be repeated in future [24].

3 Methodological Framework

This section provides an explanation of the methodology used to achieve the research objectives. The common research pipeline is structured according to the developed framework of supervised machine learning. The research workflow has four main stages as illustrated in Figure 1.

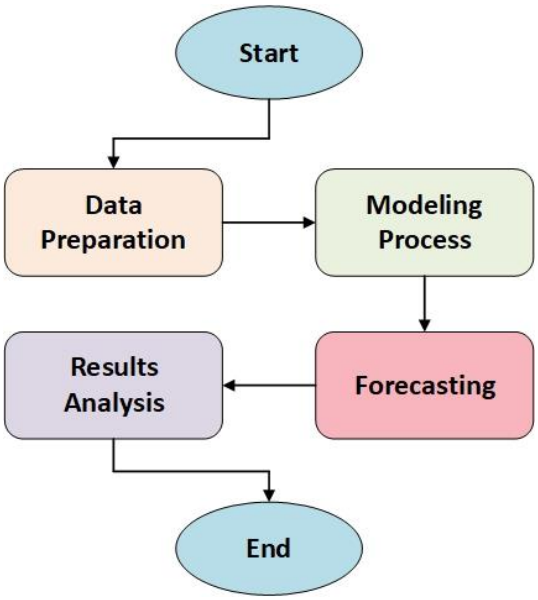

Figure 1. Research workflow

The first stage in this methodology is the data preparation. We create a closed sales deals dataset based on historical sales data. This stage requires identifying variables which describing a sales in best way and features engineering by the values aggregation on time. The objective of the second stage is to identify the features that have best predicting sales potential and training the models by different algorithms. The third stage is a forecasting process by the implementation of different models with different advanced configurations. The final stage is the evaluation results analysis and best models' description.

3.1 Research Software Tools Selection

As a preliminary stage of research was determinate software products with the help of which the research objectives will be achieved. The main criteria for selection are:

- Availability to realize research tasks on each stage.
- Preferably with full functionality free of charge.
- Minimal use of programming languages in modelling processes.

The following two software products were used: Microsoft Excel (Version 2203) and Weka (Version 3.8.3).

Microsoft Excel was chosen as data processing and simple data analysis creation tool. It is easy to get started and free of charge for students. One of the most popular software packages for business and study. It can help you understand the meaning of many operations before further learning other tools (such as Python and R). Despite all the advantages, there are also limitations that can complicate the progress of tasks. The main limitations considering this work for example is situation of stuttering when the amount of data is large [25] [26].

Weka (Waikato Environment for Knowledge Analysis), developed at the University of Waikato, New Zealand, is free software licensed under the GNU General Public License. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to these functions. The main advantages of Weka in terms of this research are free availability, ease of use due to its graphical user interfaces and a comprehensive collection of data pre-processing and modelling techniques [27].

3.2 Forecasting Time Horizon Selection

Choosing the time period for sales forecast is an important step [8]. Here are the three basic time frames for forecasts:

- Short-Term Forecast. Cover up to a year and can include monthly or quarterly forecasts. They help set production levels, sales targets, and overhead costs.

- Medium-Term Forecast. Range from one to four years and guide product development, workforce planning, and real estate needs.

- Long-Term Forecasts. These extend from five to 20 years and inform capital investment, capacity planning, long-range financing programs, succession planning, and workforce skill and training requirements.

Also, in Machine Learning the number of time steps ahead to be forecasted is important [28]. The most common are two names:

- One-Step Forecast. This is where the next time step *(t+1)* is predicted.

- Multi-Step Forecast. This is where two or more future time steps are to be predicted.

Thus, using this terminology we can determine that in this research we have deal with short-term and multi-step forecast. The practice time frame for forecasting in our case is 31 steps ahead (31 days).

3.3 Data Preparation

Data preparation stage consist of the several phases of common methodology used for machine learning project, also called as cross-industry standard process for data mining (CRISP-DM) [29]. The stage combines inside the data understanding and data preparation phases of CRISP-DM process. The data preparation process pipeline is illustrated in Figure 2.

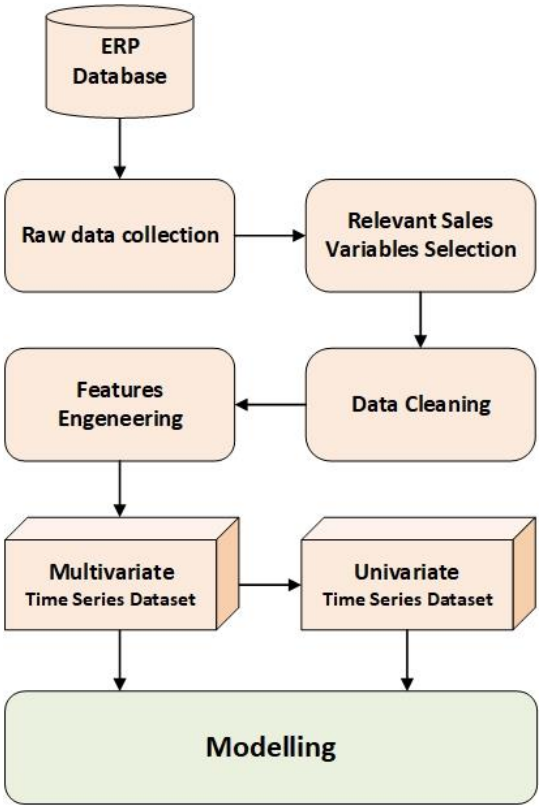

Figure 2. Data preparation stage pipeline

3.3.1 Raw Data Overview

Data for this work was obtained from a wholesale B2B company's ERP database. This refers to closed sales between January 2019 and December 2021. The data is provided on Microsoft Excel Worksheet (*.xlsx*) files for each year separately. The volume of transactions (unique rows) is presented in Table 1.

Table 1. Volume of transactions in raw data

YEAR	VOLUME OF TRANSACTIONS
2019	539 939
2020	587 536
2021	640 456
Total	**1 767 931**

The raw data under consideration has the following structure as shown in Figure 3 and variables description as shown in Table 2.

Date	OrderNr	CustomerID	SalesPerson	AccountManager	Wh	ItemID	Quantity	Net sales	Margin	ELV	Product Segment	ABC value	Customer Segment	Customer Group
2.01.2019	M00233821	11050857	Urmas Niinepuu	ELV Raamleping	JHV	660775040	1	501,1	12,53	Yes	Tele - ja elektrivõrgu va	C	UTILITY	104
2.01.2019	M00261308	10089320	Urmas Niinepuu	ELV Raamleping	MUSTAMAE	660775033	1	501,56	12,99	Yes	Tele - ja elektrivõrgu va	A	UTILITY	104
2.01.2019	M00261308	10089320	Urmas Niinepuu	ELV Raamleping	MUSTAMAE	540002185	1	557,78	142,43	Yes	Tele - ja elektrivõrgu va	B	UTILITY	104
2.01.2019	M00252507	10665798	Urmas Niinepuu	ELV Raamleping	TRT	540002022	1	214,65	-134,21	Yes	Tele - ja elektrivõrgu va	B	UTILITY	104
2.01.2019	M00237420	11445550	Urmas Niinepuu	ELV Raamleping	KURESSAARE	520161176	1	346,84	63,99	No	Tele - ja elektrivõrgu va	B	UTILITY	104
2.01.2019	M00246671	11445550	Urmas Niinepuu	ELV Raamleping	KURESSAARE	540078114	1	319,41	104,54	Yes	Tele - ja elektrivõrgu va	B	UTILITY	104
2.01.2019	M00261943	10665798	Urmas Niinepuu	ELV Raamleping	TRT	660775016	1	212,51	5,51	Yes	Tele - ja elektrivõrgu va	A	UTILITY	104
2.01.2019	M00252507	10665798	Urmas Niinepuu	ELV Raamleping	TRT	660775016	1	212,51	5,51	Yes	Tele - ja elektrivõrgu va	A	UTILITY	104
2.01.2019	M00263887	10722319	Evelin Tähepõld	ELV Raamleping	RAPLA	660775016	3	637,53	16,53	Yes	Tele - ja elektrivõrgu va	A	UTILITY	104

Figure 3. Raw Data structure. Example of first ten rows from 2019

Table 2. Raw Data description

VARIABLE LABEL	DESCRIPTION	TYPE
Date	Transaction Date	date
OrderNr	Number of orders	categorical
CustomerID	Customer unique identification number	categorical
Sales Person	Transaction operator	categorical
Account Manager	Customer Key Account Manager	categorical
Wh	Virtual stock from which the sale was made	categorical
ItemID	Product unique identification number	categorical
Quantity	Quantity of the product	numerical
Net Sales	Sales sum of the product	numerical
Margin	Margin sum of the product	numerical
ELV	Specific Product status	binary
Product segment	Type of the Product	categorical
ABC value	Inventory classification of products	categorical
Customer Segment	Type of the customer	categorical
Customer group	Customers pricing group	categorical

3.3.2 Relevant Sales Variables Selection

Relevant sales variable selection based on data understanding principles of CRISP-DM methodology. During the complex process of business understanding in this process was involved company business analytics and head of sales. The main goal was determinate those variables which have low prediction power or could not be used as prediction variable for some reasons.

3.3.3 Feature Engineering Principles

In a machine learning workflow, we pick not only the model, but also the features. The definition of a feature is a numeric representation of raw data and feature engineering is the process of formulating the most appropriate features given the data, the model, and the task. The right features are relevant to the task at hand and should be easy for the model to ingest. The number of features is also important. If there are not enough informative features, then the model will be unable to perform the ultimate task. If there are too many features, or if most of them are irrelevant, then the model will be more expensive and tricky to train. Good features make the subsequent modelling step easy and the resulting model more capable of completing the desired task. Bad feature may request a much more complicated model to archive the same level of performance [30].

Based on the above principles, the author of the research work is pays great attention to the feature engineering stage in the workflow.

Raw data exploration shows that the most of labelled variables are categorial. But categorical variables also hold a significant predictive power. This information must be encoded into features to help the model leveraging the dataset hierarchy.

The same time during the stage of raw data understanding the part of variables does not have practical sense to use during this research. Because they are consisted dynamic information that could be changed during the time and re-saved to database as new historical data with renamed labelled values.

All categorical variables and some of their combinations have been encoded with One Hot Encoding method [31] with aggregation by time (daily based). Aggregations performed are only summarization of the records by the prediction keys, where the time dimension allows the records to be aggregated on an arbitrarily granularity level which fits best the data depending on the model. Multivariate time series data architecture is illustrated in Figure 4.

As a result of raw data transformation, we have a dataset which could be named as multivariate time series dataset. We can easy transform it to univariate time series dataset by removing of unpredictable variables.

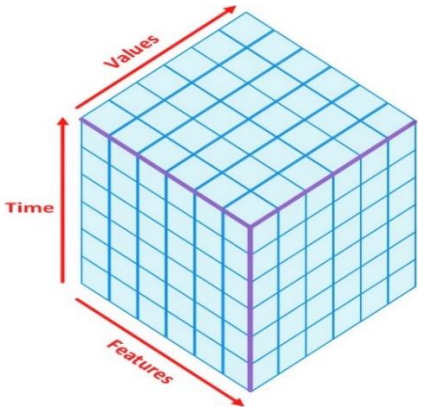

Figure 4. Multivariate time series data architecture

3.4 Modelling

Modelling stage pipeline is illustrated in Figure 5.

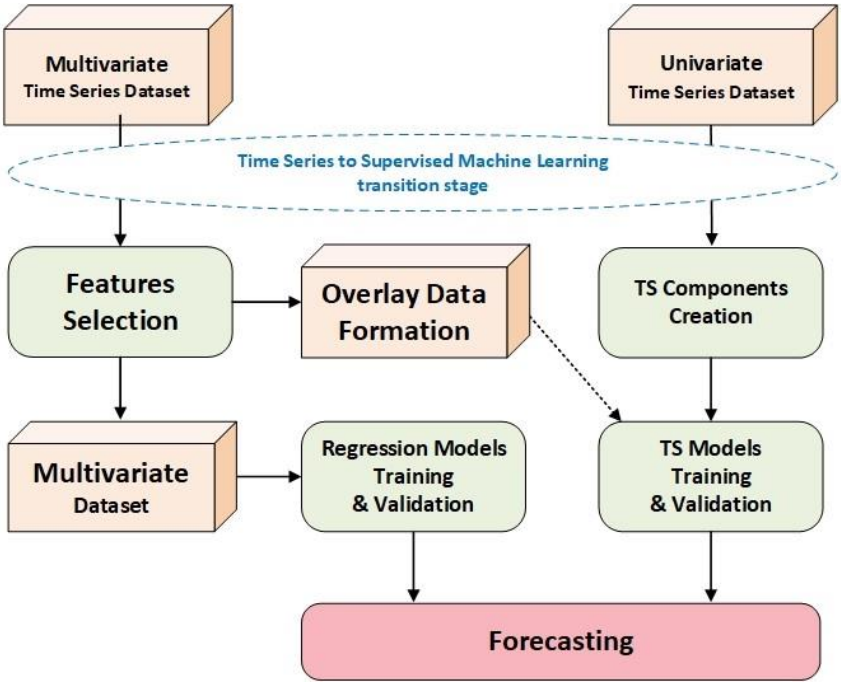

Figure 5. Machine learning modelling stage pipeline

3.4.1 Time Series Components Generation

A time series is a sequence of observations taken sequentially in time. Time series forecasting involves taking models then fit them on historical data then using them to predict future observations. Therefore, days, ago of the measurement is used as an input to predict the next days. The steps that are considered to shift the data backward in the time (sequence), called lag times or lags [32]. Lags are basically the shift of the data one step or more backward in the time (see Figure 6). Therefore, a time series problem can be

transformed into a supervised ML by adding lags of measurements as inputs of the supervised ML (see Figures 7).

Lag 1

Time	Lag-1	y
1	?	1
2	1	0
3	0	1
4	1	0
5	0	0
6	0	?

Lag 2

Time	Lag-2	Lag-1	y
1	?	?	1
2	?	1	0
3	1	0	1
4	0	1	0
5	1	0	0
6	0	0	?

Lag 3

Time	Lag-3	Lag-2	Lag-1	y
1	?	?	?	1
2	?	?	1	0
3	?	1	0	1
4	1	0	1	0
5	0	1	0	0
6	1	0	0	?

Figure 6. The example of lags creation

Supervised $y=F(x)$

x	y
5	1
4	0
5	1
3	1
6	0
3	1

Time Series

Time	Measure
1	1
2	0
3	1
4	0
5	0
6	?

Time Series to Supervised

Time	Lag-1 or x	y
1	?	1
2	1	0
3	0	1
4	1	0
5	0	0
6	0	?

Figure 7. Time Series to Supervised Machine Learning transformation

As an advanced periodic feature Weka software allows to customize which date-derived periodic attributes are created.

As a separate setup of each experimental case will be used an overlay data for the regression time series model. That is, data that is not to be forecasted, but could give significant predictive power to the model. The overlay data in our case are the same features that will be correlated to the target variable.

3.4.2 Feature Selection Method

Feature selection process of this research is based on standard ML feature selection method as **Correlation**. This method implemented to dataset for defining the feature subsets that should be considered when making prediction. The main goal of this process is feature count reduction. It allows to use only useful features that influence to the target

variable and avoid using such that are inappropriate to take into account. And avoid the situation then the final model has a too complicated structure for implementation in practice.

Correlation refers to the so-called Filter methods [33]. The Filter methods are faster and less computationally expensive than so-called Wrapper methods. When dealing with high-dimensional data, it is computationally cheaper to use filter methods.

Correlation is a measure of the linear relationship of two or more variables. Through correlation, we can predict one variable from the other. The logic behind using correlation for feature selection is that the good variables are highly correlated with the target. Furthermore, variables should be correlated with the target but should be uncorrelated among themselves. If two variables are correlated, we can predict one from the other. Therefore, if two features are correlated, the model only really needs one of them, as the second one does not add additional information. We will use the Pearson Correlation [34]. As reference value setting used absolute value equalled **0.9**, as the threshold for selecting the variables.

3.4.3 Final Dataset Splitting Principles

The final dataset is divided into three parts. To predict the margin, for the next 31 days using historical data from the last three years, the instances of from last 31 days (December 2021) were set aside as overlay data. The remaining 1065 days were used to train and validate the models on an 80/20 basis.

3.4.4 Machine Learning Algorithms Selection

The choice of ML algorithms for this research is determined by two criteria:

- Research results reviews based on scientific articles dedicated to solving similar tasks or problems.
- The possibility of implementing the selected algorithm using the selected software

By those criteria for experimental part of the work was chosen next algorithms:

- **Linear regression**

Linear regression is an approach for modelling the relationship between a continuous dependent variable y and one or more predictors X [13]. The relationship between y and X can be linearly modelled as $y=\beta TX+\epsilon y=\beta TX+\epsilon$ Given the training examples $\{xi,yi\}Ni=1\{xi,yi\}i=1N$, the parameter vector β can be learnt.

Linear regression model (LR) that captures the linear relationship between two variables (simple linear regression) or more variables (multiple linear regression), one labelled as the dependent variable and the others labelled as the independent variables. In general, multiple linear regression procedures will estimate the value of the variable based on the following equation, then α represents the coefficients.

On this research we need to talk only about multiple linear regression because we have a multivariate dataset as the core of forecasting process.

The research article published by K. Kausthub in 2021 [35], confirm good overall accuracy and low RMSE value as a result of the algorithm implementation for sales prediction problem solving.

- **Support Vector regression**

Support Vector regression (SVR) was proposed in 1996 [36] is a machine learning model that uses the Support Vector Machine, a classification algorithm, to predict a continuous variable.

While linear regression models minimize the error between the actual and predicted values through the line of best fit, Support Vector regression manages to fit the best line within a threshold of values, otherwise called the epsilon-insensitive tube. This is a tube containing the margin of error we allow our model to have, where any points of error inside the tube are of least important. The points outside the tube, however, are accounted for and the distance between the point and the tube itself are measured and labelled as the support vector.

The review of research work published by several authors in 2019 [37], where was introduced comparison that covered 95 time series datasets. Authors indicates that SVM is the most promising methods for temporal data modelling and forecasting.

- **Random Forest regression**

Random Forest [38] is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, include the following: there is no interpretability and overfitting may easily occur by selection of the number of trees to include in the model.

A Random Forest regression model (RF) is powerful and accurate. As we can see in research work of C. Shyamala and colleagues [21] the RF was confirmed to be an effective model in forecasting sales data.

3.5 Forecasting Process Description

The forecasting stage will be divided into three parallel pipelines as illustrated in Figure 8. We predict a real values of target variable by using one regression model. As a parallel process, we execute a univariate time series forecasting divided in turn into two different setups. One with use of time series components only and second one with overlay data added to the first. As a result of forecasting, we have an evaluation metrics and predicted values for each experimental case.

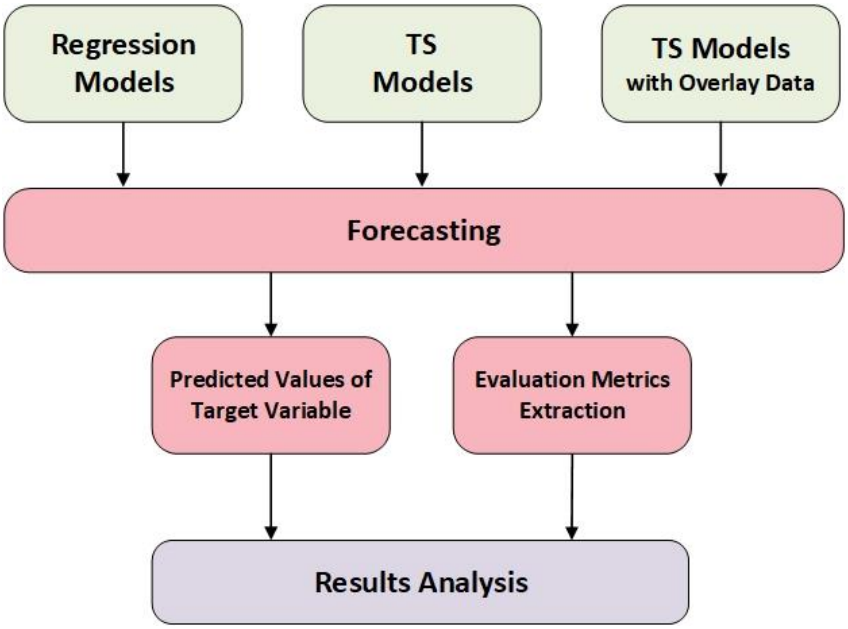

Figure 8. Forecasting stage pipeline

3.6 Evaluation Metrics Selection

Evaluation metrics are used to measure the quality of statistical or machine learning model. To evaluate the performance of ML models, the following most common metric is Root Mean Square Error (RMSE) [39]. The RMSE is a frequently used measure of the differences between values predicted by model, or an estimator and the values observed. Formally it is defined as follows:

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

Where, n = total number of data points, y_i = actual value, $\hat{y}_i$ = predicted value.

The second selection metric is Mean Absolute Error (MAE) [39] is a measure of errors between paired observations expressing the same phenomenon. MAE is calculated as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}i|$$

Where, N = total number of data points, y_i = actual value, $\hat{y}_i$ = predicted value.

MAE and RMSE can be used together to diagnose the variation in the errors in a set of forecasts. RMSE will always be larger or equal to the MAE. The greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE=MAE, then all the errors are of the same magnitude.

For practical reason, we decided to use an additional measure of the accuracy of the predicted target variable values in relation to actual values that we already know.

The additional performance metric used in this work is the accuracy of predicted value to actual value by each day from 31 step-ahead forecast. Let's consider name this metric as Predicted-to-Actual accuracy. This metric will measure the prediction error in each day.

We compare accuracy of each day by using corrected sample standard deviation [40], Its calculated as:

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

where $\{x_1, x_2 \ldots x_n\}$ are the observed values of the sample items, and $\bar{x}$ is the mean value of these observations, while the denominator *N* stands for the size of the sample.

We implement this metric using of STDEV function in Excel. Formally this function is describe now "far" predicted value from actual (where 0 value is 100% accuracy).

4 Experimentation

4.1 Data Preparation

4.1.1 Relevant Variables Selection

Before starts any transformations with data we need to know how useful is the already mined data. For this analysis was engaged specialists from company to identify each variable background. During analysis are prepared decision about several variables. This is a baseline for the elimination of those variables from dataset on next stages of research:

- The ***Wh*** variable describes the name of the virtual store from that sale was made. This virtual store is not tied geographically and may not have predictive force.

- The ***ABC_value*** variable describes an inventory categorisation. This categorization relates to product and may be changed during some period (usually quarterly or yearly). The company database is stored only real value of this variable but not historical. That means, for example, that one product could be during 2019 in one category "*A*", but today the same product already in category "*C*". The database is stored only "*C*" value and we don't know real value of the product in 2019 statistics.

- The ***Customer_group*** variable describes a customer pricing level. And this value is also having the same variative problem like *ABC_value* before.

Based on the above arguments, was decided to eliminate ***Wh, ABC_value***, ***Customer_group*** variables from the next stages of observed dataset.

4.1.2 Raw Data Cleaning

This step typically imputes the missing values and handles the outliers present in the dataset. Missing values are found in ***Sales_Person, Account_Manager, ELV, Product_segment*** and ***Customer_segment*** variables of the dataset. All those variables are categorical and here we can't find the average or assign any known category manually without any mistake. But we consider that they are still may have prediction power and

elimination before ML analysis may have a negative impact on forecasting accuracy. We assign to all of them category value named ***Unknown***.

4.1.3 Feature Engineering

Based on the research objectives and feature engineering principles we need to generate new features, there data presented aggregated by each day and all feature values have only numeric type of data. The generated dataset with new features has a structure introduced in Table 3.

Table 3. Features structure in generated dataset

	FEATURES							
		TARGET VARIABLE	**GROUP 1**			**GROUP 2**		
	TIME STAMP	**TOTAL**	**TOTAL**	**LABELLED VARIABLE …1**	**LABELLED VARIABLE …N**	**TOTAL**	**LABELLED VARIABLE …1**	**LABELLED VARIABLE …N**
INSTANCES	date...1	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value
	date...2	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value
	date...3	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value
	date...4	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value
	date...5	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value
	Date...n	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value	numeric value

Consider and describe transformation of each raw data variable separately:

- ***Date***. It has a fundamental importance in forecasting process. We use this variable as a feature aggregated by each calendar day. This feature is defining a time series and will be used as timestamp during forecasting.

- ***Margin***. This variable as a feature aggregated as sum of values (numeric) by each day. It is a target variable in research.

- ***Sales*** and ***Quantity.*** Those variables we decide to use as values (numeric data). We transform they as total feature and distributed by each labelled feature. We will count it as sum of values by each day.

- ***OrderNr, CustomerID*** and ***ItemID***. Those variables we decide to use as values (numeric data). We transform they as total feature and distributed by each labelled feature. We will count it as sum of unique values by each day.

- ***ELV.*** We will use it as feature with 3 possible labels (TRUE, FALSE, UNKNOWN) aggregated by SALES, QUANTITY, ORDERS, CUSTOMERS, and ITEMS by each day.

- ***Product_segment.*** We will use it as feature with 5 possible labels (CONSTRUCTION, UTILITY, INDUSTRY, LIGHTING, UNKNOWN) aggregated by SALES, QUANTITY, ORDERS, CUSTOMERS, and ITEMS by each day.

- ***Customer_segment***. We will use it as feature with 4 possible labels (CONSTRUCTION, UTILITY, INDUSTRY, UNKNOWN) aggregated by SALES, QUANTITY, ORDERS, CUSTOMERS, and ITEMS by each day.

- ***Sales_Person.*** This variable is combined into 4 possible labels (ELV, MANUAL, WEB, UNKNOWN) and represents the sales operator. Where ELV – old company web sales platform, MANUAL – all sales where operator was a man. WEB – new company web sales platform. UNKNOWN – unknown operator. Each label also aggregated by SALES, QUANTITY, ORDERS, CUSTOMERS, ITEMS by each day.

- ***Account_manager.*** This variable is combined into 4 possible labels (ELV_K, ELV_R, CS, KAM) and represents the status of the kay account management. Where ELV_K and ELV_R are special deal cases which does not have key account managers. CS – customer service sale, the customer does not have key account manager. KAM – The customer has a key account manager. Each label

also aggregated by SALES, QUANTITY, ORDERS, CUSTOMERS, ITEMS by each day.

4.1.4 Final Dataset Overview

After raw data transformations and feature generation process, we have a dataset with 107 features (1 timestamp, 1 target variable and 5 groups of features) and 1096 instances (dates from 1.01.2019 till 31.12.2021). Each of 5 feature groups has a similar structure consist of one total feature and 21 distributed features. In Appendix 2 we can see a list of features before selection methods implementation.

4.2 Feature Selection

As a result of correlation filter implementation to target variable we got the following coefficients as presented in Appendix 2. The whole correlation matrix is also presented in Figure 9. Based on the previously defined reference setting of absolute value equalled 0.9 as the threshold, we eliminate all features which coefficient values are below from next observations. The final dataset contains features presented in Table 4. We sort them from highest to lowest correlation coefficient. Thus, in the process of building regression models and as possible time series overlay data will used only 41 features.

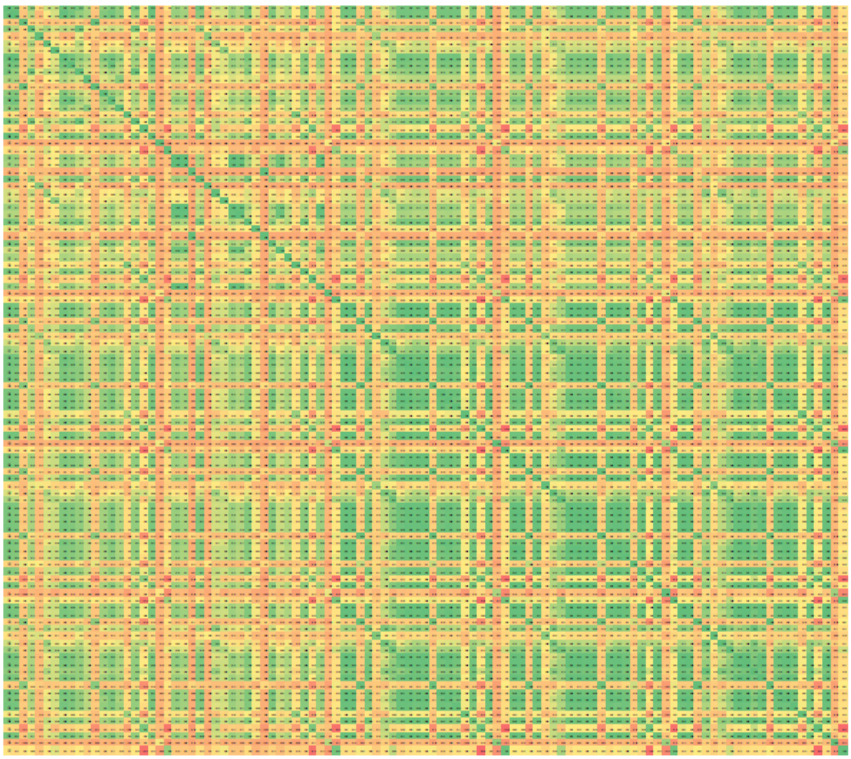

Figure 9. Features correlation matrix

Table 4. Features list with reference correlation coefficient and above

FEATURE	CORRELATION COEFFICIENT
SALES_SUM_TOTAL	0,99
SALES_SUM_OPERATION_BY_MANUAL	0,98
SALES_SUM_ELV_PRODUCT_FALSE	0,96
SALES_SUM_KAM	0,95
ORDERS_COUNT_UNIQUE_TOTAL	0,94
ORDERS_COUNT_UNIQUE_ELV_PRODUCT_FALSE	0,94
SALES_SUM_PROD_SEGM_UTILITY	0,93
SALES_SUM_CUS_SEGM_UTILITY	0,93
ORDERS_COUNT_UNIQUE_PROD_SEGM_CONSTRUCTION	0,93
ORDERS_COUNT_UNIQUE_PROD_SEGM_INDUSTRY	0,93
ORDERS_COUNT_UNIQUE_OPERATION_BY_MANUAL	0,93
QUANTITY_SUM_ELV_PRODUCT_TRUE	0,92
ORDERS_COUNT_UNIQUE_KAM	0,92
ORDERS_COUNT_UNIQUE_PROD_SEGM_UTILITY	0,92
ORDERS_COUNT_UNIQUE_CUS_SEGM_INDUSTRY	0,92
CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_INDUSTRY	0,92
ITEMS_COUNT_UNIQUE_ELV_PRODUCT_FALSE	0,92
ITEMS_COUNT_UNIQUE_KAM	0,92
ITEMS_COUNT_UNIQUE_PROD_SEGM_LIGHTING	0,92
ITEMS_COUNT_UNIQUE_CUS_SEGM_INDUSTRY	0,92
ITEMS_COUNT_UNIQUE_OPERATION_BY_MANUAL	0,92
QUANTITY_SUM_PROD_SEGM_UTILITY	0,91
ORDERS_COUNT_UNIQUE_PROD_SEGM_LIGHTING	0,91
ORDERS_COUNT_UNIQUE_CUS_SEGM_CONSTRUCTION	0,91
CUSTOMERS_COUNT_UNIQUE_TOTAL	0,91
CUSTOMERS_COUNT_UNIQUE_ELV_PRODUCT_FALSE	0,91
CUSTOMERS_COUNT_UNIQUE_KAM	0,91

CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_CONSTRUCTION	0,91
CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_UTILITY	0,91
CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_LIGHTING	0,91
CUSTOMERS_COUNT_UNIQUE_OPERATION_BY_MANUAL	0,91
ITEMS_COUNT_UNIQUE_TOTAL	0,91
ITEMS_COUNT_UNIQUE_PROD_SEGM_CONSTRUCTION	0,91
ITEMS_COUNT_UNIQUE_PROD_SEGM_INDUSTRY	0,91
ITEMS_COUNT_UNIQUE_CUS_SEGM_CONSTRUCTION	0,91
SALES_SUM_PROD_SEGM_CONSTRUCTION	0,90
QUANTITY_SUM_CUS_SEGM_UTILITY	0,90
ORDERS_COUNT_UNIQUE_CUS_SEGM_UTILITY	0,90
CUSTOMERS_COUNT_UNIQUE_ELV_PRODUCT_TRUE	0,90
CUSTOMERS_COUNT_UNIQUE_CUS_SEGM_CONSTRUCTION	0,90
CUSTOMERS_COUNT_UNIQUE_CUS_SEGM_INDUSTRY	0,90

During the several ML model building experiments we also will be used additional feature selection algorithm known as Greedy algorithm [41]. A Greedy feature selection is the one in which an algorithm will either select the best features one by one (forward selection) or removes worst feature one by one (backward selection).

4.3 Patterns in Dataset

Based on correlation filter implementation we can make the following assumptions about the relationships of variables and describe the patterns in observed data:

- The SALES group of features.

  The target variable in our research is derivative of the amount of sales. We know what margin is a several percent of sales amount. From here we observe that several features from the group of SALES have a good correlation with MARGIN. We see that the major prediction power has several factors that will described by features of customer group and product groups. It is sales amount to utility customers and sales amount by products from utility and construction groups. In

same time the sales amount of products which hasn't ELV status also was correlated as feature-predictor. Total sales amount, manual operation and sales amount produced to customers who have a key account manager are top features by prediction power in our research.

- The QUANTITY group of features.

  In contrast to the previous group, only a few features, that characterizing the volume of sales, have are significant predictive power. It is a volume of sold quantities from product group of construction and to customers from customer group of utility and the products with positive ELV status.

- The ORDERS group of features.

  Looking at this group of features we can see that 11 of 21 features has good correlation with target variable. Basically, they are characterizing product group (4 of 5 features), customer group (3 of 4 features), product positive status by ELV and operation and key account manager characteristic. Also orders total count have a prediction power too.

- The CUSTOMERS group of features.

  The customers group of features have almost the same structure of patterns like orders group. With exception of elimination the feature that describe customer from utility group and with adding the feature described non positive ELV product status. Thus, 11 of 21 group features also has a good predictive power to target variable.

- The ITEMS group of features.

  In this group we can identify only 9 of 21 preliminary generated feature. As we can see it is features related with product group (3 of 5), customer group (2 of 4), nonpositive ELV status of product, and operation, total and key account manager characteristic

Based on the above we identify next patterns in the observed data:

- Total values as generated features (excl. Quantity) play an important role in target variable formation
- Features that characterize the manual operations with sales (excl. Quantity) play an important role in target variable formation
- Features that characterize the account management status for customers (excl. Quantity) play an important role in target variable formation
- Features that characterize the nonpositive ELV status of the product (excl. Quantity) play an important role in target variable formation
- Features that characterize products and customers groups could be important. But they importance should be analysed in every case.

Here we can also note, in our opinion, negatively influencing factors characterizing the availability of data and observations. Unfortunately, available data does not include any geographical distribution of sales. Thus, we cannot determine any patterns in historical observations. We can only assume from the sales business essence that this distribution could help to define new patterns and bring to the research a new prediction power.

4.4 Regression Models

Based on selected ML algorithms and selected features, we start to build and evaluate regression models. On next subsections are introduced with description best models for each algorithm. For training the models we used 852 instances (80%) and for evaluation 213 (20%).

4.4.1 Linear Regression Models

The Linear Regression models shown next evaluation metrics as presented in Table 5. We have a two very similar result by evaluation metrics but big difference in model size. First model created without additional feature selection setup (call it next as **LR**) and second with additional Greedy selection (call it next as **LR+G**). Which allowed to reduce number of used features to the model building from 41 to 23. We also can notice that both

model setups have an R^2 value and adjusted R^2 value near to 0,94. It indicates that at least 94% of the variation in the output variables are explained by the input variables.

Table 5. Linear regression models metrics on the test sample

Model setup	Evaluation Metrics	
	RMSE	MAE
LR	5513	2836
LR+G	5554	2818

4.4.2 Support Vector Regression Models

Support Vector regression model have two setups of features scaling. In the first case data are normalized (the features are on a relatively similar scale), call it next as **SVR_N**. And in second standardized (the features are close to normally distributed), call it next as **SVR_S**. As a configuration of SVR model was chosen the polynomial kernel with exponent 1. The evaluation metrics of both cases are very similar and presented in Table 6.

Table 6. Support Vector Regression models metrics on the test sample

Model setup	Evaluation Metrics	
	RMSE	MAE
SVR_N	5226	2340
SVR_S	5155	2325

4.4.3 Random Forest Regression Model

The Random Forest regression model have only one setup (call it next as **RF**) and evaluation metrics are presented in Table 7.

Table 7. Random Forest Regression model metrics on the test sample

Model setup	Evaluation Metrics	
	RMSE	MAE
RF	5420	2375

4.5 Time Series Regression Models

Based on selected ML algorithms and univariate time series dataset we start to build and evaluate time series regression models with different setups. On next subsections are introduced with description best models for each algorithm. For training the models we used 80% of 1065 instances and for validation remaining 20%. The splitting of dataset implemented a strictly on time scale.

4.5.1 Linear Regression Time Series Models

The Linear Regression time series models shown next evaluation metrics as presented in Table 8. We have a two different setup of model based on additional Greedy features selection. The first setup contains only target variable lags 1-7 and additional periodic attributes of weekend and day of week (call it next as **LR TS**). The second is additionally to the first are overlay data for forecasted period 1.12.2021 - 31.12.2021 (call it next as **LR TS+O**).

Table 8. Linear regression time series models metrics on the test sample

Model setup	**Evaluation Metrics**	
	RMSE	**MAE**
LR TS	19465	12737
LR TS+O	7480	4938

4.5.2 Support Vector Regression Time Series Models

The Support Vector regression time series models shown next evaluation metrics as presented in Table 9. We have a four different setup of model which have:

- Normalized features with lags and periodicity (call it next as **SVR_N TS**)
- Standardized features with lags and periodicity (call it next as **SVR_S TS**)
- Normalized features with lags (1-7), periodicity (weekend and day of week) and overlay data (call it next as **SVR_N TS+O**)
- Standardized features with lags (1-7), periodicity (weekend and day of week) and overlay data (call it next as **SVR_S TS+O**)

As a configuration of SVR model was chosen the polynomial kernel with exponent 1.

Table 9. Support Vector regression time series models metrics on the test sample

Model setup	Evaluation Metrics	
	RMSE	MAE
SVR_N TS	19741	9579
SVR_S TS	19719	9461
SVR_N TS+O	6607	3752
SVR_S TS+O	6542	3469

4.5.3 Random Forest Regression Time Series Models

The Random Forest regression time series models shown next evaluation metrics as presented in Table 10. We have a two different setup of model based on our forecasting methodology. The first setup contains only target variable lags 1-7 and additional periodic attributes of weekend and day of week (call it next as **RF TS**). The second is additionally to the first are overlay data for forecasted period 1.12.2021 - 31.12.2021 (call it next as **RF TS+O**).

Table 10. Random Forest regression time series models metrics on the test sample.

Model setup	Evaluation Metrics	
	RMSE	MAE
RF TS	19805	10645
RF TS+O	8238	4593

4.6 Target Variable Forecasting

As a result of forecasting we need to have a numeric values of target variable for each period (day) ahead. For this we implement each model equation and calculate whose numbers.

We used as a reference for regression model without time series component only Linear Regression. The evaluation metrics of SVR and RF regression models are slightly better but model equation is very difficult to implementation on practice.

For regression time series models, we used best setup for each algorithm and separately adding the overlay data to them. Forecasting stage results are presented in Table 11.

Table 11. Forecasted values of target variable by algorithms and setup

DATE	ACTUAL VALUES	FORECASTED VALUES BY ALGORITHMS AND SETUP						
		LR+G	LR TS	LR TS+O	SVR_S TS	SVR_S TS+O	RF TS	RF TS+O
2021-12-01	24891	20416	16002	16708	16900	22281	46365	21154
2021-12-02	17262	16416	21143	19107	17687	16688	30650	16532
2021-12-03	54927	48730	73091	47382	71897	50300	104140	48373
2021-12-04	0	-54	3363	-1371	303	-210	5553	75
2021-12-05	0	-54	-5945	-3469	-429	-1190	4807	73
2021-12-06	18986	17325	26510	21352	17466	18232	51342	18670
2021-12-07	16696	16343	47447	18986	38786	22620	64002	17539
2021-12-08	17755	17905	21817	19419	16897	20867	46500	17697
2021-12-09	8802	11890	25375	12116	17919	9580	62350	12312
2021-12-10	78777	65313	75956	68796	70073	63464	70624	72787
2021-12-11	0	-54	3186	-16	397	-121	17714	75
2021-12-12	0	-54	130	1018	43	-222	14962	73
2021-12-13	15748	14335	25620	17025	16633	18157	44777	15350
2021-12-14	18268	13808	33571	14688	22613	16187	70349	17303
2021-12-15	22673	20937	23624	21026	17059	22413	46508	22269
2021-12-16	30353	29899	27572	33130	18163	27990	53429	23291
2021-12-17	26118	25607	76999	25349	69992	23793	71592	20032
2021-12-18	0	-54	3077	2319	414	-99	19284	75
2021-12-19	0	-54	3140	635	276	-278	16905	73
2021-12-20	42261	44499	25182	44400	16711	41540	44674	29540
2021-12-21	113661	87133	30420	87783	18915	87882	71513	98111
2021-12-22	53680	53067	24182	53294	17168	47799	46295	27106
2021-12-23	17778	20617	28601	19713	18305	17317	53185	20242
2021-12-24	17910	7945	77406	8725	70210	11794	72830	11472
2021-12-25	0	-54	3159	-953	419	-503	19284	75
2021-12-26	0	-54	4291	-3357	362	-1319	17549	73
2021-12-27	36296	31341	25174	52860	16833	60612	45272	19638
2021-12-28	11946	12053	29826	16018	18108	18522	70445	12165
2021-12-29	19135	16512	24465	19050	17248	18652	46424	23477
2021-12-30	10460	20881	29156	23828	18402	19805	53284	15700
2021-12-31	9551	9848	77773	11989	70475	13062	72830	18402

5 Results Analysis

5.1 Evaluation Metrics Comparation

5.1.1 RMSE and MAE

All basic evaluation metrics related with research are combined in Table 12 and presented graphically on Figure 10. As we can see from the metrics values clearly distinguish three subgroups of results. Regression models have RMSE and MAE lower than same time series regression models. The time series regression models with overlay data have a slightly higher RMSE and MAE compared to regression models, but significantly better performance, almost in three times, than time series regression models based only on time series components.

Table 12. Evaluation metrics by ML models and setup

METHOD	MODEL SETUP	RMSE	MAE
Regression	LR	5513	2836
	LR+G	5554	2818
	SVR_N	5226	2340
	SVR_S	5155	2325
	RF	5420	2375
Time Series regression	LR TS	19465	12737
	SVR_N TS	19741	9579
	SVR_S TS	19719	9461
	RF TS	19805	10645
Time Series regression with Overlay data	LR TS+O	7480	4938
	SVR_N TS+O	6607	3752
	SVR_S TS+O	6542	3469
	RF TS+O	8238	4593

The comparation of ML algorithms shows that Support Vector regressor applicable to regression problem solving has a better performance than Linear regressor or Random

Forest regressor. We see exactly same correlation in every subgroup of metrics. SVR is also slightly better for TS and much better for TS+O approach.

The best performance by evaluation metrics shown Support Vector regression model with standardized features. This model has lowest RMSE = **5155** and lowest MAE = **2325** on the test data.

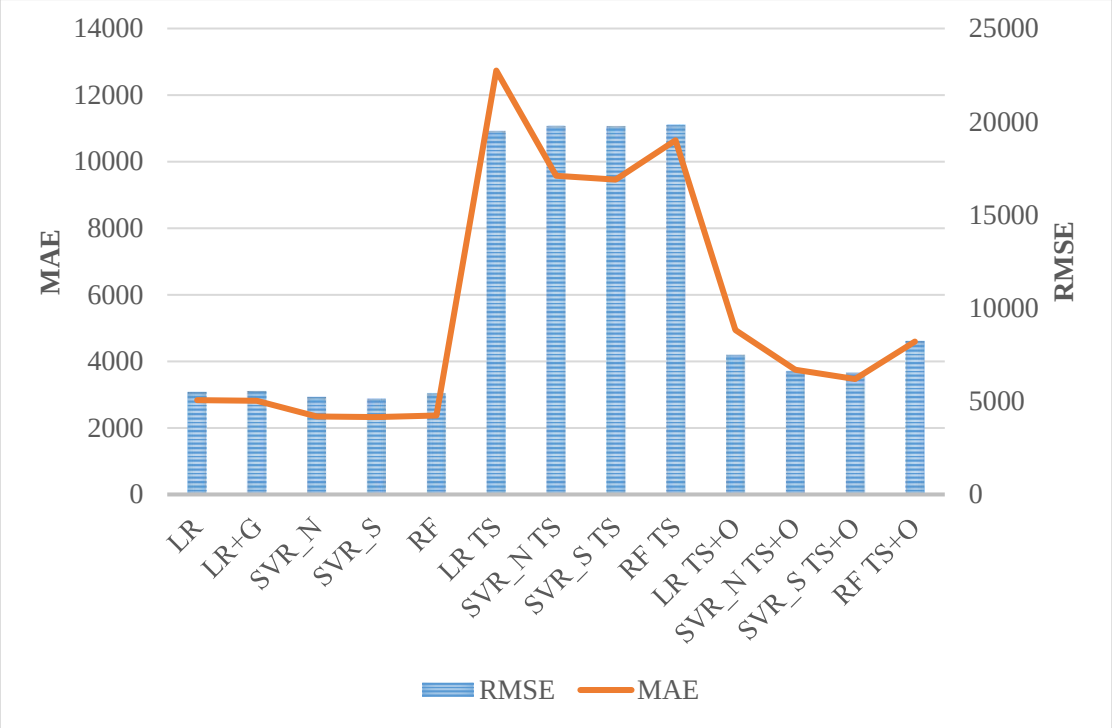

Figure 10. Evaluation metrics by ML models and setup

5.1.2 Predicted-to-Actual Accuracy

A comparison of the actual and predicted values for the test data are presented in Figures 11-13. We combine the results in three different graphs for comparation the methods of forecasting.

In Figure 11 we can see a comparison of actual values to predicted values by Linear Regression model with additional Greedy selection method. As we can see that curve of the predicted values very good describe the actual values curve. This means that the prediction has good performance.

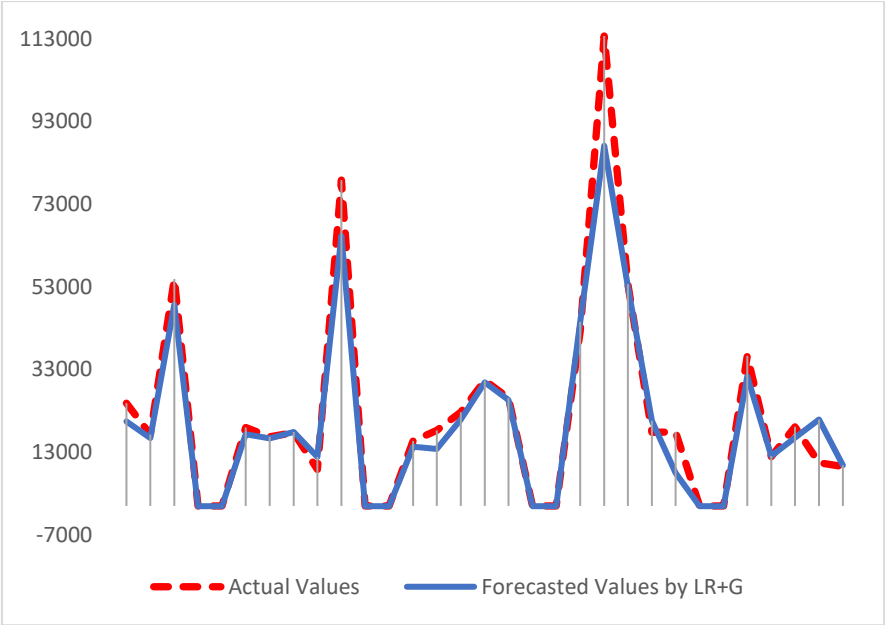

Figure 11. Predicted-to-Actual accuracy curve for LR model

In Figure 12 we can see a comparison of actual values to predicted values by TS models with different ML algorithms. As we can see curve of the predicted values from most points has a big difference with actual values curve. This means that the prediction has poor performance. The reason of that is not enough only target values and its time series components to predict this process. The TS models take in account only temporal characteristics and do not consider other complex aspects that affect the forecast.

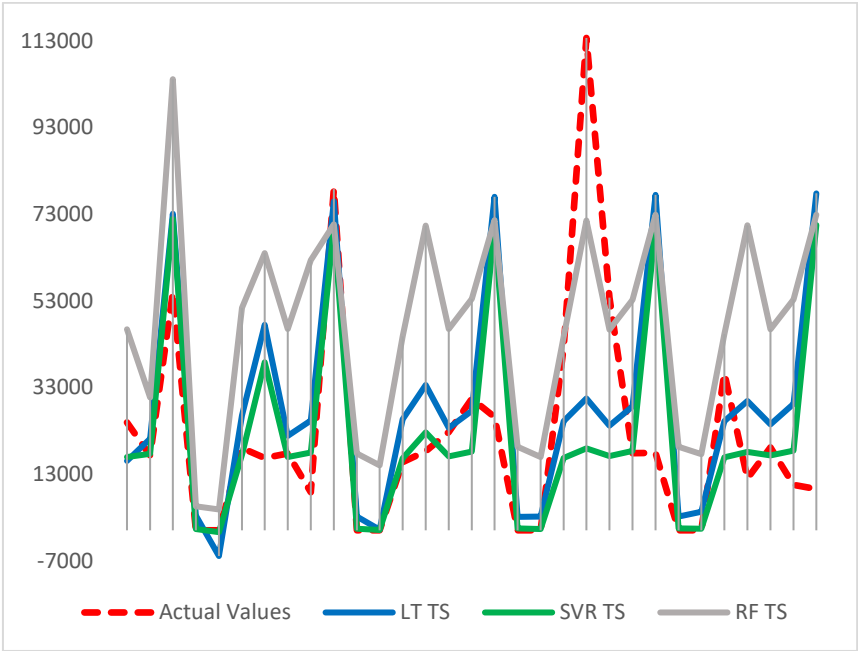

Figure 12. Predicted-to-Actual accuracy curves for TS models

In Figure 13 we can see a comparison of actual values to predicted values by TS+O models. As we can see that curve of the predicted values from most points describes the curve of the actual values well. This means that the prediction has good performance.

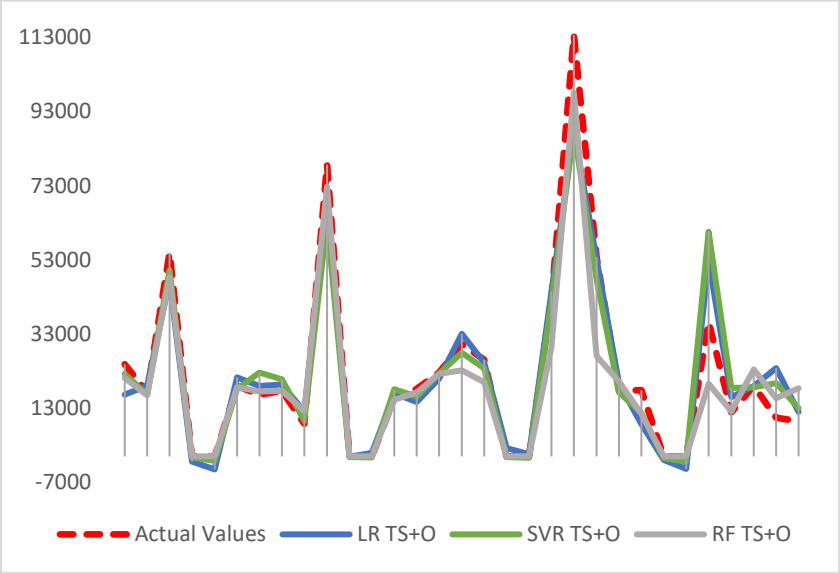

Figure 13. Predicted-to-Actual accuracy curves for TS+O models

The errors on the forecasting we can estimate to each step of forecast period. In our case is 31 days. The standard deviation values of each day are presented in Figures 14-15 as graphs. The zero level is absolute accuracy. We estimate separately regression time series approach to regression model (see Figure 14) and regression time series with overlay data approach to regression model (see Figure 15). Both figures have identical axis values inside group but different in separately. We take in account the maximum of predicted values in each group for the best graphical representation and comparation.

The analysis of Figure 14 confirms our earlier statements that regression model without TS components (in our case LR model as a reference) has significantly better accuracy than regression time series model created by the same ML algorithm.

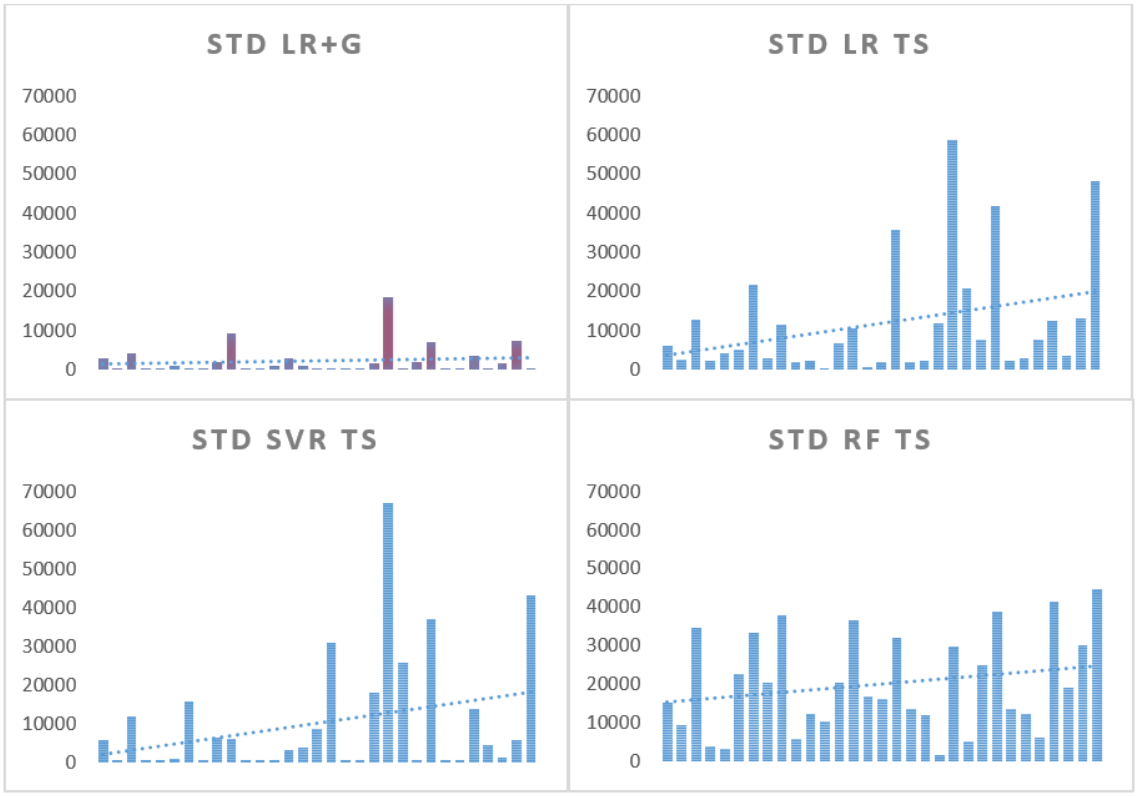

Figure 14. Regression and TS models comparation by STD

Our main observations are:

- The STD of each predicted value by RF TS model does not have top accuracy. Errors distribution is chaotic. There are no obvious bursts of inaccuracies. The average line of error in predicted values is growing by time smoothly.

- The STD of each day value predicted by LR TS model has several top points of accuracy and several near to that. but these points represented weekends when the numerical value of the predicted variable should be close to 0. We also observe high peaks in errors at several time points. Which may be caused by the uncertainty of the sales process flow. The average line of error in predicted values rapidly growing.

- The STD of each day value predicted by SVR TS model has more than ten points near to top of accuracy. That means the good accuracy in particular days. But the same time we see a lot of high peaks of predicted values. Which indicate to a clear lack of prediction power in those days, due to lack of additional observations. The average line of error in predicted values rapidly growing.

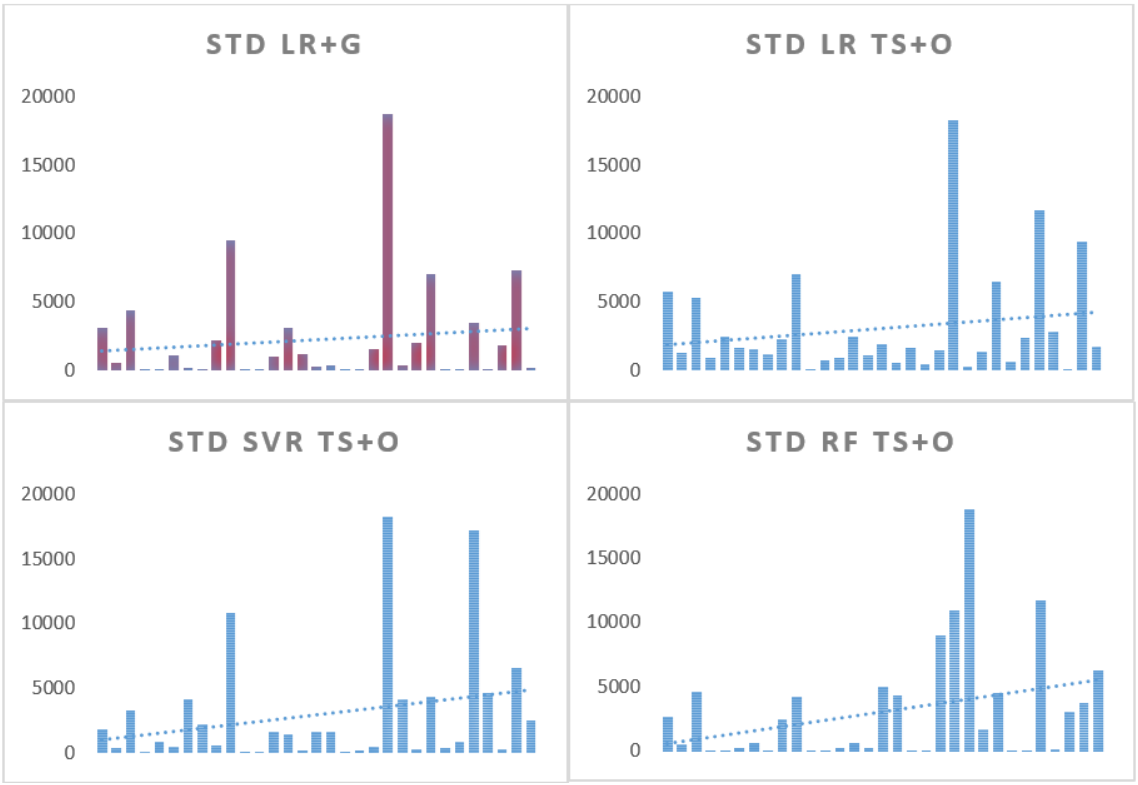

Figure 15. Regression and TS+O models comparation by STD

The analysis of Figure 15 confirms our earlier statements that regression model (in our case LR model as a reference) has similar accuracy as TS+O models. Our main observations are:

- The STD of each predicted value by RF TS+O model has good accuracy. Errors distribution is periodical. The error peaks already not so high like in cases with TS models. The reason of peaks in particular case may be that we did not indicate these days as holidays. The average line of error in predicted values is growing by time quite smoothly in comparison with the error value.

- The STD of each day value predicted by LR TS+O model have several points near to top of accuracy. We observe several high peaks in errors at time points. The reason of peaks in particular case may be that we did not indicate these days as holidays. The average line of error in predicted values is growing by time quite smoothly in comparison with the error value.

- The STD of each day value predicted by SVR TS+O model have more than ten top points of accuracy. That means the very good accuracy in particular days. The same time we see a several peaks of predicted values. Which indicate to a lack of

prediction power in those days, due to lack of additional observations or holidays indication. The average line of error in predicted values is growing by time quite smoothly in comparison with the error value.

5.2 Best Models Insight

In the previous section, we identified the models with the best predictive accuracy to target variable. Further we will consider in more details the equations of the selected models.

Regression models created by Linear regression and Support Vector regression algorithms has the lowest RMSE and MAE values in our research. But they equations have a significant difference. Due to the fact that in LR model (see Table 13) was applied additional feature selection algorithm, we have a much shorter and less complicated equation than in the case of using a SVR model (see Table 14) with data standardization.

Table 13. LR model equation with additional Greedy feature selection method

		MARGIN_SUM_TOTAL	=
-0,0853	*	*SALES_SUM_TOTAL*	+
0,1186	*	*SALES_SUM_ELV_PRODUCT_FALSE*	+
0,0568	*	*SALES_SUM_PROD_SEGM_CONSTRUCTION*	+
0,0644	*	*SALES_SUM_PROD_SEGM_UTILITY*	+
0,0379	*	*SALES_SUM_CUS_SEGM_UTILITY*	+
0,0578	*	*SALES_SUM_OPERATION_BY_MANUAL*	+
0,4057	*	*QUANTITY_SUM_ELV_PRODUCT_TRUE*	+
-0,0778	*	*QUANTITY_SUM_PROD_SEGM_UTILITY*	+
-0,1218	*	*QUANTITY_SUM_CUS_SEGM_UTILITY*	+
109,1553	*	*ORDERS_COUNT_UNIQUE_TOTAL*	+
-59,2408	*	*ORDERS_COUNT_UNIQUE_ELV_PRODUCT_FALSE*	+
-78,1563	*	*ORDERS_COUNT_UNIQUE_PROD_SEGM_UTILITY*	+
-47,0230	*	*ORDERS_COUNT_UNIQUE_CUS_SEGM_CONSTRUCTION*	+
306,7184	*	*CUSTOMERS_COUNT_UNIQUE_TOTAL*	+
166,9494	*	*CUSTOMERS_COUNT_UNIQUE_ELV_PRODUCT_TRUE*	+
-129,6570	*	*CUSTOMERS_COUNT_UNIQUE_KAM*	+
139,0464	*	*CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_LIGHTING*	+
-106,3956	*	*CUSTOMERS_COUNT_UNIQUE_CUS_SEGM_INDUSTRY*	+
-231,3501	*	*CUSTOMERS_COUNT_UNIQUE_OPERATION_BY_MANUAL*	+
39,5897	*	*ITEMS_COUNT_UNIQUE_ELV_PRODUCT_FALSE*	+
17,5334	*	*ITEMS_COUNT_UNIQUE_KAM*	+
-31,2250	*	*ITEMS_COUNT_UNIQUE_PROD_SEGM_CONSTRUCTION*	+
-95,1587	*	*ITEMS_COUNT_UNIQUE_PROD_SEGM_LIGHTING*	+
-17,5043	*	*ITEMS_COUNT_UNIQUE_CUS_SEGM_CONSTRUCTION*	+
-14,6884	*	*ITEMS_COUNT_UNIQUE_OPERATION_BY_MANUAL*	+
-53,8528			

Table 14. SVR model equation with data standardization

			MARGIN_SUM_TOTAL	=
0,3838	*	*(standardized)*	*SALES_SUM_TOTAL*	+
0,2826	*	*(standardized)*	*SALES_SUM_ELV_PRODUCT_FALSE*	+
0,1989	*	*(standardized)*	*SALES_SUM_KAM*	+
0,0311	*	*(standardized)*	*SALES_SUM_PROD_SEGM_CONSTRUCTION*	+
0,0333	*	*(standardized)*	*SALES_SUM_PROD_SEGM_UTILITY*	+
0,1142	*	*(standardized)*	*SALES_SUM_CUM_SEGM_UTILITY*	+
-0,3027	*	*(standardized)*	*SALES_SUM_OPERATION_BY_MANUAL*	+
0,1212	*	*(standardized)*	*QUANTITY_SUM_ELV_PRODUCT_TRUE*	+
-0,0448	*	*(standardized)*	*QUANTITY_SUM_PROD_SEGM_UTILITY*	+
-0,0478	*	*(standardized)*	*QUANTITY_SUM_CUM_SEGM_UTILITY*	+
0,5237	*	*(standardized)*	*ORDERS_COUNT_UNIQUE_TOTAL*	+
-0,0432	*	*(standardized)*	*ORDERS_COUNT_UNIQUE_ELV_PRODUCT_FALSE*	+
0,3163	*	*(standardized)*	*ORDERS_COUNT_UNIQUE_KAM*	+
-0,1421	*	*(standardized)*	*ORDERS_COUNT_UNIQUE_PROD_SEGM_CONSTRUCTION*	+
-0,0239	*	*(standardized)*	*ORDERS_COUNT_UNIQUE_PROD_SEGM_UTILITY*	+
0,0556	*	*(standardized)*	*ORDERS_COUNT_UNIQUE_PROD_SEGM_INDUSTRY*	+
-0,0817	*	*(standardized)*	*ORDERS_COUNT_UNIQUE_PROD_SEGM_LIGHTING*	+
-0,1354	*	*(standardized)*	*ORDERS_COUNT_UNIQUE_CUM_SEGM_CONSTRUCTION*	+
-0,1124	*	*(standardized)*	*ORDERS_COUNT_UNIQUE_CUM_SEGM_INDUSTRY*	+
-0,1943	*	*(standardized)*	*ORDERS_COUNT_UNIQUE_CUM_SEGM_UTILITY*	+
-0,0407	*	*(standardized)*	*ORDERS_COUNT_UNIQUE_OPERATION_BY_MANUAL*	+
0,1929	*	*(standardized)*	*CUSTOMERS_COUNT_UNIQUE_TOTAL*	+
0,1657	*	*(standardized)*	*CUSTOMERS_COUNT_UNIQUE_ELV_PRODUCT_FALSE*	+
0,0227	*	*(standardized)*	*CUSTOMERS_COUNT_UNIQUE_ELV_PRODUCT_TRUE*	+
-0,4161	*	*(standardized)*	*CUSTOMERS_COUNT_UNIQUE_KAM*	+
0,0254	*	*(standardized)*	*CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_CONSTRUCTION*	+
-0,0452	*	*(standardized)*	*CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_UTILITY*	+
-0,0451	*	*(standardized)*	*CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_INDUSTRY*	+
0,0317	*	*(standardized)*	*CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_LIGHTING*	+
-0,0002	*	*(standardized)*	*CUSTOMERS_COUNT_UNIQUE_CUM_SEGM_CONSTRUCTION*	+
-0,0245	*	*(standardized)*	*CUSTOMERS_COUNT_UNIQUE_CUM_SEGM_INDUSTRY*	+
-0,0547	*	*(standardized)*	*CUSTOMERS_COUNT_UNIQUE_OPERATION_BY_MANUAL*	+
0,1928	*	*(standardized)*	*ITEMS_COUNT_UNIQUE_TOTAL*	+
-0,1648	*	*(standardized)*	*ITEMS_COUNT_UNIQUE_ELV_PRODUCT_FALSE*	+
0,0502	*	*(standardized)*	*ITEMS_COUNT_UNIQUE_KAM*	+
0,0353	*	*(standardized)*	*ITEMS_COUNT_UNIQUE_PROD_SEGM_CONSTRUCTION*	+
0,0055	*	*(standardized)*	*ITEMS_COUNT_UNIQUE_PROD_SEGM_INDUSTRY*	+
0,0133	*	*(standardized)*	*ITEMS_COUNT_UNIQUE_PROD_SEGM_LIGHTING*	+
-0,0781	*	*(standardized)*	*ITEMS_COUNT_UNIQUE_CUM_SEGM_CONSTRUCTION*	+
0,0468	*	*(standardized)*	*ITEMS_COUNT_UNIQUE_CUM_SEGM_INDUSTRY*	+
0,1263	*	*(standardized)*	*ITEMS_COUNT_UNIQUE_OPERATION_BY_MANUAL*	+
-0,0012				

The most accurate ML TS+O model was created by using SVR algorithm with data standardization. But adding the time series components to the regression model creates only more complicate equation (see Table 15) and don't add some performance to the model.

Table 15. SVR TS model equation with data standardization

			MARGIN_SUM_TOTAL	=
0,5484	*	*(standardized)*	*SALES_SUM_TOTAL*	+
0,3109	*	*(standardized)*	*SALES_SUM_ELV_PRODUCT_FALSE*	+
0,0197	*	*(standardized)*	*SALES_SUM_KAM*	+
0,0552	*	*(standardized)*	*SALES_SUM_PROD_SEGM_CONSTRUCTION*	+

*0,0562 * (standardized) SALES_SUM_PROD_SEGM_UTILITY +*
*0,0484 * (standardized) SALES_SUM_CUM_SEGM_UTILITY +*
*-0,3500 * (standardized) SALES_SUM_OPERATION_BY_MANUAL +*
*0,1526 * (standardized) QUANTITY_SUM_ELV_PRODUCT_TRUE +*
*-0,0192 * (standardized) QUANTITY_SUM_PROD_SEGM_UTILITY +*
*-0,0435 * (standardized) QUANTITY_SUM_CUM_SEGM_UTILITY +*
*0,0063 * (standardized) ORDERS_COUNT_UNIQUE_TOTAL +*
*0,1671 * (standardized) ORDERS_COUNT_UNIQUE_ELV_PRODUCT_FALSE +*
*0,2947 * (standardized) ORDERS_COUNT_UNIQUE_KAM +*
*-0,0162 * (standardized) ORDERS_COUNT_UNIQUE_PROD_SEGM_CONSTRUCTION +*
*-0,1151 * (standardized) ORDERS_COUNT_UNIQUE_PROD_SEGM_UTILITY +*
*-0,0185 * (standardized) ORDERS_COUNT_UNIQUE_PROD_SEGM_INDUSTRY +*
*-0,0917 * (standardized) ORDERS_COUNT_UNIQUE_PROD_SEGM_LIGHTING +*
*-0,1580 * (standardized) ORDERS_COUNT_UNIQUE_CUM_SEGM_CONSTRUCTION +*
*-0,0262 * (standardized) ORDERS_COUNT_UNIQUE_CUM_SEGM_INDUSTRY +*
*-0,0605 * (standardized) ORDERS_COUNT_UNIQUE_CUM_SEGM_UTILITY +*
*0,0834 * (standardized) ORDERS_COUNT_UNIQUE_OPERATION_BY_MANUAL +*
*0,0953 * (standardized) CUSTOMERS_COUNT_UNIQUE_TOTAL +*
*0,0998 * (standardized) CUSTOMERS_COUNT_UNIQUE_ELV_PRODUCT_FALSE +*
*0,0061 * (standardized) CUSTOMERS_COUNT_UNIQUE_ELV_PRODUCT_TRUE +*
*-0,1875 * (standardized) CUSTOMERS_COUNT_UNIQUE_KAM +*
*-0,1562 * (standardized) CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_CONSTRUCTION +*
*-0,0535 * (standardized) CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_UTILITY +*
*0,0167 * (standardized) CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_INDUSTRY +*
*0,0976 * (standardized) CUSTOMERS_COUNT_UNIQUE_PROD_SEGM_LIGHTING +*
*0,0794 * (standardized) CUSTOMERS_COUNT_UNIQUE_CUM_SEGM_CONSTRUCTION +*
*-0,0318 * (standardized) CUSTOMERS_COUNT_UNIQUE_CUM_SEGM_INDUSTRY +*
*-0,0787 * (standardized) CUSTOMERS_COUNT_UNIQUE_OPERATION_BY_MANUAL +*
*0,1310 * (standardized) ITEMS_COUNT_UNIQUE_TOTAL +*
*0,0083 * (standardized) ITEMS_COUNT_UNIQUE_ELV_PRODUCT_FALSE +*
*0,0197 * (standardized) ITEMS_COUNT_UNIQUE_KAM +*
*0,0481 * (standardized) ITEMS_COUNT_UNIQUE_PROD_SEGM_CONSTRUCTION +*
*-0,0071 * (standardized) ITEMS_COUNT_UNIQUE_PROD_SEGM_INDUSTRY +*
*-0,0346 * (standardized) ITEMS_COUNT_UNIQUE_PROD_SEGM_LIGHTING +*
*-0,1220 * (standardized) ITEMS_COUNT_UNIQUE_CUM_SEGM_CONSTRUCTION +*
*0,0148 * (standardized) ITEMS_COUNT_UNIQUE_CUM_SEGM_INDUSTRY +*
*0,2006 * (standardized) ITEMS_COUNT_UNIQUE_OPERATION_BY_MANUAL +*
*0,0018 * (standardized) DayOfWeek=sun +*
*0,0006 * (standardized) DayOfWeek=mon +*
*0,0024 * (standardized) DayOfWeek=wed +*
*0,0015 * (standardized) DayOfWeek=wed +*
*0,0054 * (standardized) DayOfWeek=thu +*
*-0,0123 * (standardized) DayOfWeek=fri +*
*0,0007 * (standardized) DayOfWeek=sat +*
*0,0019 * (standardized) DayOfWeek=sat +*
*-0,0047 * (standardized) DATE-remapped +*
*-0,0005 * (standardized) Lag_MARGIN_SUM_TOTAL-1 +*
*-0,0003 * (standardized) Lag_MARGIN_SUM_TOTAL-3 +*
*-0,0002 * (standardized) Lag_MARGIN_SUM_TOTAL-4 +*
*0,0010 * (standardized) Lag_MARGIN_SUM_TOTAL-5 +*
*-0,0081 * (standardized) Lag_MARGIN_SUM_TOTAL-6 +*
*-0,0123 * (standardized) Lag_MARGIN_SUM_TOTAL-7 +*
*0,0108 * (standardized) DATE-remapped^2 +*
*-0,0075 * (standardized) DATE-remapped^3 +*
*0,0008 * (standardized) DATE-remapped*Lag_MARGIN_SUM_TOTAL-1 +*
*-0,0003 * (standardized) DATE-remapped*Lag_MARGIN_SUM_TOTAL-2 +*
*0,0002 * (standardized) DATE-remapped*Lag_MARGIN_SUM_TOTAL-3 +*
*0,0001 * (standardized) DATE-remapped*Lag_MARGIN_SUM_TOTAL-4 +*
*0,0006 * (standardized) DATE-remapped*Lag_MARGIN_SUM_TOTAL-5 +*
*0,0064 * (standardized) DATE-remapped*Lag_MARGIN_SUM_TOTAL-6 +*
*0,0269 * (standardized) DATE-remapped*Lag_MARGIN_SUM_TOTAL-7 +*
0,0049

6 Summary

The master’s thesis focuses on the application of machine learning methods to sales forecasting. For B2B sales organisations, the topic of this work adds value across an organization. With the accurate forecasting, sales organizations can make correct decisions with respect to sales resource allocation.

The objective of the master's thesis was to find the best features and models for sales margin forecasting using machine learning methods and historical data from a particular B2B company. Using the supervised machine learning approach, regression models were created by different ML algorithms and based on different data setups.

As a preliminary stage of modelling processes, we create a multivariate time series dataset from raw data using feature engineering techniques. With the help of feature selection methods, the optimal number of features was determined to optimize the performance of the models. Created dataset we will later transform to multivariate dataset without time series component for regression models building, and to univariate time series dataset for time series regression models training and testing. As an intermediate conclusion, we stated good data quality but the absence of variables with possible predictive power.

From the available algorithms in Weka software, we have chosen for experimentation the ones that are most widely recommended in the research literature - Linear Regression, Support Vector Regression and Random Forest Regression.

As a result of modelling, created ML models were compared by RMSE and MAE evaluation metrics. The most optimal models trained by each algorithm were used for short-term multistep forecasting of target variable.

Comparative metrics analysis showed that, the Linear regression and Support Vector regression models without time series components have the best performance. In terms complexity of the equation, we choose the Linear regression model as a reference.

Time series regression model, regardless of the used algorithm, shown the significantly lower predictive power than regression models based on correlated features. The reason for this is the apparent inadequacy in review of only target variable and its time series components, without taking into account the influence of external factors.

In combination of time series components with additional overlay data, we have a better result than just based in time series components, but still slightly worse compared to pure regression models.

As a result, the final choice was made in favour of the Linear regression model and Support Vector regression time series model with overlay data. Both models have comparable accuracy of target variable, but last one is overfitted. Based on experience gained in this research, we can confirm that the B2B sales forecasting is mostly the regression than time series problem. The using of time series components in most experimental case doesn't give additional predictive power to the model, but rather degrades it.

Analysing the business and available raw data we assume that data sufficiency could be improved. The additionally stored historical data, such as geographical sales distribution, can help to improve the model performance and show additional pattern in data.

The machine learning algorithms in practice can help to improve the quality of data-driven decision-making processes by discovering patterns in big data.

References

[1] M.Kaput, “AI for Sales: What You Need to Know”, Marketing Artificial Intelligence Institute, *[www]*, accessed 08.11.2021.

[2] P.Martin, M.Castaño, R.Lopez, “Building a sales prediction model for a retail store”, Neural Designer platform blog, *[www]*, accessed 17.01.2022.

[3] Amazon Web Services Whitepaper, “Time Series Forecasting Principles with Amazon Forecast”, *[www]*, accessed 08.11.2021 .

[4] M.Döring, “Prediction vs Forecasting”, Data science blog, *[www]*, accessed 17.01.2022.

[5] R.J.Hyndman, G.Athanasopoulos, “Forecasting: Principles and Practice (2nd edition)”, online textbook, *[www]*, accessed 20.02.2022.

[6] J.C.Chambers, S.K.Mullick, D.D.Smith, “How to Choose the Right Forecasting Technique”, magazine article, *[www]*, accessed 20.02.2022.

[7] L.Soares, “Sales Forecasting: from Traditional Time Series to Modern Deep Learning”, The Medium blog, *[www]*, accessed 17.01.2022.

[8] K.Eby, “The Last Guide to Sales Forecasting You’ll Ever Need: How-To Guides and Examples”, *[www]*, accessed 08.11.2021.

[9] Pennsylvanian State Univercity of Science, “Regression Methods”, open lesson, *[www]*, accessed 20.02.2022.

[10] B.Hidalgo, M.Goodman, “Multivariate or Multivariable Regression” *[DOI:10.2105/AJPH.2012.300897]*.

[11] N. S. Arunraj, D. Ahrens, M. Fernandes, “Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry”, International Journal of Operations Research and Information Systems, Vol.7 Is.2, *[DOI:10.4018/IJORIS.2016040101]*.

[12] B.Billaha, M.L.King, R.D.Snyder, A.B.Koehler, “Exponential smoothing model selection for forecasting”, International Journal of Forecasting, Vol. 22, Is 2, *[DOI:10.1016/j.ijforecast.2005.08.002]*.

[13] D.H.Maulud, A.M.Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning”, Journal of Applied Science and Technology Trends. Vol. 1, Is. 4, *[DOI:10.38094/jastt1457]*.

[14] G. Zhang, B. Eddy Patuwo, M. Y. Hu, “Forecasting with artificial neural networks: The state of the art”, International Journal of Forecasting. Vol. 14, Is. 1, *[DOI:10.1016/S0169-2070(97)00044-7]*.

[15] M.Rakhra, P.Soniya, D.Tanwar, P.Singh, D.Bordoloi, P.Agarwal, “Crop Price Prediction Using Random Forest and Decision Tree Regression: A Review”. Journal Materials Today: Proceedings, *[DOI:10.1016/j.matpr.2021.03.261]*.

[16] M.d.C.Moura, E.Zio, I.D.Lins, E.Droguett, “Failure and reliability prediction by support vector machines regression of time series data”, Journal Reliability Engineering & System Safety. Vol. 96, Is. 11, *[DOI:10.1016/j.ress.2011.06.006]*.

[17] A.Ibrahim, R.Kashef, M.Li, E.Valencia, E.Huang, “Bitcoin Network Mechanics: Forecasting the BTC Closing Price Using Vector Auto-regression Models Based on Endogenous and Exogenous Feature Variables”, Journal of Risk and Financial Management. Vol. 13, Is. 9, *[DOI:10.3390/jrfm13090189]*.

[18] M.Bohanec, M.K.Borštnar, M.Robnik-Šikonja, “Integration of machine learning insights into organizational learning. A case of B2B sales forecasting”. Conference Paper, *[www]*, accessed 17.01.2022.

[19] B.Ramya, K.Vedavathi, “An Advanced Sales Forecasting Using Machine Learning Algorithm”, International Journal of Innovative Science and Research Technology, Vol. 5, Is. 5, *[www]*, accessed 13.03.2022.

[20] A.Krishna, Akhilesh V, A.Aich, C.Hegde, “Sales-forecasting of Retail Stores using Machine Learning Techniques”, 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, *[DOI:10.1109/CSITSS.2018.8768765]*.

[21] Dr.C.Shyamala, P.Sabarish, T.Vignesh, S.Yogeendran, “Walmart Sales Prediction Using Machine Learning Algorithms”, Annals of RSCB, Vol. 25, Is.4, *[www]*, 20.02.2022.

[22] G.Nunnari, V.Nunnari, "Forecasting Monthly Sales Retail Time Series: A Case Study” 2017 IEEE 19th Conference on Business Informatics, *[DOI:10.1109/CBI.2017.57]*.

[23] S.Cheriyan, S.Ibrahim, S.Mohanan, S.Treesa, “Intelligent Sales Prediction Using Machine Learning Techniques”, 2018 International Conference on Computing, Electronics & Communications Engineering, *[DOI:10.1109/iCCECOME.2018.8659115]*.

[24] B.M. Pavlyshenko, “Machine-Learning Models for Sales Time Series Forecasting”, 2018 IEEE Second International Conference on Data Stream Mining & Processing, *[DOI:10.3390/data4010015]*.

[25] “Comparison of Data Analysis Tools: Excel, R, Python and BI”, *[www]*, accessed 02.04.2022.

[26] H.Zhou, “Learn Data Mining Through Excel: A Step-by-Step Approach for Understanding Machine Learning Methods”, *[DOI:10.1007/978-1-4842-5982-5]*.

[27] Waikato University webpage “About Weka tool”, *[www]*, accessed 17.09.2021.

[28] PhD J. Brownlee, “Time Series Forecasting as Supervised Learning”, *[www]*, accessed 17.01.2022.

[29] R.Wirth, J.Hipp, “CRISP-DM: Towards a Standard Process Model for Data Mining”, *[www]*, accessed 10.11.2021.

[30] A.Zheng, A.Casar, “Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists”, *[www]*, accessed 21.01.2022.

[31] H.Bui, “How to Encode Categorical Data”, Towards Data Science, *[www]*, accessed 20.02.2022.

[32] Pourya, “Time Series Machine Learning Regression Framework”, Towards Data Science, *[www]*, accessed 17.01.2022.

[33] A.J.Ferreira, M.A.T.Figueiredo, “Efficient feature selection filters for high-dimensional data”, *[DOI:10.1016/j.patrec.2012.05.019]*.

[34] Olukoya, B.Musiliu, “Comparison of Feature Selection Techniques for Predicting Student’s Academic Performance”. International Journal of Research and Scientific Innovation. Vol. 7, Is. 8, *[www]*, accessed 16.02.2022.

[35] K.Kausthub, “Commercials Sales Prediction Using Multiple Linear Regression”. International Research Journal of Engineering and Technology. Vol. 8, Is. 3, *[www]*, accessed 21.01.2022.

[36] C.Wu, J.Ho, D.T.Lee, “Travel-time prediction with support vector regression“, IEEE Transactions on Intelligent Transportation Systems journal. Vol. 5, Is. 4, *[DOI:10.1109/TITS.2004.837813]*.

[37] A.R.S.Parmezana,V.M.A.Souzaa, G.E.A.P.A.Batista, “Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model”. Information Sciences. Vol. 484, *[DOI:10.1016/j.ins.2019.01.076]*.

[38] L.Breiman, “Random forests“. Mach. Learn, Vol. 45, Is. 1, *[DOI:10.1023/A:1010933404324]*.

[39] C.J.Willmott, K.Matsuura, “Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance” Climate Research, Vol. 30, Is. 1, *[www]*, accessed 16.02.2022.

[40] Wikipedia contributors, "Standard deviation," Wikipedia, The Free Encyclopedia, *[www]*, accessed 02.04.2022.

[41] I.Tsamardinos, G.Borboudakis, P.Katsogridakis, “A greedy feature selection algorithm for Big Data of high dimensionality”. Mach Learn, Vol 108, *[DOI:10.1007/s10994-018-5748-7]*.

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Igor Roos

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “Sales Forecasting Using Machine Learning Methods: The Case of B2B Company”, supervised by Avar Pentel
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

10.05.2022

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2

Link to the repository containing the data transformation tables and files with datasets that took part in the project.

https://gitlab.cs.ttu.ee/igroos/sales-forecasting-using-machine-learning-methods/

- Raw data sample of one month in 2019
- Transformed multivariate time series dataset
- List of features with correlation coefficient
- Weka
    - Datasets
    - Model files