Ministry of Education and Science of the Republic of Kazakhstan
Suleyman Demirel University

UDC 50.10

Dias Nurmambetov

Planning and data analysis system using machine learning algorithms

THESIS

Presented in Partial Fulfillment for the
Degree of Master of Science in Computing Systems and Software
(degree code: 6M070400)
Department of Computer Sciences
Faculty of Engineering and Natural Sciences

Supervisor: **Assist. Prof. PhD Andrey Bogdanchikov**

Kaskelen, 2020

Abstract

The information system, which includes data analysis, data clustering, as well as sales amount prediction using machine learning algorithms will be described in this study. The system is intended for work in a manufacturing company, that sells certain products to retail outlets. According to the Forbes article, "Ten Ways Big Data Is Revolutionizing Supply Chain Management" [1], companies tend to integrate big data and advanced analytics into optimization tools. Thus, the idea of the system was proposed to increase efficiency of business processes. The system includes a method which can be used for outlets clustering using unsupervised and supervised machine learning algorithms comprising 2 steps – data partition using unsupervised Gaussian Mixture (GM) Model clustering algorithm based on outlets sales amount and further partition for one of them using Logistic Regression (LR) and Neural Networks (NN) classification algorithms, which predict whether outlets will achieve monthly sales plan. The proposed algorithm was tested on real sales data and formed 3 clusters according to business needs. Sales plan achievement prediction gave up to 74% accuracy. Also, system should include plans settling algorithm based on sales forecasting per each outlet. This part was implemented using regression ensembling algorithm XGBoost, time series model SARIMA and recurrent neural network LSTM model. Sales forecasting was based on weekly sales data for 5 years in the outlet and predicts future sales amount values. Models were compared in terms of Mean Absolute Percentage Error (MAPE) and running time. From the results, it was decided to choose LSTM model which gave 17% MAPE and good running time to embed to the system with possible improvements.

Аңдатпа

Бұл жұмыста деректерді талдауды, мәліметтерді кластерлеуді, сондай-ақ машиналық оқыту алгоритмдерін қолдана отырып сату көлемін болжауды қамтитын ақпараттық жүйе сипатталады. Бұл жүйе белгілі бір өнімді сауда орындарына сататын өндірістік компанияда жұмыс істеуге арналған. «Үлкен мәліметтердің жеткізілім тізбегін басқаруды жаңартатын он тәсілі» [1] Forbes мақаласына сәйкес, компаниялар үлкен мәліметтер мен озық аналитикаларды оңтайландыру құралдарына біріктіруге бейім. Осылайша, бизнес-процестердің тиімділігін арттыру үшін жүйенің идеясы ұсынылды. Бұл жүйеде бақыланбайтын және бақыланатын машиналық оқыту алгоритмдерін қолдана отырып, екі сатыдан тұратын бақыланбайтын Гауссиан қоспасы (GM) модельді кластерлік алгоритмді қолдана отырып, сауда нүктелерінің сату көлеміне негізделген және логистикалық регрессияны қолдана отырып, оның бөлігін кластерлеу үшін қолдануға болатын әдіс бар. (LR) және Neural Networks (NN) жіктеу алгоритмдері, олар сауда нүктелерінің ай сайынғы сату жоспарына жететіндігін болжайды. Ұсынылған алгоритм нақты сату деректері бойынша сыналды және бизнес қажеттіліктеріне сәйкес 3 кластер құрады. Сатылым жоспарының жетістігін болжау 74 % дәлдігіне дейін жеткіздік. Сондай-ақ, жүйеде әр сауда нүктенің сатылымын болжауына негізделген жоспарларын құратын алгоритм қажет. Бұл бөлім XGBoost деген құрама регрессиялық алгоритмін, SARIMA уақыттық модельдерін және LSTM қайталанатын нейрондық желіні қолдану арқылы орындалды. Сатылымдарды болжау бөлімшесі 5 жылдағы апта сайынғы сатылымдар туралы мәліметтерге негізделген және болашақта сату көлемінің болжамын жасайды. Модельдер орташа абсолюттік қате (MAPE) және жұмыс уақыты бойынша салыстырылды. Нәтижелерден LSTM моделін таңдау туралы шешім қабылданды, ол 17 % MAPE және жақсы жұмыс уақытын көрсетті де, жетілдірулерден кейін жүйеге ене алады.

Аннотация

В этом исследовании будет описана информационная система, которая включает анализ данных, кластеризацию данных, а также прогнозирование объема продаж с использованием алгоритмов машинного обучения. Система предназначена для работы в производственной компании, которая продает определенные продукты в розничные магазины. Согласно статье Forbes, «Десять способов как большие данные революционизируют управление цепочками поставок» [1], компании стремятся интегрировать большие данные и расширенную аналитику в инструменты оптимизации. Таким образом, идея системы была предложена для повышения эффективности бизнес-процессов. Система включает в себя метод, который можно использовать для кластеризации торговых точек с использованием алгоритмов машинного обучения без учителя и с учителем, состоящих из двух этапов - разбиение данных с использованием алгоритма кластеризации без учителя модели Гауссовой смеси (GM) на основе объема продаж торговых точек и дополнительного разделения для одного из них с использованием логистической регрессии (LR) и алгоритма классификации на базе нейронных сетей (NN), которые предсказывают, будут ли торговые точки достигать ежемесячного плана продаж. Предложенный алгоритм был протестирован на реальных данных о продажах и сформировано 3 кластера в соответствии с потребностями бизнеса. Прогноз достижения плана продаж дал точность до 74 %. Кроме того, система должна включать алгоритм расчета планов, основанный на прогнозировании продаж для каждой торговой точки. Эта часть была реализована с использованием алгоритма регрессии ансамблей XGBoost, модели временных рядов SARIMA и модели рекуррентной нейронной сети LSTM. Прогнозирование продаж основывалось на еженедельных данных о продажах за 5 лет в торговой точке и прогнозировало значения будущих продаж. Модели сравнивались по средней абсолютной ошибке в процентах (MAPE) и времени выполнения. Из результатов было решено выбрать модель LSTM, которая дает 17% MAPE и хорошее время вычисления для встраивания в систему с возможными улучшениями.

Acknowledgements

The author wants to thank thesis supervisor for regular support and new challenges, which helped to improve this work. Thanks to my family for their patience and full support. Thanks to colleagues and supervisor for their support and trust.

Contents

Nomenclature

ACF Auto-Correlation Function

AI Artificial Intelligence

ARIMA Autoregressive Integrated Moving Average

BI Business Intelligence

CRM Customer Relations Management

ERP Enterprise resource planning

GM Gaussian Mixture

HDFS Hadoop Distributed File System

IoT Internet of Things

IQR Interquartile Range

KNN K-nearest neighbors

LGBM Light Gradient Boosting Model

LiR Linear Regression

LR Logistic Regression

LSTM Long Short-Term Memory

MAPE Mean Absolute Percentage Error

MLP Multilayer Perceptron

NN Neural Networks

PACF Partial Auto-Correlation Function

RNN Recurrent Neural Network

SARIMA Seasonal Autoregressive Integrated Moving Average

SoE Systems of Engagement

SoR Systems of Record

VA Visual analytics

1. Introduction

1.1 Motivation

Segmentation of retail outlets in terms of manufacturing companies' strategy applied in sales amount and trade activities for each of them is very important. A Harvard Business School professor said that of 30,000 new consumer products launches each year, 95% fail because of ineffective segmentation. [2]

Directed investments into outlets help companies to make more profit and decrease expenses. According to Harvard Business review, sales teams adopting AI are seeing an increase in leads and appointments of more than 50%, cost reductions of 40–60%, and call time reductions of 60–70% [3]

Gartner published a paper titled, "Demand Forecasting Leads the List of Challenges Impacting Customer Service Across Industries". [4] From the title of the article, they conclude that accurate demand forecasts across all industries is very important for business.

Thus, an idea of developing the information system came up, for our manufacturing company which consists of outlets segmentation part and sales amount forecasting one.

1.2 Aims and Objectives

The main idea of the research is to create complex system which will make company's work more efficient and increase return of it's directed investments. Which helps to analyze sales data for manufacturing or retail companies with implementing segmentation of outlets and sales amount prediction for selected segments. Since there are studies related to forecasting or retail segmentation itself, company needs a combined solution, which could be accessed by different users in

different locations. For the company it is also outlets segmentation accuracy - any pair of outlets from one segment is more similar to each other than to a POS from another one. Groups are calculated in a few minutes instead of a few days, which decreases labor cost.

During developing the system, main objectives are:

- To study machine learning algorithms and select ones, which could be used for system's models development
- Get a pool of machine learning and clustering algorithms to be used
- Conduct a literature review in terms of thesis field to find out similar works or works in the field
- To use data wrangling and analysis tools and algorithms to define necessary parameters and dependencies of data
- Define data to be used, including historical data
- Implementation of data analysis algorithms and testing
- Get results and evaluate performance of algorithms

1.3 Working progress

The first step was to review machine learning general types and principles to choose appropriate algorithms and methods to the problem. There are many different types of machine learning algorithms, which can be useful depending on purposes and used data.

Supervised machine learning algorithms are those algorithms, which require algorithms' training before it can predict or classify outputs. The input dataset can be divided into train and test datasets. Cross-validation is one of the approaches, which is to divide input data into training and test sets with the aim of avoiding overfitting and use one of the parts as testing set one by one. If the output of implemented algorithm is from continuous set, that kind of problems are called regression problems, and it is called classification if the output set is discrete.

Supervised algorithms are used, if there are "correct" answers for the input data. Then classification accuracy can be estimated comparing "correct" answers with algorithm's output as share of correctly predicted from total number of answers. Unsupervised learning algorithms do not learn any features from training data, and it is mainly used for clustering or decreasing dimension of data. [5]

The Neural Networks are based on a biological concept of neurons. A neuron is a cell in a brain, which communicates other ones by exchanging signals. NN include input and output layers and can also have several hidden layers. It can discover complex dependencies based on input data and find out non-linear functions to predict target data. [6]

The preparation phase began with a selection of clustering algorithms to segment the points according to sales amount criteria. Many clustering algorithms and methods were considered during this thesis preparation [7],[8],[9], since there is a huge variety, including the most popular K-means algorithm [10], as well as the Gaussian mixture algorithm [11] (an algorithm based on probabilities for each sample of relation to each cluster) and complex models with ensembles of clustering algorithms [12].

Further:

1. Using clustering algorithms, system can create outlets clusters based on certain parameters

2. Determine patterns and dependencies between the data

3. Additionally, develop a model of prediction the fact of achieving settled plan for outlets using the previous implementation of plans information and sales in the outlet dynamics data

4. To build a model of predicting monthly plans (sales amount) for each outlet, considering the purchase historical data and characteristics of outlets or its' sellers.

   For the sales amount prediction, other papers will be studied to define best performing regression time series algorithms. [13]

There are a lot of different separate problems regarding market segmentation and forecasting. The goal of the thesis is to combine developing a complex infor-

mation system for sales analytics including segmentation (stage 1) and forecasting (stage 2) machine learning models.

1.4 Thesis Outline

The first chapter is Introduction chapter. It is the current chapter and it gives work overview and background. In Chapter 2 we review related literature and find out the main objective. Chapter 3 is describing proposed methods and algorithms used in the information system. Chapter 4 is discussing results, and in Conclusion chapter the main conclusions and further steps are discussed.

2. Literature review and preliminaries

2.1 Literature review

2.1.1 Advanced analytics systems

During the study, papers describing the approaches in building and comparing different analytics systems were reviewed.

The first question to be answered was - what if a software ready for all our business needs already exists. In paper "Visual Analytics for the Big Data Era - A Comparative Review of State-of-the-Art Commercial Systems"[14] comparative analysis of most popular commercial systems, which provide automatic data analysis is presented. However, it does not include Microsoft Power BI, because the research was made before its appearance. It were compared within functional capabilities, visualization methods and system performance.

Visual analytics (VA) is becoming very popular and software companies which develop such systems are growing exponentially. Also, it is the trend that big enterprise solutions vendors are acquiring VA technology companies to integrate it with their existing tools. Development of the area follows from the increasing volume of data generated by business. This data is generated from everywhere - social media, different sensors in a hardware applied in any industries, GPS data and other text and media files from all around the Internet. VA is using visualization of data together with analytics capabilities of the computer and data analyst[15] to solve business tasks in the area of its application.

With the increasing number of analytics solutions, including open-source and commercial ones, there is a need of overviews of the best ones to allow users and

businesses to select the appropriate for their needs. Before, there were studies related to the comparison of open-source applications and commercial ones.

Indeed, using available solutions is more efficient than creating a new system, but it can be hard to find the appropriate system for our business needs and existing software and analytics environment. The paper "Comparison of Open Source Visual Analytics Toolkits"[16] provides a comparative review of open-source VA solutions with respect to the quality of visualization, capabilities for analysis and the interface of applications itself. It concluded from the comparison of number of toolkits that there are both advantages and flaws for each of them, most of them are good in one area of application or in some functionalities and finding the best one for the user implies defining their needs and prioritizing the most important ones. And such studies can help with making a choice. [14] is about commercial systems which usually have wider range of functions than open source tools, but acquainting with the reviews we can decide and choose free or commercial one.

The main difference of commercial systems from open-source ones is that commercial systems are much more easier to integrate with existing environment or the service and integration are included in the cost of maintenance from vendor. There are several research companies which provide a review of commercial Business Intelligence (BI) systems. BI systems typically are commercial analytics systems, which are integrated tools including data extraction, analysis, visualization, prediction and all other necessary analytics functionalities. Other VA tools can be a system which solves only some part of analytics problems and usually are integrated with other systems. Every year Gartner publishes "Magic Quadrant for Analytics and Business Intelligence Platforms"[17] with the review of the best-known commercial BI systems with respect to its capabilities and integrity. Capabilities is about 15 areas, including security, cloud calculations and advanced analytics ability, auto insights generation or natural language using for building reports. Nowadays, such advanced features like ability to create a report when query is made as just a business request text or cloud computations are becoming very common and must-have for good commercial systems.

In Figure 2.1 there is a quadrant of companies for 2020. Microsoft is presented with their Power BI product and it has been a leader for the last several years together with Tableau (which was acquired by Salesforce which is presented on

Visionaires quarter), Qlik (with Qlik Sense) and ThoughtSpot. The report is enriched with the description of strengths and cautions for each of the companies presented.

Gartner's rating includes only the BI companies which met Gartner's standards in terms of trends, number of clients and potential of their product. It is based on Gartner's analysts opinion, online survey of real users and some other questionnaires. And two axes presented which are ability to execute and completeness of vision are presented with their own list of parameters. Ability to execute assesses the capabilities of the system and user experience which are available now in the products; whereas completeness of vision is about the strategy of companies expressed in the future plans for product development, market potential understanding and so on.

Comparing to the existing reviews of VA tools, [14] provides:

- a wider research including not only BI tools but VA products in general
- evaluation of systems capability to work with Big Data, i.e. huge amounts of data with irregular structure, whereas existing surveys are researching usability of the systems
- testing on real datasets to compare the performance of systems, whereas existing surveys are based on users opinions

In the paper VA tools from different categories were selected, to cover maximum of the whole market. As it was mentioned, besides BI systems, it were also data visualization software, network analysis products or some other niche products. First, systems were surveyed in terms of data management, visualization, analysis and performance. Then, systems were tested for performance with a stress test and data analysis, visualization with 2 real datasets.

According to [15], there are three steps in visual analytics:

- data management - it is about extracting data from different sources, merging, preprocessing it to create a dataset which is ready to be used in analysis
- data modeling - analytics itself including models creation, connection of data

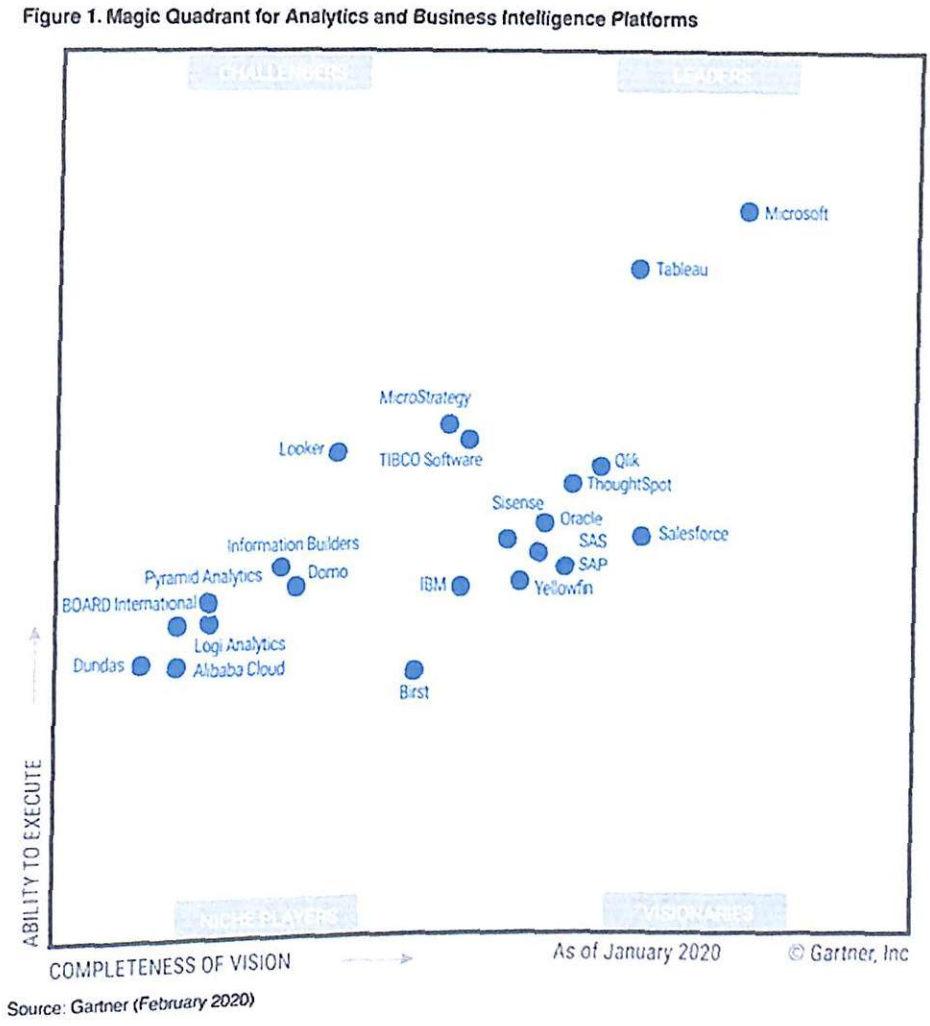

Figure 2.1: Gartner's Magic Quadrant for Analytics and BI Platforms 2020.

- visualization - using best appropriate data visualization techniques and approaches, to create visual representation of data in order to be able to make insights from it

Let's look at each of them some more detailed.

Data management includes not only the functionality of data loading and connection to the data sources, but also data handling which is preprocessing, pivoting or other data transformation according to our needs. Analytics systems can differ in terms of data management by the list of available data sources to connect or possibility to read from different distributed sources in parallel which is aligned with the development of Big Data. Data handling or preprocessing is very important part of analysis, which predetermines a success of data analysis. In the paper, systems were compared with respect to data preprocessing functionalities available, in the same time in our study some of the mentioned steps and types of preprocessing from the paper were also implemented:

- data cleaning including replacing missing values with zeros or average values and outliers detection
- data normalization
- columns and rows aggregation, create new columns based on existing ones
- joining tables
- grouping, pivoting and filtering

In terms of data modeling the main categories which are reviewed in the paper are - statistics functions, machine learning algorithms and dimension reduction algorithms. All of these tools help to analyze data from different views and to extract insights. Most of the reviewed systems are able to implement statistics and machine learning, the difference is in the variety of algorithms for each type. And only two systems (JMP and Spotfire) support dimension reduction algorithms.

Statistics are about calculating statistics functions for one variable such as mean, standard deviation and so on, for two variables such as correlation, and less common multivariate analysis. This type of data modeling is not implemented in our system since the main idea was to apply machine learning algorithms to effective forecasting according to the business needs. Thus, all types of machine learning algorithms such as clustering, classification and regression algorithms were implemented in our system. Dimension reduction algorithms are usually used to change the multidimensional data to 2D or 3D to be able to visualize it on a graph. Such kind of algorithms imply that data retain the mutual distance between data samples or their orientation in space. In our system, dimension reduction were also tested in classification problem to reduce the dimension to 2D space and plot the data by clusters.

In terms of visualization the systems in the paper were compared in a set of available graph visualizations and the interface capabilities such as advanced interactive filtering, drilling up and down. Most commonly used visualized data type in numeric and it is the only type used in our system. There were some categorical data, but it was translated to the numeric with discrete values.

And most of the systems support the most common and standard graph types used for plotting not very high dimensional numeric data - line charts, bar charts, pie charts, column charts, scatter charts and histograms. In our system we also

used scatter plots and line charts to visualize data. And only few systems support graphs for multivariate data. Each type of graphs is used depending on data and what the author wants to show. For example, to show the dynamic change of the variable through time line charts are used with time put on horizontal axis. Pie charts are used to show the parts of the whole, whereas bar and column charts are for comparison of values clustered by some another parameter. Scatter charts and histograms are used to show the relations between variables and distribution of the variable respectively. In [18] there is a chart suggestion diagram - an universal tool to find the appropriate type of chart for numeric data. Tableau and JMP are both also able to create visualization recommendations based on data (by the date of research made). It is also available in Power BI, which was created later than the research, as it was mentioned. Capabilities of interactive interface are the main strengths of some systems in terms of visualization. It is about filtering data interactively (simultaneously for several charts), changing the level or scale of the data, ability to view the data in different scales at the same time. This kind of functionality can be also added to our system, but the most effective way is to integrate with existing VA systems which support the most common types of visualization since it is not the main purpose of our system.

System performance is also important if we deal with big amount of data. And in the paper authors provided stress tests with various sizes of CSV files to be loaded into the system. And only Spotfire was able to load 50 GB file, whereas Tableau and JMP were able to load 20 GB and Qlik only 10 GB. Similar approach will be used in our study - all forecasting algorithms will be tested for running time and it will be also considered to choose the best one.

The tasks which are supported by all systems are in these categories:

- data exploration, which is mostly about creating hypotheses based on data available. It includes data visualization (usually histograms, box plots or scatter plots to find the structure of data, sometimes heat maps), statistical models and even machine learning algorithms.

- dashboards, which is about set of graphs and visual representation of data. Dashboards usually implies an interactive interface with connected graphs which can be filtered, selected to extract insights from the data dependencies among graphs

- reporting, which are used for results overview. Reports are usually communicated to interested parties with request or on a regular basis. It comprises the main indicators that describe the state of the business at that time
- alerting, which means making signals to user when something important happens.

Mostly, our study is related to data exploration - to find out patterns from the data using available machine learning algorithms and visualization techniques. In terms of exploration methods only few systems, which are QlikView, Spotfire, JMP, ADVIZOR, support machine learning and predictive algorithms. Spotfire, comparing to other systems, has advanced predictive analytics, ADVIZOR has some not common interactive graphs. And the list of algorithms available there is limited with the standard ones. Because of it, defining business needs and adding all necessary functionality to the separate system product has become a better choice.

According to "A systems approach to big data technology applied to supply chain"[19], which describes the approaches and methods regarding new analytics systems integration to business processes, there are 2 types of systems - SoR (Systems of Record) and SoE (Systems of Engagement). SoR are the systems, which support basic supply chain operations - production, planning and sales itself. And SoE is a systems group which refers to using advanced analytics and getting insights from it to increase business efficiency or find opportunities to grow. It includes monitoring, prediction, visualization and other methods described in the paper.

Also, integration process of SoR and SoE systems is described in a detailed way, which could be also useful for connecting our system with sales database. Nowadays, penetration of digital technologies into other aspects of business is called as a digital transformation. And it is seen as integration of business processes and advanced analytics. For example, smart ordering systems are using demand forecasting to create an optimal number of orders and products for the retailers. Sales forecasting is also used as a complement for planning of production by manufacturing companies.

The transformation of applying digital is also going on. Before, only static reporting and analysis results were used for improving product strategy or making decisions. Now it is a trend of dynamic analytics when the results of analysis are

applied automatically, without a human. Then, strong settled business processes together with integrated analytics lead to effective and optimal business decisions and profits increase.

So, for SoR it is important to be stable and reliable since all business processes are based on it. On the other hand, SoE is becoming connected with many sensors, data sources and with increasing amount of data, it must be able to retain stable and scalable even with extension of the system.

According to the paper there are some important points that should be considered when integrating these 2 types of systems:

- system integration method - it is better to consider API connection if it is necessary to ensure high synchronicity and to use files or intermediate servers when it is about large amount of data because it could be hard to provide reliable integration with big amounts of data

- code architecture - it can be an issue when it is necessary to analyze merged data from different systems and the structure, names of columns and export methods, so it requires additional preprocessing from analysts or even additional transforming rules in case of automatic integration during data exchange

- division of microservices - it is about the connection between the systems which implies separate integration between the parts of the system because it can differ in requirements for stability, reliability or frequency of use. For example, microservices of searching in on ecommerce site and ordering have different requirements since searching is more frequent operation and ordering should not allow any mistakes

- system integration model - it can be centralized when the whole SoR and SoE are integrated within one integration platform and distributed when the parts of each system are integrated separately as in case of microservices architecture. Centralized model allows to control and analyze the system as a whole, but it could need higher requirements to reliability since it manages more data simultaneously

Concluding, systems have evolved from static integration to dynamic analytics (SoE) integrated with business systems (SoR) and it will continue with the de-

velopment of Artificial Intelligence (AI) and Internet of Things (IoT). And such integration applied to any companies leads to the increase in profit and work efficiency.

There are also some related works which describe sales analysis tools developed for the specific purposes of the company. The paper called "Analysis and Optimization of Online Sales of Products" [20] tells about the similar system as in this thesis but for online shopping companies. Since, e-commerce companies are having more and more data from website transactions, clients actions and payments, it is very necessary to create such tools which are helpful in achieving business and sales goals.

The system described in the paper is based on affinity analysis to determine connections between some transactions and simple logistic, linear regression.

The system can be used to analyze data comprising of steps listed:

- export data from the sales transactions database
- segmentation of the data
- extract some insights from sales data such as trends or seasonality and so on
- project the insights into recommendations within sales and business improving tasks

The importance of using analysis tools for increasing sales and work effectiveness in e-commerce is mentioned in the paper. It is the same as for supply chain or retail companies which is this thesis is dedicated for.

Several data mining approaches are used for the system development. Data mining helps to find out some patterns in sales or transaction logs and then use the obtained insights to rearrange stocks or come up with some promo. For example, to increase the stock or purchasing for fast selling products is also the tool which is used in retail companies. Patterns recognition in the sales data and creating insights are very important and can be useful for decision making or strategic planning by the company. The difference of this thesis is that we do not consider every item specifically, but sales for all products within the achievement of monthly full portfolio plans. In the paper, also affinity analysis is used to find out connections between the sales of different products.

The system described in the paper used sales data for each item, data about basket consisting of items combination and customers' feedback regarding each product. Affinity analysis used in the system implies analyzing relations between products in the basket to define main products and the ones which are always bought as an addition. As well as defining the most effective promo campaigns or trade activities. Logistic regression is used to define whether the specific item will be sold or not, since the algorithm is binary classification model. It is similar to the approach used in stage 1 of this thesis to define whether outlet will achieve the plan or not. And Linear Regression (LiR) is used to forecast some indicators which have a roughly linear relation with sales data and this dependency can be estimated in the future.

There is also a review of the existing analytics systems presented in the paper. Listing their advantages, the described systems are:

- Mindtree - is a good tool for managers to review main indicators since it has reporting visualization and to settle plans for future transactions
- Micro Strategy - it has great capability for advanced predictive analytics and advanced data visualization
- Tableau - is one of the most popular data visualization tools which is also able to provide Big Data analytics, including widely used Hadoop based systems

As it was mentioned before, the system described in the paper uses data mining approaches to select data from the database. Then it stores and segments tables for items sales, combination of items, ratings and feedback from customers. The obtained data is processed into graphs which are used to find out insights and patterns in the data such as defining the amount of sales for each product or trends during the time. The system creates recommendations in terms of key profit and sales drivers, namely:

- sales strategy - to decide which products' sales should be increased and to provide portfolio optimization in order to increase overall sales
- price strategy - to define products which have to be discounted or price increased

- stock management - to forecast demand in order to provide more efficient stock creation by products to avoid lost sales

Strategically, our system has the same purposes which are increasing overall sales amount and profit by increasing effectiveness of business processes working. The difference is in the approaches, which are used and specifics of the field of application. For manufacturing company, stock management and sales itself can be not connected directly, since they are implemented as separate supply chain parts and can be independent.

The analysis tool in the paper is realised on a back-end of the clothes selling website, created specially for the research. It took data from sales tables and segments products on fast and slow sold ones. Then in terms of price strategy the price of fast selling products are increased and slow selling ones decreased. It confirms the importance of using segmented approach in contemporary business, as it was mentioned above and applied in this thesis.

The paper also presents the results of survey created to estimate the user experience of their analysis tool and this approach also can be used to get a feedback from our system users and enhance its work.

Concluding, the described system in the paper is useful in a field of e-commerce or some other fields if it has the same indicators or needs. It is able to segment products in order to implement some actions or activities on some of them to increase overall sales and profit. It uses some business reporting tools but not advanced predictive analytics algorithms such as deep neural networks or sales time series prediction models which are able to define covert features and dependencies in data. And there are some differences in manufacturing company's working specifics and analysis needs, so it requires personalized analytics system, ideally created for each company itself.

Another paper which is called "Walmart's Sales Data Analysis - A Big Data Analytics Perspective" [21] also describes the analytics system which was developed within the retail company which is close to the area of application of our system. It is mentioned that nowadays retailers, despite being very traditional business models, are also racing in implementing advanced analytics and using all potential of the data available. And it is also important to provide customers with some promo activities or offers to increase overall sales and reach the goals.

The paper presents the analysis of the big amount of data of Walmart which

includes defining most important parameters influencing on sales, visual representation of the insights with using Scala and Spark Big Data Analytics frameworks. Also there is an option to forecast future sales based on previous years data using Big Data application, described in the paper. Reviewing this paper may be beneficial in terms of using some of the Big Data tools and approaches in finding dependencies among the data in our system.

The most adaptive businesses are trying to understand the willing of customers, to predict it in order to have a competitive advantage. It also includes reviewing your current progress and ability to settle future plans in terms of sales and marketing strategy. It is also mentioned in the paper that the more important is to be able to process the data, analyze it rather than just having it. That is why the authors present their approach of getting insights from Walmart's data using stack of Big Data technologies and data visualization.

Walmart is a retail chain of supermarket or hypermarket like stores, which has a lot of branches in different locations and cities. The main idea of the paper is to find out the most influencing parameters on sales in order to analyze this features and provide recommendation on increasing sales. The parameters can be the level of unemployment, average salary in a locality, holidays or time of the year. All of these factors can impact on the sales itself and market potential in the locality or city, and the similar approach was used in our system where outlets were segmented based on the company's sales amount and the amount of sales of the whole category. The factors listed above are not widely used in our study for outlets, since outlets are traditional independent stores [22], whereas Walmart is a chain and has a wider audience. Independent stores are very local and could be influenced by so many factors, for example, being close to the big store nearby or financial potential of the owner. Though, such global factors as holidays have an impact both on big chain stores as Walmart and independent stores, so it could be useful to reallocate sales or resources on the eve of the holidays. Authors explain that they chose Walmart as an object of analysis, because all big retailers have the same problems and tasks. After becoming a big retailer it is important to sell more and simultaneously reduce the costs. And if there are some sales increases happening flexible companies manage to have enough product or enough staff. It is also about allocation between the shops of the chain, so getting some analytics regarding each shop can be useful for the retailers. The main purpose of the

retailer is to find out customer's needs and possible future behaviour in order to be able to provide effective actions so that customers come back to purchasing at the same retailer's shops.

The described approaches analyze the data from the next points of view:

- different scale - analytics can be applied for the whole chain or for any of shops separately which allows to define best or worst performing teams and to make decisions based on the analysis
- dynamics - it is possible to apply predictive analytics to forecast future weeks sales based on previous sales history in order to align the result with the plans

It was mentioned in the paper, that there was already a study related to sales analysis in Walmart called "Forecast of sales of Walmart store using Big Data application"[23]. This paper described data analytics based on Hadoop, Hive Big Data technologies which were applied on the open Walmart dataset, the same as used by the authors of [21]. Then, R statistics tool was used to apply a standard forecasting model and also Tableau data visualization tool was used to show the patterns in the data and obtained results. The main idea of that paper was to apply Big Data frameworks to be able to act with huge amounts of data and create some meaningful insights based on it.

The authors provided deep research in other works to find out the better choice of Big Data technology to use in their study. A paper called "Leveraging MapReduce with Hadoop for Weather Data Analytics"[24] shows an application of Big Data for forecasting and analysis of weather and temperature. It showed an efficient and accurate analysis for climate area, which is crucial for some other related areas such as agricultural planning, using MapReduce architecture. Another paper [25] also provides weather forecasting but with comparison of MapReduce and Spark technologies. A huge amount of data from different sensors and devices, from other climate parameters was analyzed in order to create accurate and early forecasts for weather. Spark was proved as better performing in that paper. Finally, the paper "A review: MapReduce and Spark for Big Data analysis"[26] directly provides a comparative review of two mentioned technologies. The authors concluded, that it depends on the area of application but mostly Spark is much more efficient in terms of time and memory. Thus, authors of the [21] used

Spark with Hadoop for performing Big Data analytics.

Apache Spark is a framework for dealing with Big Data using distributed calculations on the clusters, and programs can be written in Python or Java. The main data object is Spark Dataframe which is a kind of table with named columns and was used to extract data from Walmart datasets stored in CSV format. Spark can also use Hadoop Distributed File System (HDFS) as an architecture for data management. Spark support the most widely used data types, but for reading CSV files authors used a special library. And after extracting data, transforming it and saving into separate tables Pandas[27] library was used to work with the datasets.

The analysis was made in the paper using 3 years weekly sales dataset for the each store, also using additional features such as temperature, fuel price, unemployment level for the each store and week. In our system it will be also weekly sales used for forecasting by several models. Also, in the paper Spark performance for different stages of analysis was compared with widely used systems for such kind of operations.

The paper provides weekly sales data graph for each of 45 Walmart stores during 3 years. From this graph, it is clearly seen that:

- sales trends are the same among stores, since graphs for each of 45 stores are almost parallel to each other and differ only with the level of sales amount
- sales trends are the same for each year, since sales amount increases and decreases in the same period every year. It stays low for the first quarter, then increase in the second, decrease slightly in the third quarter and highly increase at the end of the each year

As mentioned before, sales data in our research are not similar to the ones presented in the paper regarding Walmart. Mostly, because in our study we consider more local stores when in case of Walmart each store presents macro parameters such as temperature, solvency of the population and so on and this parameters impact on a big stores directly. For example, on figures 2.2 and 2.3 you can see the weekly sales dynamics from our dataset and there is no any stability in trends depending on the time of the year and the behaviour of each store are not similar to each other. That is why we considered weekly sales for forecasting model for each outlet separately.

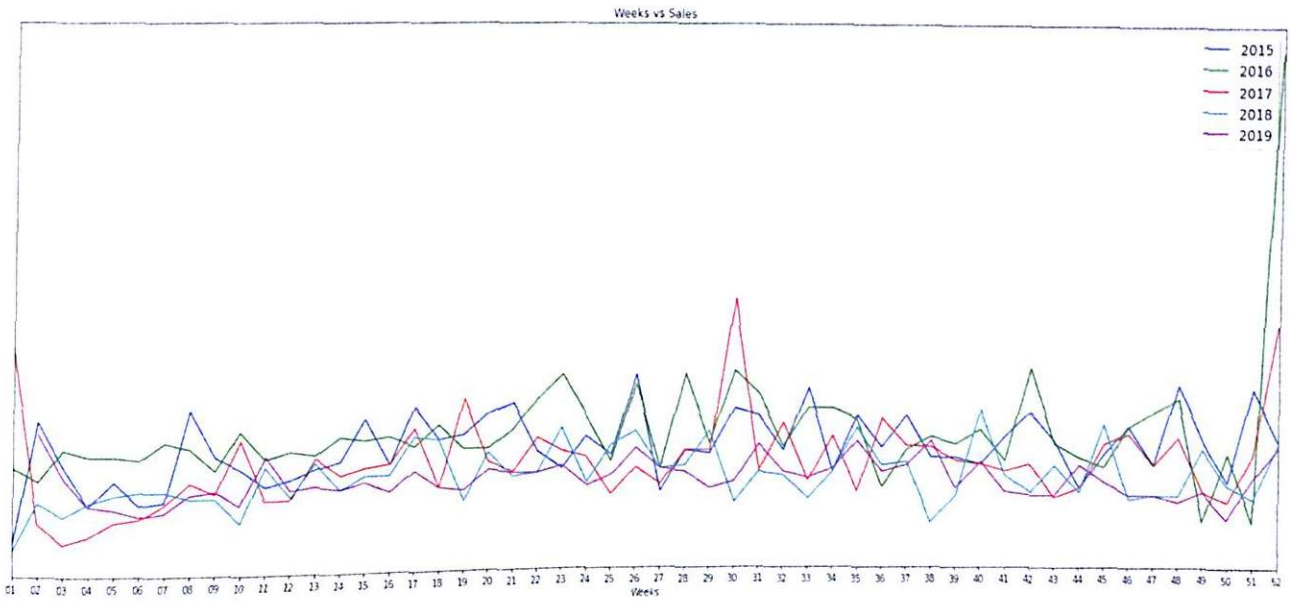

Figure 2.2: Weekly sales in the store A from our study dataset.

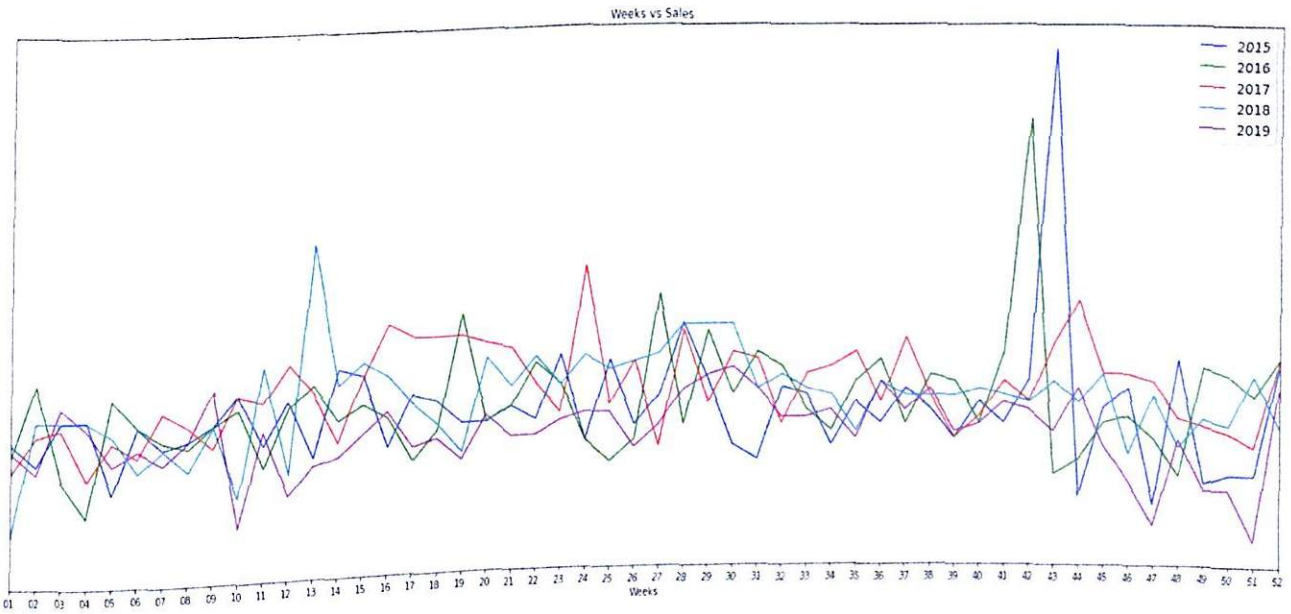

Figure 2.3: Weekly sales in the store B from our study dataset.

From the obtained data also it was concluded that there is a range of fuel prices and outside temperature when sales amount are maximum values amongst other values. So, using Big Data or other analysis tools more and more insights from the sales data, like it were presented in the paper, can be extracted in order to increase business competitiveness which in turn leads to the better customer experience.

2.1.2 Machine learning algorithms' review

The important step of literature review was to study machine learning general types and principles. It will help us to define which ones are the most relevant for segmentation and forecasting objectives during the work.

On Figure 2.4 there are most popular types of algorithms reviewed, which can be used in system development.[5]

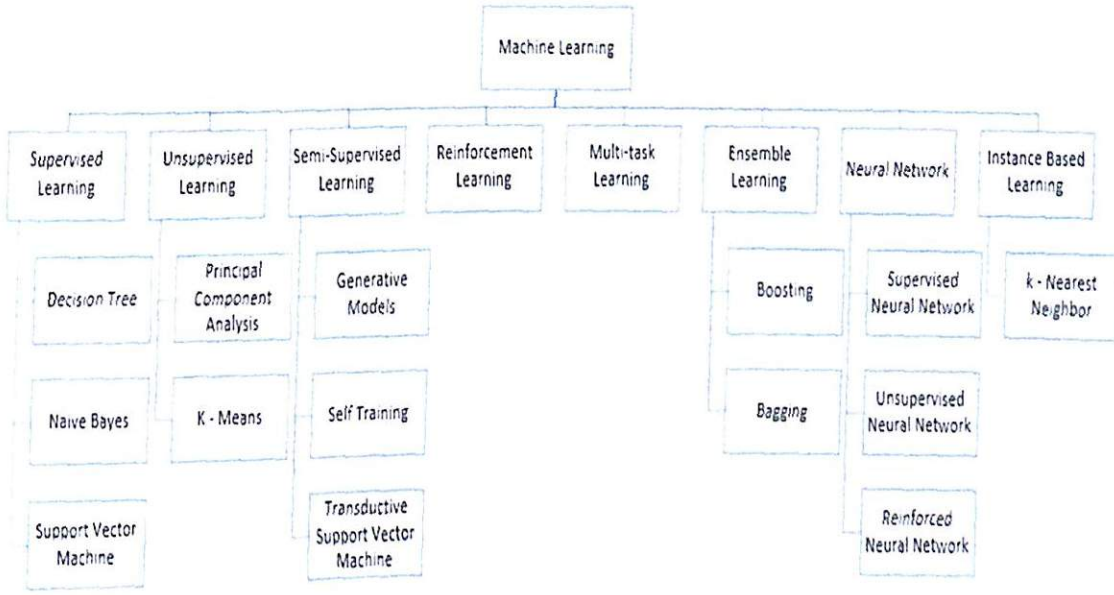

Figure 2.4: Types of learning.

The supervised machine learning algorithms need algorithms' training on truth labels (correct answers) before it can predict or classify outputs. The input dataset can be divided into train and test datasets to train a model with the train dataset and check its performance with the test dataset.

Cross-validation is one of the approaches, when you divide data into k folds and consequentially use each fold as a test set, while others are used for training.[28] All algorithms learn some kind of patterns from the training dataset and apply them to the test dataset for prediction or classification.

If the task is to find an output from continuous set, the algorithm is called regression algorithm and classification, if the output is from discrete set.

Supervised algorithms are used, if there are "truth labels"("correct" answers) for the input. Comparing similarity of the output and correct answers, accuracy of the algorithm can be defined.

Supervised machine learning algorithms can be:

1. Decision Tree: is used mainly for classification purpose. This kind of algorithms means building a tree based on data attributes sorting and using this tree to predict the output for test data.

2. Naive Bayes: it creates trees based on the probability of happening some scenarios and mainly used for classification and clustering problems. It takes one assumption: each parameter of the classified data is considered independently of the other class parameters.

3. Support Vector Machine: works on the principle of margin calculation with increasing the dimension of data. It basically, draw margins between the classes. Builds a hyperplane, i.e. allows you to project your data into a larger dimension space and divide it into 2 classes of data.

Unsupervised learning algorithms do not learn any features from training data, and it is mainly used for clustering new data without any "correct" answers or decreasing dimension of data.

1. K-Means Clustering: self-learning algorithm that builds K clusters from input data, so that at each point in the cluster there are minimum distance to the cluster center.

2. Principal Component Analysis (PCA) – the dimension of the data is reduced to make the computations faster and easier.[29]

Ensemble learning is a combination of various learners to create a new one to solve a particular task. They are:

1. Boosting: combines weak learners in order to create one strong learner. AdaBoost is the most popular. In each round, AdaBoost selects a portion of the training set and checks how accurate each classifier is. As a result, the ensemble consists of one classifier. Next, AdaBoost again tries to identify the best classifier.

2. Bagging: is applied where the accuracy and stability of a machine learning algorithm needs to be increased. It is applicable in classification and regression.

In instance-based learning, learner tries to apply the same pattern to the new data without doing anything with training data.

K-Nearest Neighbor (kNN): is a "lazy" classifier; it begins the classification only when new unmarked data appears. First, he searches for the nearest marked points. Then, using the neighbor classes, kNN decides how best to classify the new data. The decision on whether a new element belongs to a cluster is based on belonging to the majority of neighbors.

The neural networks (NN) are based on a biological concept of neurons. There are a lot of different types of neural networks, such as Multilayer Perceptron (MLP), Concolutional Neural Networks, Recurrent Neural Networks (RNN) an so on. The main idea is to train each of the set of neurons, which is some kind of linear function followed with activation function. Different types of NNs are used for different data and objectives.

Following algorithms were studied deeper [30] and some of them were implemented with using real data and according to the objectives of research.

Algorithm C4.5 / C5.0 and CART algorithm are algorithms based on decision trees. Both of it require training, based on the construction of a decision tree. After building the tree, it makes a decision to classify the test data by passing through the tree. CART algorithm uses evaluation functions to find delimiters when building a binary decision tree. After building the tree, new data is added to the class that corresponds to the leaf, that we got after walking through the tree using the data given.

Apriori algorithm - a self-learning algorithm that uses association rules to study the relationships between database variables. Often it is related to example of "consumer basket" in which it is possible to define the relations between products that are most often bought together. Described algorithms can be used further in the system for searching relations between attributes (Apriori) or classification of data (CART).

Following algorithms were tested with real data to divide outlets for clusters and implemented within the work. K-means finds the optimal centers for the clusters the data to be divided into. EM algorithm is a self-learning algorithm, commonly used for data clustering. Within a statistical model is built — a normal distribution that describes the data as accurately as possible, and then assigns the belonging of points to clusters. These two clustering algorithms were used

and tested to be used in a system during this work.

2.1.3 Forecasting models

There are papers related to sales forecasting systems, however only sales forecasting are described there, not full analytics of sales data. The paper "Hitting your number or not? A robust & intelligent sales forecast system"[31] introduces a sales forecasting system with highly accurate predictions and describes it as a product. Sales forecasting is using past sales data to define future sales or other indicators for the week, month or year.

The paper also describes sales funnel and its stages to give a general understanding of how the process of selling works nowadays. In case of manufacturing company, route to consumer is slightly different. There can be different intermediate stages such as wholesalers, distributors or chains and the topic of selling is the volume rather than the fact of sale. But traditionally sales funnel consists of leads, opportunities and customers stages. Leads are coming from different traffic sources, e.g. internet website, events or personal calls. After expressing interest, lead could become an opportunity which has also different stages of readiness to become a customer. All of the leads (independently if it has reached the next stages) usually are gathered in Customer Relations Management (CRM) system and are detailed with some set of features. Generally, the same process is used by manufacturing companies on the stage of distribution chain building, but our work is more related to volumes increasing based on the specifics of market trends or each outlets capabilities.

Forecast requirements listed there could be used as a benchmark in this paper - forecasts must be:

- accurate - able to learn from many past periods even with outliers in data
- stable - no high volatility within some period of time
- fast - available at any levels of specification and have minimum delay for calculating any of it
- consistent - sum of predictions for subsets of time series gives a prediction for its union, for any level of specification also

- interpretable - logic is understandable by any users

In our system, several models for sales forecasting were reviewed and compared for being used in the system. They are - traditional time-series models, neural nets and decision-tree based ensembling models.

The similar approach was used in [31], where probability models, Opportunity Score Aggregation models, Time Series models, Neural Network models and Hybrid models are combined in the system as a model library. Since, it is described as universal system for any company, with data fed to the system, it estimates the best model from model library in terms of Mean Absolute Percentage Error (MAPE) and implements it. But in our study, since it is for the specific field and company, the best model will be estimated also with MAPE beforehand and only the best model to be realized in the system.

There are the challenges that meet us on the way to developing sales forecast system according to the paper:

- business process change - with changing of data structure such as number of categories, redefining steps of route to consumer it could lead the forecasting method to become inapplicable

- bias to quota - every outlet understands its plans and the result can be directed to achieve this plan, and it can be peak sales at the end of month or vice versa decreasing volumes to leave it for the next month

- human factor - there could appear mistakes on any stage due to manually uploaded data such as importing data from Excel files to the database

- sales pattern change - patterns can change drastically with any outer factors such as changing prices due to the crisis, new products launch, thus making learned patterns from historical data inapplicable

Figure 2.5 shows a deployment diagram from the paper which describes their sales forecasting system. Clients are any devices which user can connect by to create their forecast, fill the data or make some other actions. All of the actions are recorded in CRM which is able to update data in a database on a regular basis with some pre-defined intervals. Data base is stored in Data Store and the Data Store can be connected with the "heart" of the system which is Forecast

Engine through Data Collector. Data Collector is the set of APIs, protocols or database queries. So, when the request comes from the client to Forecast Engine Forecast Engine through the network, Forecast Engine creates database queries to get required data for the forecast in a real-time mode. This architecture could be useful for further developing of our system, but for now this study is focused on developing analytics core which are modeling algorithms itself.

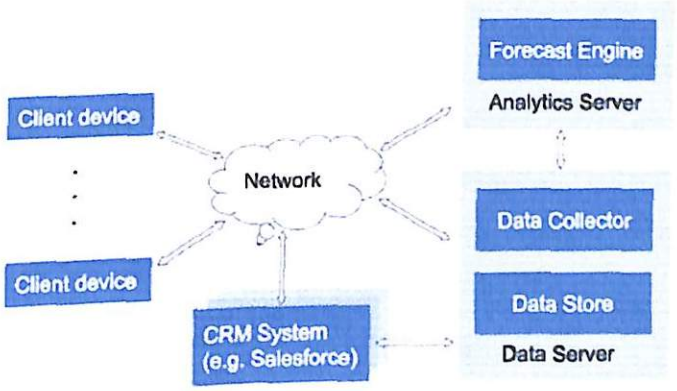

Figure 2.5: Deployment diagram for sales forecasting system.[31]

Forecast Engine contains a model library, which applies all models from the set and use the most accurate for given data. Models from the paper are developed to predict revenue from the opportunities (as a stage of sales funnel) for future quarters and based on revenue and other data from previous quarters. There are:

1. Probability models - after aggregating opportunities conversion rate is used as a probability to calculate revenue. Revenue for each category of products in the current quarter are multiplied with probabilities (of the opportunity to be successfully closed) and summed up. Probability is weighted mean of conversion rates for previous quarters. And the weights can be defined in different ways:

   - just equal - then it is just a mean of previous quarters conversion rates
   - bigger weights for the same quarter in previous years - if the change is seasonal
   - bigger weights for the latest quarters - if there is a persisting trend

   Additionally, there is another term related to the opportunities newly created opportunities during the quarter. This model is easier to implement,

does not need training but could be not consistent.

2. Opportunity Score Aggregation models - here the model first calculates the chance to be closed by the end of quarter for each opportunity. Then multiplying this chances with the price of each opportunity and summing up it receives the forecast revenue. The chance of being closed can be calculated with binary classification algorithms such as Logistic Regression (LR) using parameters of the opportunity as features and training on past quarters opportunities. Also, Long Short-Term Memory (LSTM) can be used instead, if there are some dependencies on opportunities historical data. These models in the system are consistent, but need more resources for training and debugging since it trains a lot of separate machine learning models. The similar modeling is used in our system for predicting plans achievement fact with LR and predicting sales amount per each outlet with LSTM.

3. Time Series models - revenue data from quarter to quarter are considered as time series with some trend or seasonality in it. Then, time series models, which are common for sales forecasting problems are used. In the system from the paper, Seasonal Autoregressive Integrated Moving Average (SARIMA) is implemented if considering quarters or other time ranges as a seasonal ranges. Moreover, there are Simple Linear and Exponential regression models in the library which can give decent results, since many plans for the quarter could be settled by management considering just linear or exponential market growth. Formulas for Linear and Exponential regression models of revenue based on historical revenue values are shown below in 2.1 and 2.2 respectively.

$$\mathbf{y = a * x + b} \tag{2.1}$$

$$\mathbf{y = a * b^x} \tag{2.2}$$

Time series models are also easy to implement and run since training is not required, but it show good results in case of strong patterns in terms of seasonality or trend in data. Also, it may be not consistent.

4. Neural Network models - LSTM is used for predicting the whole revenue

and it is the output of the model and the input is different features related to the whole funnel. LSTM as a neural network is able to capture complex non-linear functions and is a common model used for time series prediction.

5. Hybrid models - it is based on linear or exponential regression models for the opportunities together with any other model listed above and applied for the aggregated data in the current quarter. Linear or exponential regression model is applied for past quarters only, so it give enough accuracy if revenues from past quarters are in a good linear or exponential trend. In the system, determination coefficient[32] is calculated for checking this fact and it has to be greater than some threshold. Then the output of hybrid model is weighted mean from regression model and any model from listed above applied for the current quarter. And the weight for linear or exponential regression model is decreasing from 1 to 0 depending on the day of the current quarter.

Concluding the paper, authors provided results of testing with the data of two organizations, and their system showed 98% and 93% accuracy whereas some other baseline model gave 88% accuracy for both. In terms of volatility revenue forecasts for each day of the quarter did not deviate for more than 2% from the value at the end of quarter which means that the system gives not volatile forecasts. And computational time for forecasting itself was much less than time for querying data from database. So, the robust sales forecasting system was developed which complies with all the forecast requirements listed above. In the future, authors want to implement confidence intervals for forecast values, which also could be useful in our system.

For diving into traditional time series models "Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA process: Theoretical Basis and Empirical Results" [33] was also reviewed. Using SARIMA model high enough accuracy level can be achieved, also MAPE is used there as a measure for forecast accuracy estimation.

As it was already mentioned, one of the most popular methods in forecasting are neural networks. NNs have several advantages, which are ability to learn more complicated dependencies than linear models, flexibility in terms of fine-tuning opportunities, though it requires technical knowledge and experience working with NNs. Paper "A Comparative Study of Machine Learning Frameworks for Demand

Forecasting"[13] provides a review and comparison of several models applied to sales forecast for next 16 days. There are 3 NN approaches described, which are single NN considering 16 days as one variable, 16 models approach, when single NN predicts only one specific day, and 16 model approach with adding categorical features and sequence-to-sequence model as preprocessing step.

Light Gradient Boosting Model(LGBM) also described in [13] is one of the boosting (or ensembling) algorithm based on decision trees. It runs several decision tree based models in order to unite it as one strong model. It achieves the same accuracy as other boosting algorithms such as XGBoost but in better time.

The paper is specifically about demand forecasting and as it was described in the review of analytics systems, accurate demand forecasting can lead to right stock management as well as sales management and finally to increasing profit. It is worthy to mention that perishable items have higher weight in loss function because it is more important to provide right stock management for such products. The main feature of the paper is that the authors provide forecasting at the very low level such as forecasting per each day, each product and at any store. In our study we use weekly sales data because sales per each outlet are not made every day and prediction on a day level is not possible.

Data used in the paper is sales for each store and each item, store data, items, oil prices and holidays since it is the factors which can impact on sales amount. On preprocessing step, windowing was implemented - when time values change from rows to columns and any additional features combined from the initial ones could be added such as mean sales.

These are the models used in the paper described:

- Moving Average Method - it is merely average sales for some number of previous days, corrected with a parameter responsible for seasonality. Quite accurate for short-term forecast, but not for long-term

- Light Gradient Boosting Model(LGBM) - it is ensembling algorithm based on decision trees and gives the same accuracy as other similar algorithms such as XGBoost in less running time

- Factorization Machine - it is general model like SVM, but is better for high-sparse data as it is in the paper. It finds latent factors which are showing interaction between other factors.

- Deep Neural Networks - in the paper it was used with additional preprocessing which are categorical embedding (mapping values to the categories where similar categories have similar mapped values) and sequence to sequence model (uses LSTM to encode sequence with unknown length to high fixed dimension sequence and then another one to decode into another sequence)

In the paper LGBM and NNs achieve same order accuracy and run time, and bigger comparing to traditional methods. And similar approach in comparison of different algorithm types will be used in our system.

The paper called "Egyptian Case Study - Sales forecasting model for automotive section"[34] presents a sales forecasting approach applied to automotive sales in Egypt using macro parameters such as inflation rate, gasoline price, GDP, and sales data such as car prices and previous sales on a monthly basis. The paper describes using NNs, multiple Linear Regressions and Adaptive Neuro-Fuzzy inference system (ANFIS) and it resulted that NNs applied for the problem and data features used gave better accuracy.

Table 2.1: A summary on forecasting literature

Studies	**Motivation**	**Result of the Research**
Hitting your number or not? A robust & intelligent sales forecast system [31]	To get acquainted with existing sales forecasting system and its integration process	Robust forecasting system which chooses the best model in a real-time. There could be different approaches in forecasting aggregated revenue - to forecast it as one parameter or forecast separately for each item and then sum it up. Running time is important in terms of applying for huge amount of data in a real-time mode.

Continued on next page

Table 2.1 – continued from previous page

Studies	Motivation	Result of the Research
Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results [33]	To study traditional time-series methods	SARIMA model gives good accuracy. Can be slow if it needs analysis for each observation separately. Root Mean Square Error(RMSE) and MAPE are common to be used in forecast performance estimation.
A Comparative Study of Machine Learning Frameworks for Demand Forecasting [13]	To get acquainted with the possible time series forecasting methods and its performance	Standard time series model, ensembling or boosting algorithm and Neural Networks were tested for demand forecasting. LGBM and NNs achieve the best accuracy
Egyptian Case Study - Sales forecasting model for automotive section [34]	To see some other application of forecasting models and the possible features to be used	Besides NNs and Linear Regression, Adaptive Neuro-Fuzzy inference system was also tested Besides sales data itself, macro parameters such as GDP, oil prices on some time span can be added to train a model NNs showed the best accuracy

2.2 Preliminaries

Aim is to create a system, which will increase effectiveness of analysis and solving trade marketing issues, using contemporary achievements in machine learning and data science. Though within this thesis, only modeling part is implemented and the system itself could be developed additionally.

At the stage of analyzing the functional requirements for the system, use case diagram was developed for the system:

1. Users can use the system for outlets segmentation
2. Also, users can calculate plans for retail outlets, as well as view and build reports, including carry out "what-if" analysis
3. The administrator is responsible for maintaining the database used by the algorithms and for developing calculation algorithms.
4. The system is connected to the website where the entire interface will be displayed.
5. Associated with a distribution system where calculations can be performed using Hadoop technology
6. Also connected directly to the database or to intermediary FTP server used for data exchange.

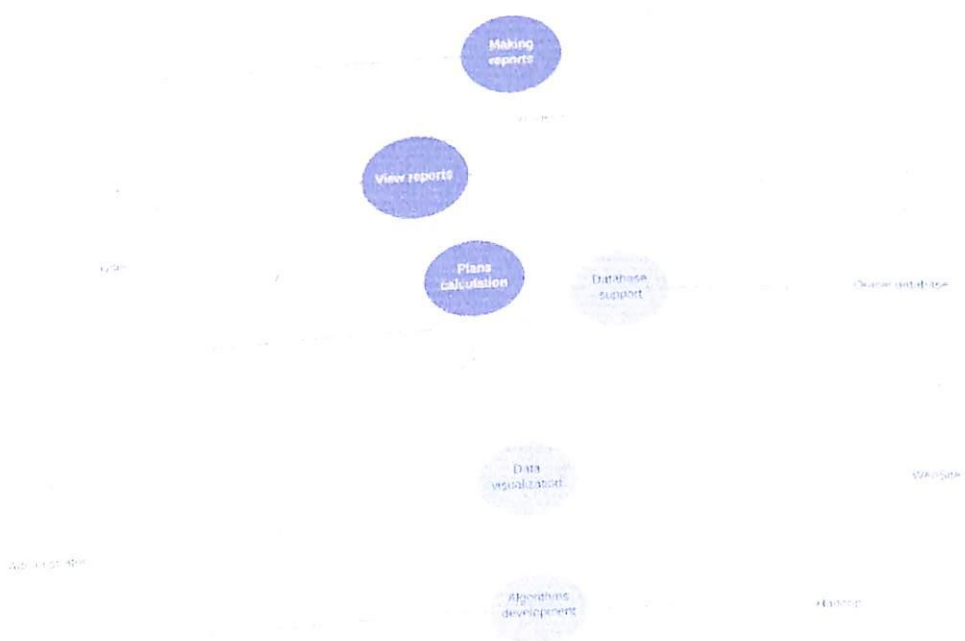

Figure 2.6: Use case diagram for the system.

Software and libraries to be used:

- Python 3.8

- Jupyter Notebook and Google Colab with GPU - environment for coding
- R: possible for the initial identification of statistical patterns or data wrangling
- Pandas, numpy, scikitlearn, scipy, matplotlib — Python libraries for machine learning and visualization
- PyTorch - Python library for deep learning

3. Proposed methods and algorithms

3.1 Stage one - data segmentation

As to recap, during this work we use sales data from Enterprise resource planning (ERP) system, which is available for downloading with any level of granulation, such are sales data on a daily, weekly, monthly basis. In our case, we use average monthly company's sales data. And we also use the data, collected by surveying each outlet on how many items they sell by all brands presented on market. Which also denotes monthly sales data but for all the category.

Using this two parameters for each outlet, a dataset can be divided for some number of clusters, with the aim to provide separate company strategy for each of it. Firstly, we provide data preprocessing, which includes removing extremely high or low sales data.

For the company's sales data there can be some outliers with extremely high sales amount comparing to other outlets. Though we use outlets with the same retailer type for segmentation, some of outlets can be denoted with the wrong type, thus sales data is strongly different from other outlets. The most prevalent types based on their operational structures are independent stores, cash-n-carry stores and retail chains or franchising stores.[22] In this work, only independent stores are considered, because of the highest contribution of this retailers type to the category sales. And since outlets type belonging is defined manually by company employee, its type can be wrongly indicated leading to the sales data outliers.

For detecting outliers we used an approach with box plots logic in it. Box plot is a diagram which is used to plot a distribution of data in dataset. It

shows the values of median, quartiles and outliers separately.[35] Median is the average value, which is also called middle value of the dataset - it means, that in the dataset there are equal number of values, which are higher or lower than the median value. The first quartile (Q1) is the value, which represents the 25th percentile - 25% of values from the dataset are lower than the first quartile, while the third quartile (Q3) represents the 75th percentile.

According to box plot logic, outliers are all data values, which are further than 1.5 times interquartile range (IQR) from the third quartile (for extremely high outlier values) and the first quartile (for extremely low outliers). In other words, outliers are higher than maximum or lower than minimum, where:

$$\mathbf{maximum} = \mathbf{Q3} + 1.5 * \mathbf{IQR} \tag{3.1}$$

$$\mathbf{minimum} = \mathbf{Q1} - 1.5 * \mathbf{IQR} \tag{3.2}$$

After detecting outliers, it were removed from dataset to eliminate wrong influence on other sales data.

For the data, which was collected by surveying it is validated using comparison of company's real sales data and surveyed data. In case of too high deviation, data was also removed from the dataset. Some special cases were clarified manually with the interviewers.

The proposed method for stage 1 is to make two-step segmentation, including:

1. Using selected unsupervised clustering algorithm to create 3 main clusters
2. Predicting whether outlet will achieve monthly sales plan based on additional features and patterns learnt from data

A similar approach was used in [36], where students prediction problem was solved by partition of students and using specific algorithms or parameters for each part based on students' previous performance and behavior.

Previously, clustering was made without any special methods and clusters were formed using some agreed threshold values for outlets sales amount.

Data used at step 1 are company's sales amount (from Oracle database) and estimated sales of total industry for each outlet gained during Census. Following

algorithms were tested with real data and results were compared within step 1. Using this data, outlets can be divided into such clusters as:

- high – outlets with high sales amount, where company should hold their position
- potential – outlets with lower sales amount, but with a big potential to grow. It is about low company's sales and high amount of the whole industry sales
- low – outlets with low sales amount, where company should attend or invest less

On Figure 3.1 the main idea of the clusters described above is illustrated.

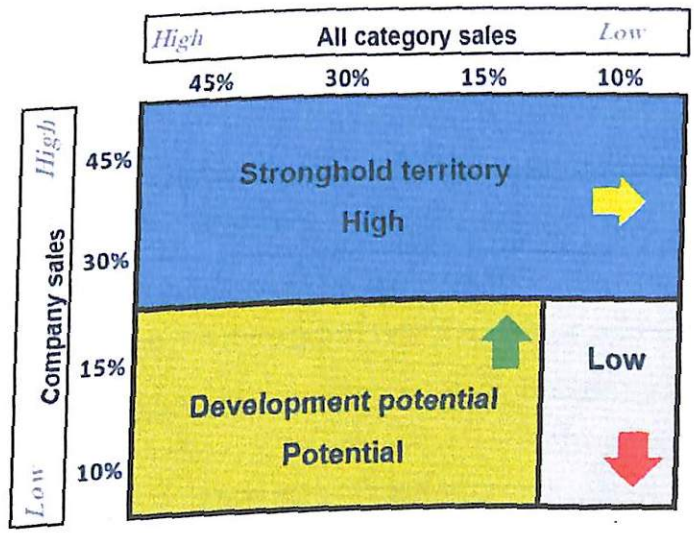

Figure 3.1: Outlets clusters

K-means finds the centers of the cluster, so that the distances from each point in the cluster to the center are minimal at each iteration. It stops, when cluster centers have not changed at the iteration. Susceptible to the choice of initial centers and noises in the data [37]. Three initial centers according to the business needs of the segmentation were chosen manually – high sales amount outlet, outlet with the potential and low sales amount outlet.

GM Model is a self-learning algorithm, commonly used for data clustering. Within it a statistical model is built — a normal distribution, that describes the data as accurately as possible, and then assigns the belonging of points to clusters. It is worth noting, that data has little amount of noises and such outlets were dropped from the dataset. They can be classified after the main segmentation using K-nearest neighbors (KNN) algorithm.

For step 2 also variation of company's sales for outlet, outlet's past plans' achievement for the whole year and last 3 months were used as additional features. Sales plan achievement for current month was used as ground truth labels.

Then data was scaled with Standard Scaler to increase the accuracy and divided to train, test and validation sets. Proportion was 80% – train validation, 20% - test, where train validation was divided as 70% for training and 30% for cross-validation as it is the most common way for split. According to the purpose, accuracy was reviewed, and which is more important - precision, recall and F1 (harmonic mean of precision and recall) scores to track how many of positive predicted outlets were actual positive (precision) and how many of actual positive were predicted correctly (recall), as the model can be used for real business purpose.

Then Logistic Regression model and Neural Network were used to predict plans achievement and scores were compared. Similar approach was described for students' performance prediction in [38]. Multilayer Perceptron (MLP) Neural Network topology was used, since it is usually used for static data and for data, which is not too large for using complex topologies.

Also, it is important, that Polynomial features were added in LR model, since the data has non-linear dependencies to increase the accuracy [39]. Using cross validation method, the best polynomial power, regularization parameters for LR [40] and number of hidden layers and neurons, activation functions for NN [41],[42],[43] were defined.

3.2 Stage two - forecasting machine learning models

At the stage of sales prediction amount of sales on a weekly basis was used per each outlet. Data for last 5 years at least should be used to be able to capture seasonality and trend. Pandas library is used to obtain data, combine it to the dataframe and make some preprocessing, which is described below. A goal is to forecast weekly sales amount based on historical data. After, as an additional application, summing up weekly sales amount, we can obtain monthly sales and check if it is greater than settled plan for outlet. Or use predicted monthly sales amount as a baseline for settling plans itself.

In this study we will consider sales data for one specific outlet only, so forecasting will base on its own sales patterns. In order to compare algorithms performance, a random outlet was chosen without lose of generality. As a potential improvement, algorithms can be tested with all other outlets and predictions compared with real sales data. For all models in this study data was split on training, validation and testing set. Split of data for time series should be implemented not randomly, but handling a structure. Thus, validation set is sequenced part of data which is going consequently after training data, and test data is a sequence at the end of the whole time series.

Before applying any algorithms data will be explored using statistical tools which is actually obligatory for using time series models such as Autoregressive Integrated Moving Average (ARIMA) or Seasonal ARIMA (SARIMA). All of these statistics functions or time series models are available in R, but in this study we will use Python library called statsmodels[44].

Using histogram plot, data was checked for normality of distribution. Figure 3.2 shows the distribution of weekly sales data in the outlet during 5 years. We can observe from the histogram that it looks similar to normal distribution and also there is one outlier with too big value of sales amount. Indeed, Figure 3.3 shows that there is an abnormal sales value at the end of 2016 and overall it is seasonal increase at the end of the year.

For checking normality of distribution, also Jarque–Bera test can be applied which checks the null hypothesis of the distribution to be normal based on comparison of skew and kurtosis with the ones of normal distribution.[45]. Test represents results as p-value, which confirms the null hypothesis of the distribution being normal, if p-value is less than some fixed value. For our data, p-value confirmed that the distribution of data is normal. Normal distribution of data implies that there will be no abnormal model behaviour since there is no skewness and imbalance in data.

In time series modeling the important property of the time series is stationarity. Stationarity is understood as the property of the process to have its statistical characteristics constant over time, such as mean, dispersion and covariance function. Preserving the mean constant is there is no trend in time series, dispersion is about the deviations of data values and constant covariance function means during whole time span data values do not beceom closer or further to each other[46].

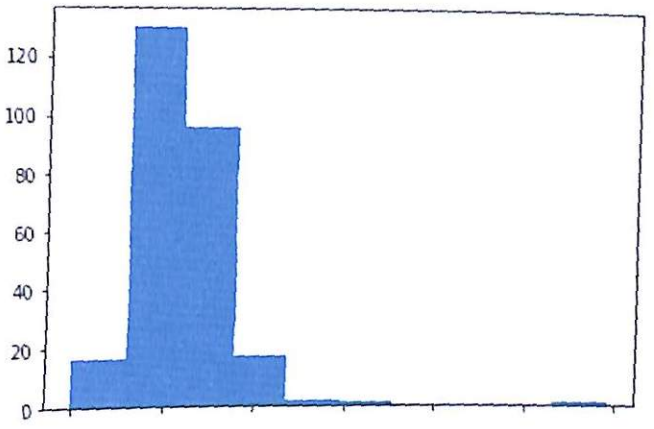

Figure 3.2: Histogram of outlet sales data samples

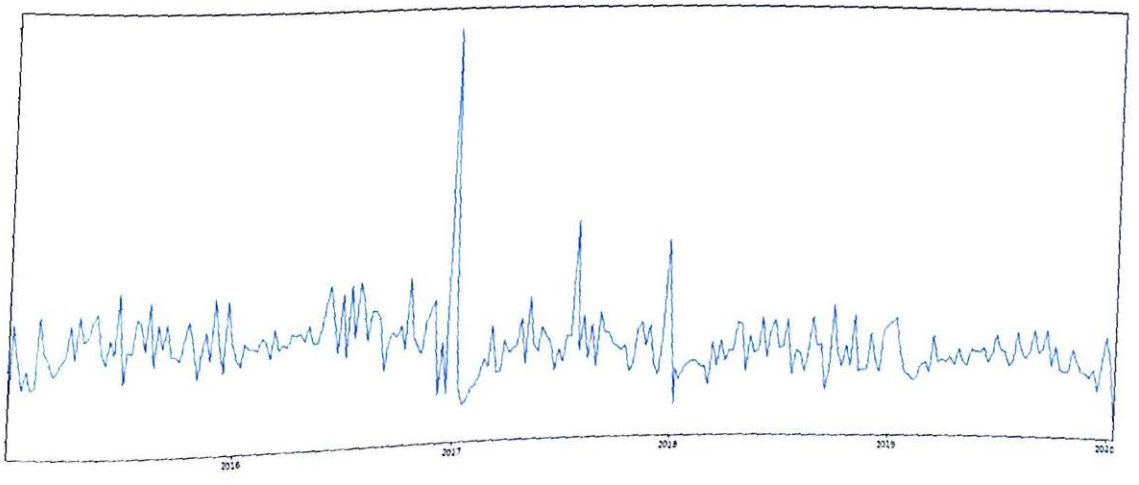

Figure 3.3: Weekly sales in the outlet

Stationarity is so important because constant statistical measures allow to make predictions for the future data since it will not change with time.

As it is seen from Figure 3.3, our sales data has no trend, has constant covariance function and pretty constant dispersion except few abnormal values. There is Dickey–Fuller test which defines the stationarity of time series based on the null hypothesis of unit root presence.[46] So, if p-value is less than some fixed value (usually, 0.05) then time series are stationary and the null hypothesis is rejected. Also, auto-correlation function (ACF) and partial auto-correlation function (PACF) plots are usually used to explore the structure of time series as well as to define coefficients for ARIMA/SARIMA models. Dickey–Fuller test for our data confirmed the stationarity by rejecting the null hypothesis, but ACF plot has too many significant lags. So, we will use the first difference (lag the data for one step and find a difference) to deal with stationarity. Differences of various orders, getting rid of the trend or seasonality, smoothing, Box-Cox algorithm are the methods of dealing with non-stationarity. More detailed information is given

in [46], and is recommended to look through, since it is not the main topic of this study.

After, difference time series should be also tested for stationarity, and the order of the first stationary time series (called order of integration) is used as a parameter d for ARIMA/SARIMA models. In our case first difference time series is stationary and in Figure 3.5 shows the difference time series itself and ACF, PACF plots for it.

The first model to be tested in this study is ARIMA, which is an integrated combination of Autoregressive (AR) model and Mean Average(MA) model. It has 3 parameters which are:

- p - the order of AR model
- q - the order of MA model
- d - the order of integration discussed above

The orders of AR and MA can be obtained from ACF and PACF plots. Parameter p is the largest index of significant lag from PACF plot, whereas q can be defined as number of significant lags from ACF. Thus, ARIMA model with parameters 5, 1, 1 as p, q, d was trained and tested on our sales data.

Figure 3.4 shows the result of prediction of a trained ARIMA model on test data and it is seen that the model did not learn all dependencies well from data and will not be considered for comparison in this study. As it was mentioned, data has some seasonal patterns and using SARIMA instead can be an appropriate way.

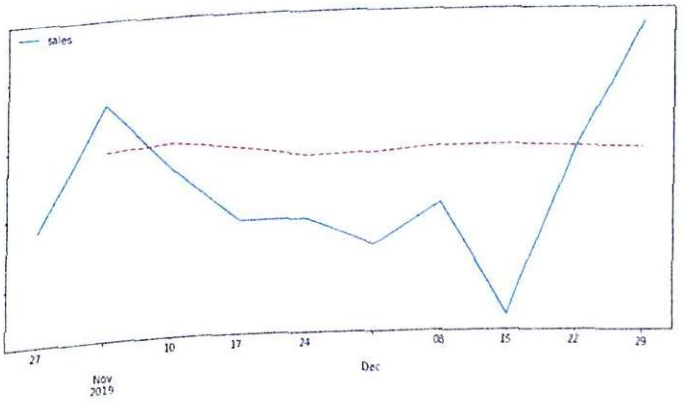

Figure 3.4: ARIMA model prediction on test data

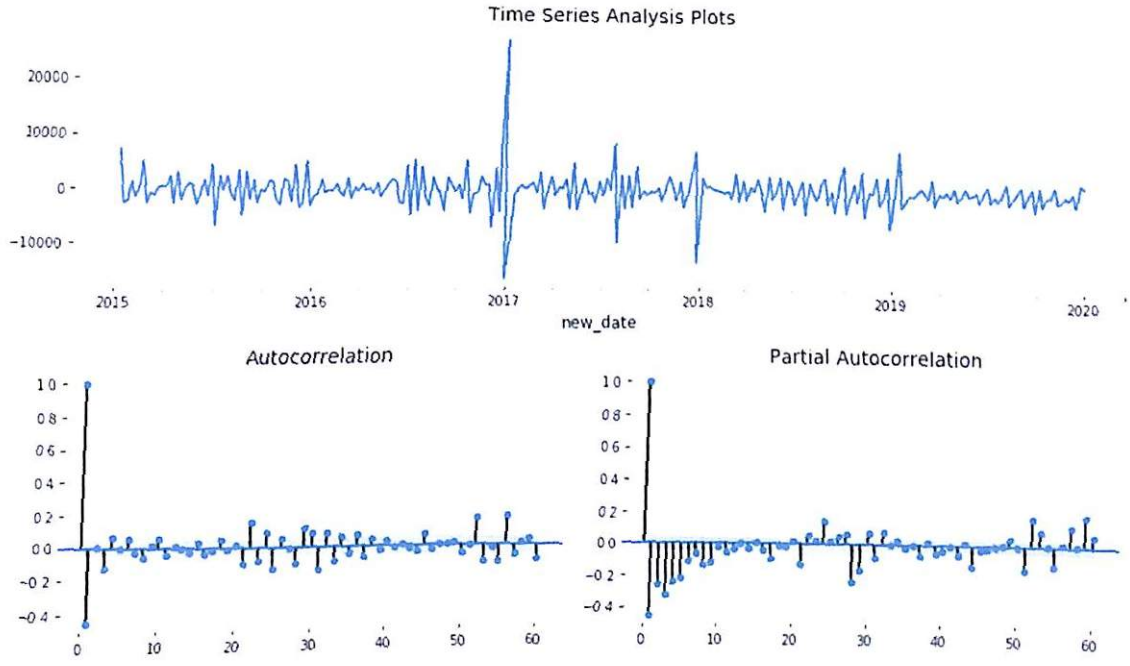

Figure 3.5: First difference time series for the outlet

SARIMA is similar to ARIMA model but with some additional improvements. It also has parameters which are responsible for seasonality. And the preprocessing is similar. The main difference is that for SARIMA usually seasonal difference is made to achieve stationarity and the integration order of seasonally differenced time series is used as an additional parameter in the model. Overall, SARIMA has 3 additional parameters:

- P - the order of seasonal AR model
- Q - the order of seasonal MA model
- D - the order of integration for seasonally differenced time series

For SARIMA we also added Box-Cox transformation to stabilize dispersion applied to the initial dataset. Then used 52 (number of weeks in one year, since it looks like having one year seasonality) as seasonal difference and get one lag difference from the obtained time series. And finally, resulted time series was checked and confirmed as stationary, thus D is equal to 1. Then, using ACF and PACF for resulted time series we can obtain most probable values of other parameters p,q,P,Q. We will also make a brute force search of best model parameters by training the model using different values for its parameters and comparing the measure called Akaike information criterion (AIC)[47] score which is specifically used for comparing time series models.

After having best parameters, we use it to train a SARIMA model. The important part of checking the performance of ARIMA and SARIMA (or any time series) models is to check the stationarity of residuals which is difference between actual and predicted data. After training the model, SARIMAX model from statsmodel library already has built-in resid function to get residuals dataset. After checking it, stationarity was confirmed, thus the difference between predicted and actual values can be considered as random.

As it was described, SARIMA may need too much additional preprocessing steps and checking for stationarity which may be inappropriate in case of prediction for many outlets separately in the analytics system.

The other set of algorithms effective in times series forecasting is Gradient Boosting Machines (GBM) which implies boosting the implementation of classic gradient descent. Basically, the logic of such algorithms is to combine a set of weak algorithms as one strong ensemble which will be able to detect strong dependencies. It is similar to classic machine learning algorithm in terms of minimizing a loss function that penalizes errors in finding a function as a linear combination of other functions within function set. Such kind of algorithms became popular for solving machine learning contests and giving best accuracy results. [48]

XGBoost is an extremely effective implementation of the classic GBM. It can be an ensemble of decision tree based or linear algorithms. In this study we will xgboost library in Python.

The first step in applying XGBoost will be to add some additional features such as month, year, week of the year, week of the month and also lag features (obtained with shifting the time series values with some lag). Any of these features can be useful for the ensembling model. If there were categorical features in our dataset it would be useful to apply categorical embeddings. After testing of different number of lags it was decided to add only lag 1 feature. Figure 3.6 obtained after training the model shows that the most important features are lag 1 feature and week of the year. Also it was decided to use decision tree based version of the algorithm.

Cross-validation for defining the best number of trees in the ensemble, eta parameter which is responsible for regularization and avoiding overfitting and maximum tree depth was implemented on validation dataset. Using stratified k-fold validation was used instead of simple k-fold since it gave better accuracy score.

Stratification in terms of cross-validation for regression problems is described in [49].

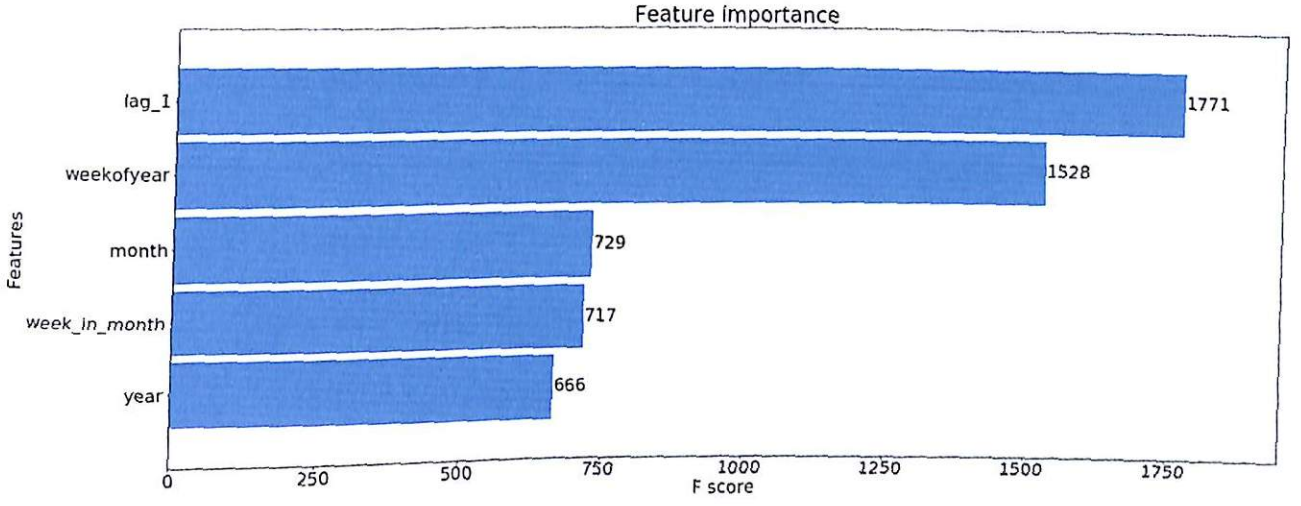

Figure 3.6: Feature importance for XGBoost algorithm

Another approach to be applied for our sales data is deep neural networks. Long Short Term Memory (LSTM) model is Recurrental Neural Networks (RNN) which main feature is remembering some state during processing even a long sequence. RNNs consist of hidden units which are looped with other hidden units in a sequence, where units also take the hidden state input from previous step besides input data. LSTM is a specific of RNNs, which together with the hidden state also inputs such called cell state which contains information to be remembered from some previous steps. For example, it can be useful in prediction for time series, when model bases on the previous value and also on some old value which could be responsible for seasonality.

LSTM in this study was implemented using Python Pytorch library which is used for deep learning tasks and allows to build neural networks on a low level with manual description of all components. As a preprocessing, for LSTM model, the actions listed below were done:

- create a difference dataset - in order to provide stationarity for better prediction, data was shifted with the lag 1 and substracted from initial data. After prediction, inverse transform will be applied to get initial dataset

- transforming to supervised learning task view - adding a lag 1 feature such as dataset represents a set of pairs of consequent sales amount values

- splitting to training, validation and test data - as it was described above and the same as other models
- scaling data - scaling in order to improve forecasting, fitted only on training data and applied for validation and test data
- creating sequences - data was split on a set of sequences consisting of data values resulted from previous steps and the next week data value as a label for each data value. I.e. it is the list of 20 sales amount values and 20 labels which are actually the value of the next week
- batching - sequences were split into batches to feed the network with mini-batches which is more effective in optimization process

LSTM in our study consists of 2 layers for learning more deep dependencies, and has dropout regularization layer to decrease overfitting. Validation was also made for finding the best learning rate.

Then, traditional time series SARIMA model, LSTM neural net and decision-tree based ensembling XGBoost model were used for prediction and the performance compared in terms of MAPE and running time.

4. Results discussion

4.1 Clustering results

Below is the result of division into 3 clusters by rationing and using the Gaussian mixture algorithm. It was divided to 3 clusters because of business needs, which is related to company's strategy implementation in every outlet.

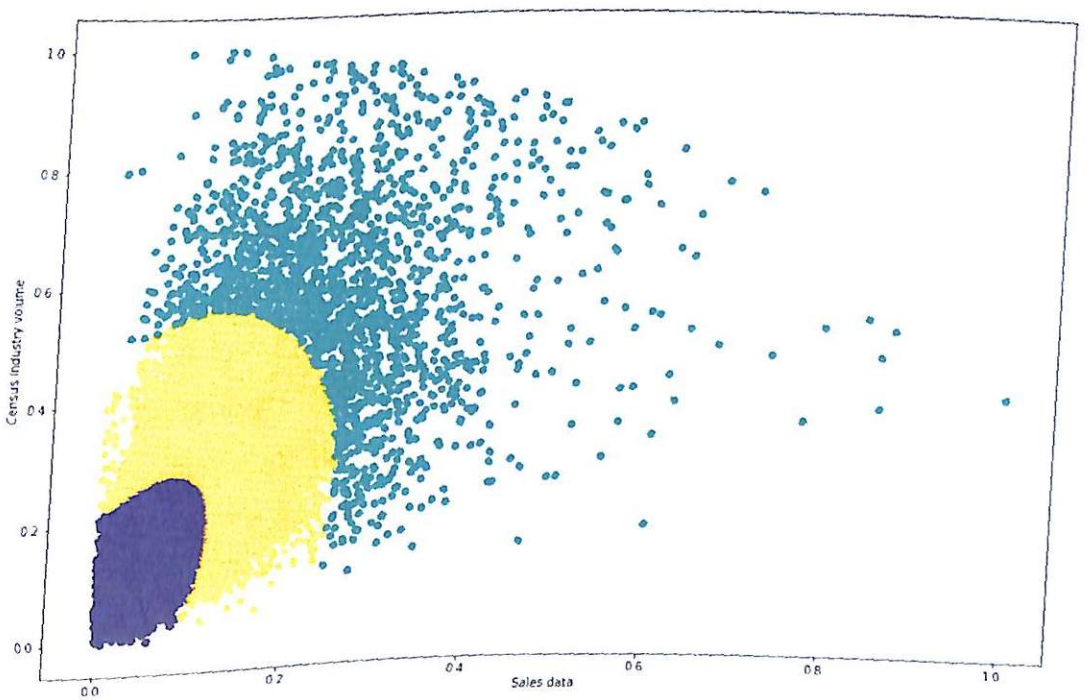

Figure 4.1: GM Model clustering results

Since all clustering algorithms results can be estimated only visually, after comparing the results of clustered data by two algorithms, GM Model algorithm was chosen as more corresponding to business means of clusters, where every cluster presents company's strategy of in-vestments in every outlet. Figure 4.1

shows the obtained clusters which were chosen because the second cluster (yellow-colored) is more corresponding to outlets, which were mentioned above as "potential".

Also, the Silhouette Coefficient (SC) [50] was used to estimate how well the both algorithms clustered the dataset. It is calculated using the mean distance inside the cluster a and the mean nearest-cluster distance b for each outlet by the following:

$$\mathbf{SC} = (\mathbf{b} - \mathbf{a}) / \max(\mathbf{b}, \mathbf{a}) \tag{4.1}$$

The best value is 1 and the worst is -1. The results are shown in Table 4.1, The scores are similar, but K-means was not chosen despite better scores, because the K-means was implemented using 2 Python libraries – scipy and sklearn. All obtained clusters had unsuitable form as mentioned above.

Table 4.1: SC for K-means and GM Model

K-means scipy	**K-means sklearn**	**GM Model**
0.457	0.465	0.416

As an additional estimation of clustering quality Figure 4.2 shows the box-plot with distribution of outlets sales data by clusters formed with GM Model algorithm. It shows, that obtained clusters fit to the needs as described above – clusters differ with the amount of sales.

During the next step LR and NN were used to predict sales plan achievement for lower cluster as it can be clustered additionally.

As it was mentioned above, polynomial features were added to the data before fitting to LR model and to define best polynomial power, training accuracy for different powers was plotted. Figure 4.3 shows, that 5 is the best choice as a global maximum of training accuracy. Feature engineering, including selection of most important features was also made to trade-off accuracy and computational speed.

After using cross validation and defining best hyperparameters for the model, learning curve was plotted to estimate if the model underfits or overfits the data. Figure 4.4 shows, that with increasing number of samples training and validation accuracy scores converge and are not too close. It means model is not underfitting nor overfitting. [51]

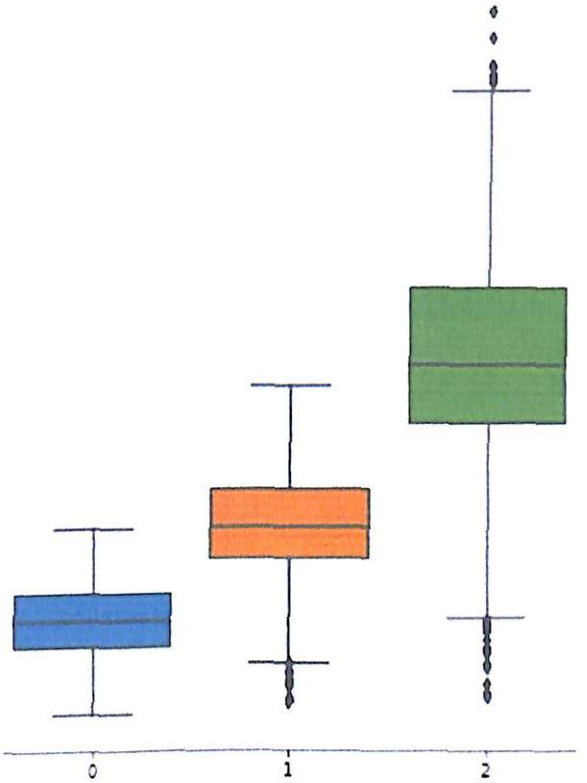

Figure 4.2: Distribution of outlets sales data by clusters

For NN using Multilayer Perceptron the most important is to define the number of hidden layers and number of neurons in each. For this work Growing Method [6] was used, according to which model begins with small number of layers and neurons, then it increases and obtained results are compared to choose the best.

Most of problems and datasets does not need more than 2 hidden layers, and the number of hidden neurons should be less than twice the size of the input layer [52]. In Table 4.2 best scores after hyperparameters tuning for testing data using LR and NN are shown. NN gives slightly better accuracy and precision, but LR gives better recall. Precision and recall are very important in terms of business needs – trying to predict sales plan achievement with aim to focus efforts on such outlets.

Table 4.2: Testing scores for LR and NN.

Metric	Logistic Regression	MLP Neural Network
Accuracy	71%	**74%**
Precision	57%	**62%**
Recall	**84%**	71%
F1	**68%**	66%

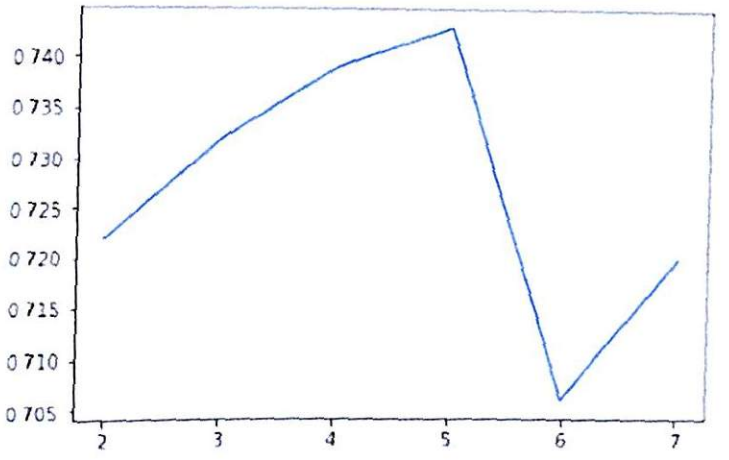

Figure 4.3: Training accuracy of LR model depending on polynomial power

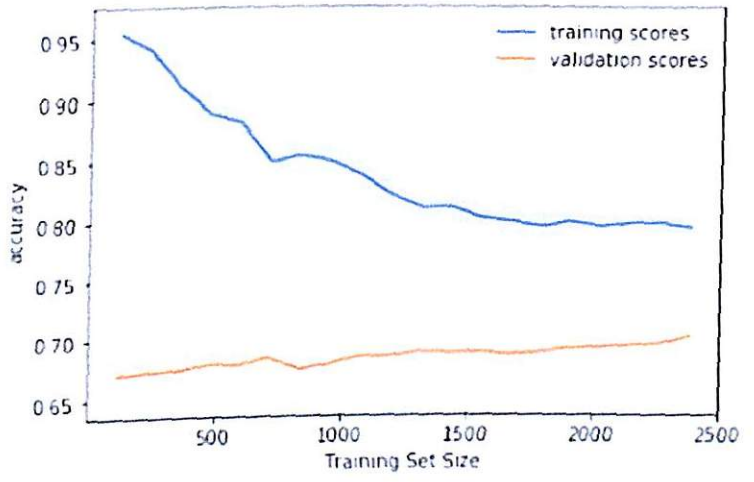

Figure 4.4: Learning curve for LR model

Accuracy scores are like other works, using NN for predicting students' performance [6], [38] and show approximate potential of these models in terms of predicting plan achievement.

4.2 Sales Forecasting results

During the research we could face with some accuracy limitations due to data structure, its sparsity or insufficiency. But the applicability of the forecasting models will be checked with real outlets and improved with more data and many outlet approaches to increase an accuracy in case of necessity.

At the stage of sales prediction amount of sales on a weekly basis was used per each outlet. Data was used for each week from 2015 till the end of 2019.

Then, as it was mentioned, traditional time-series model SARIMA, neural

nets and decision-tree based ensembling model XGBoost were used for prediction and the performance compared in terms of MAPE and running time. Each of the model was tested on last 7 weeks of 2019 (excluding last 2 weeks of December).

As it was mentioned, SARIMA model was trained with the parameters which were found by estimating AIC on validation phase. It was able to learn patterns from training data and for test data also could predict values quite close to actual ones and recognize trends. Figure 4.5 shows the result of prediction of sales amount on test data which are 7 weeks values of sales to the outlet. Root Mean Squared Error (RMSE) was used as a loss function in all 3 models.

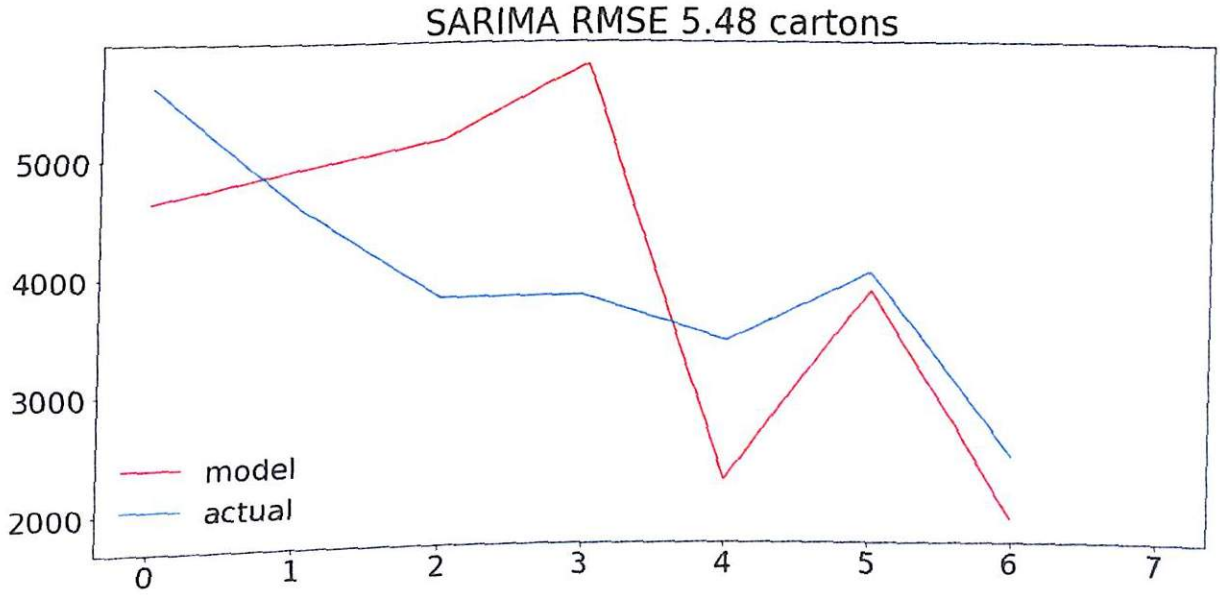

Figure 4.5: SARIMA model prediction on test data.

XGBoost cross-validation was stopped after 200 epochs (or trees) because after it was beginning to overfit and validation set error starting to increase. Stratified k-fold cross-validation was implemented on validation data. Testing and training loss was converging as it is shown in Figure 4.6.

After training with using the best parameters from cross-validation, XGBoost was able to learn sales trends well and give a better prediction than SARIMA on test data for 7 weeks. See Figure 4.7.

LSTM model was also checked on validation data and it achieved quite low error on training and validations sets both. Training of deep neural networks is conducted for some number of epochs, where each epoch is the whole process of training applied for the training data. Though the training loss stopped to decrease abruptly after 20 epochs, it was necessarily slowed down due to regular-

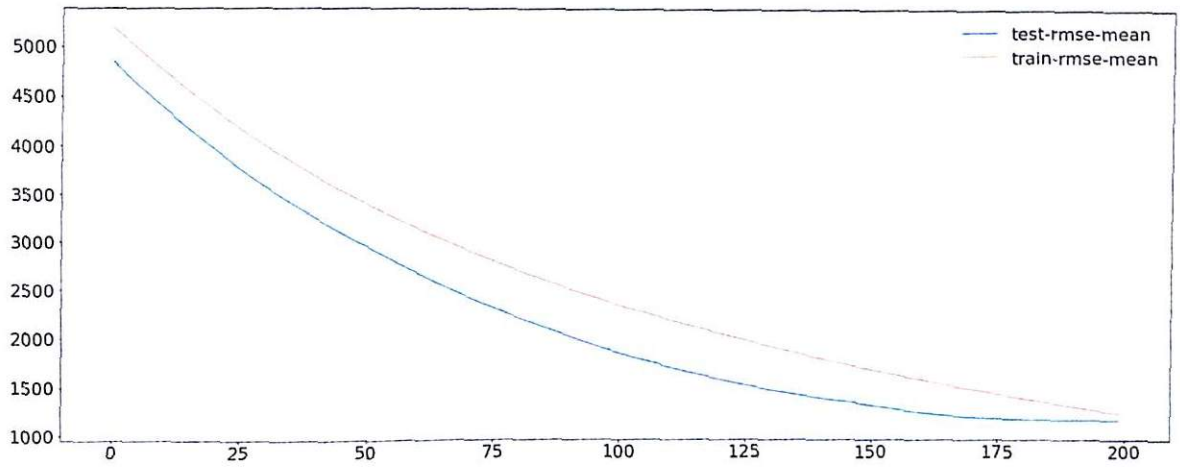

Figure 4.6: Validation error of XGBoost.

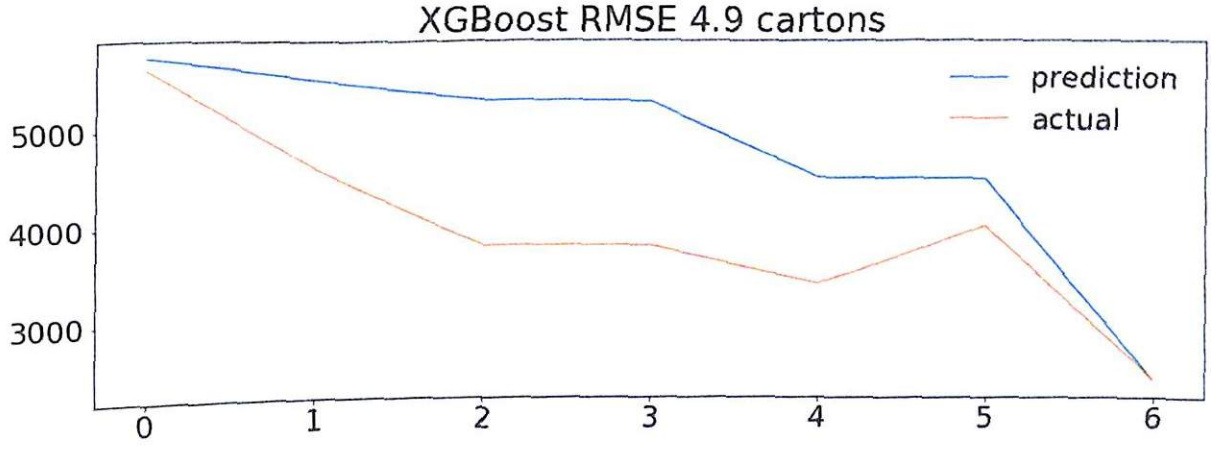

Figure 4.7: XGBoost model prediction on test data.

ization, as it is shown in Figure 4.8. Without regularization model start to overfit strongly with validation loss going up while training loss decreasing. Training loss is higher than validation loss because regularization is not applied for validation and only applied on a training stage. It is also worthy to mention that training loss is estimated during the epoch, whereas validation loss is estimated after the training epoch.

Figure 4.9 shows the prediction built by LSTM on the same 7 weeks. And it seen that LSTM has the lowest RMSE amongst all models, it was able to predict the values close to the actual value which can be useful for the application of the models.

LSTM and XGBoost models achieved the same order accuracy in terms of RMSE, though LSTM was slight better. But XGBoost was better at predicting trends or values change directions. It can be beneficial to add some features to LSTM model, which will be responsible for trends detection. The main difference

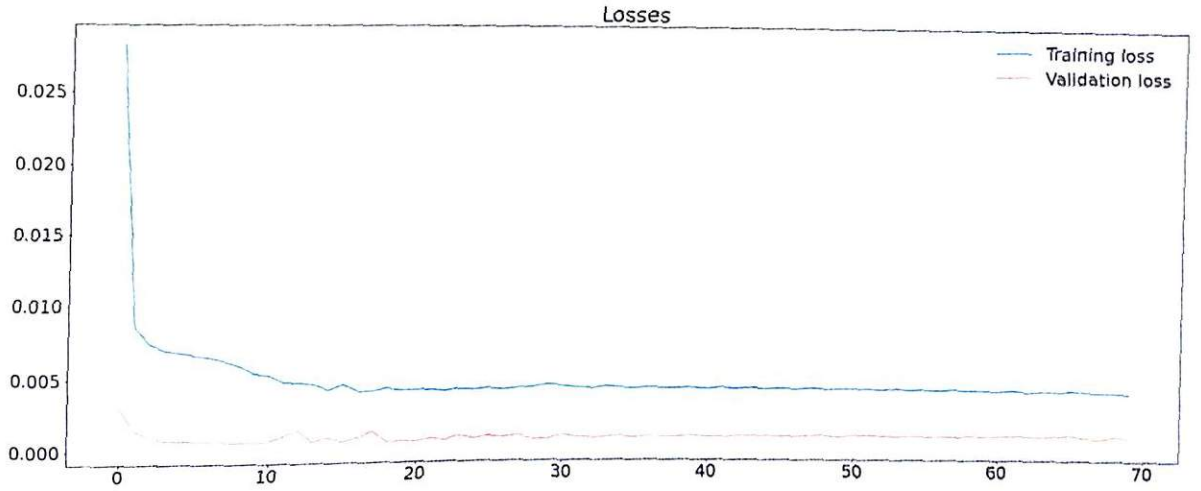

Figure 4.8: Validation error of LSTM.

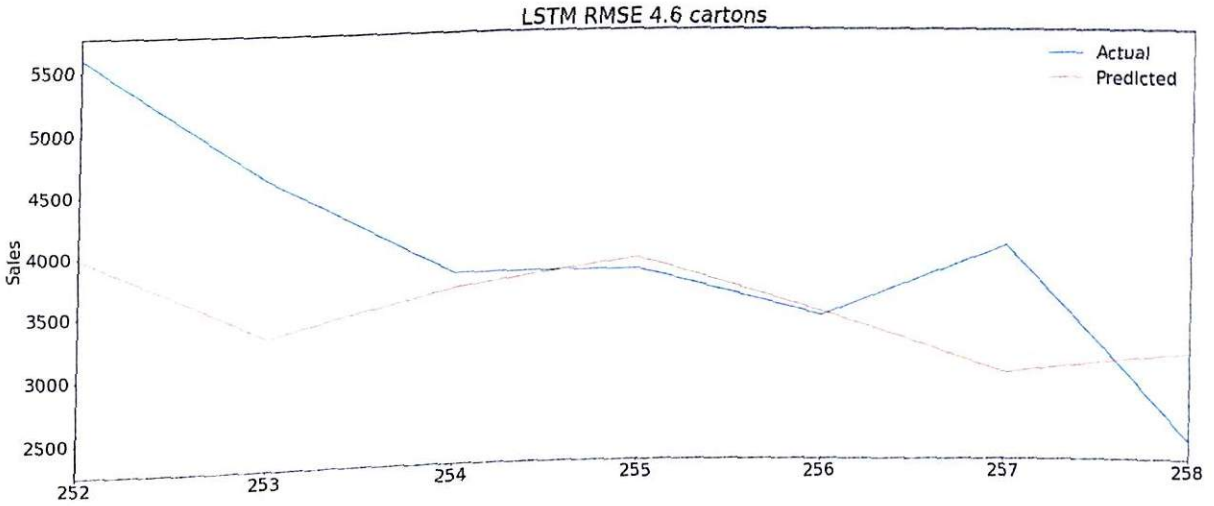

Figure 4.9: LSTM model prediction on test data.

is that LSTM was learning from sequences of consequent pairs of values, whereas XGBoost was learning on each value sample with respect to its week number, month and other parameters. The length of sequence is also a hyper-parameter which can improve the model accuracy.

In Figure 4.10 comparison of the models in terms of MAPE is plotted. It shows, that LSTM has the least error within observed models with a quite good 17% on test set. Which is comparable with similar works described in [31] and [13]

Though in Figure 4.11 it is shown, that LSTM is slower than XGBoost, it both have the same order comparing to SARIMA which need 1 second to predict the data for 7 weeks and in terms of thousands of outlets will not be applicable. LSTM is 6 times slower than XGBoost but it has to be mentioned that prediction for each of 7 values was made in a loop manually adding each next real sales value

from test data as the last item in sequence for each iteration. Running time can be decreased using some other approaches with predicting for the whole sequence at once, so XGBoost and LSTM indeed can be considered as having the same order running time.

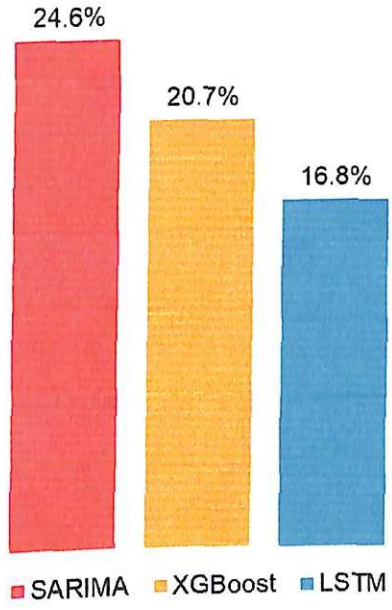

Figure 4.10: Comparison of predicting models, MAPE, %.

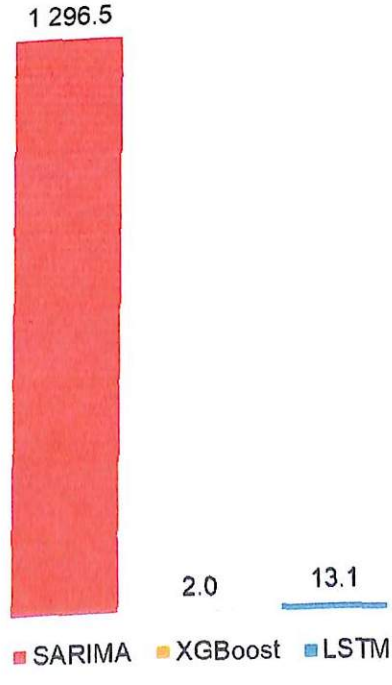

Figure 4.11: Models running time, ms.

5. Conclusion

Clustering results satisfy business needs and allow to use certain investments strategy for each outlet. Predicting sales plan achievement gave enough accuracy results, like other works solving such problem. But since predictions will lead to business decisions corresponding to it – it is very important to get higher precision and recall. Using these models if they predict outlets as achieving plan, only around 60% will indeed achieve it, and only 84% of plan achieved outlets will be predicted correct. Also, considering that only 30% of the dataset have positive labels, such result is not so valuable. Benchmark for this problem is >90% accuracy and F1. Main advantage of predicting model is that it is not overfitting or underfitting. Its accuracy can be increased with creating ensemble classification model or using additional features in the dataset. Proposed clustering method is already used and its second step with plan achievement prediction will be improved.

LSTM model is the best option to use with our data and for improving prediction accuracy additional features can be added to the model.

Next, the web part will be created to create a complete system in which different users of the interface will have different levels of access and it will be possible to perform the necessary actions with the data. In addition, it can be completed with distributed calculations for historical data, which could increase calculation effectiveness.

To use additional segmentation and create a cluster of outlets based on its sales time series transformed as for classification problem, and using some metric, then to train our LSTM model with all sequences from all outlets in the cluster can increase the accuracy of sales forecasting instead of forecasting for each outlet separately.

Future plans are to check up forecasting values for all outlets with settled

plans and facts itself and to estimate whether it is applicable for regular use. It could be also used for forecasting amounts for some territory aggregating amounts per each outlet. Also, forecast values can be compared with plans settled for the outlet to find out whether it will achieve it which was the objective of stage 1.

Implementing confidence intervals calculations for the forecasts can be also useful since there can be some distortions in forecasts and creating intervals will give more information for decisions.

Also it is possible to add visualization to our system, but the most effective way is to integrate with existing VA systems which support the most common types of visualization since it is not the main purpose of our system.

After, to integrate with existing sales reporting system in order to implement dynamic digital transformation and increase the effectiveness of the solution. It allows to define anomalies in sales or predict in a much smaller time since there is no need of data export and manual data handling. It also depends on the ability of the existing system to integrate with such solutions and this fact should be checked beforehand. In terms of integration, intermediate data server might be necessary to export data from sales system and to be accessed with the analytics system.

References

[1] L. Columbus, "Ten Ways Big Data Is Revolutionizing Supply Chain Management", *Forbes*, 2015.

[2] A. Botibol. (). The importance of customer segmentation, [Online]. Available: `https://www.bluevenn.com/blog/the-importance-of-customer-segmentation`. (accessed: 23.06.2016).

[3] T. Baumgartner and et al., "Why Salespeople Need to Develop "Machine Intelligence", *Harvard Business Review*, 2016.

[4] S. Steutermann, "Demand Forecasting Leads the List of Challenges Impacting Customer Service Across Industries", *Gartner*, 2016.

[5] A. Dey, "Machine Learning Algorithms: A Review", *(IJCSIT) International Journal of Computer Science and Information Technologies*, vol. 7, pp. 1174–1179, 2016.

[6] V. Oladokun, A. Adebanjo, and O. Charles-Owaba, "Predicting students academic performance using artificial neural network: A case study of an engineering course", *The Pacific Journal of Science and Technology*, vol. 9, no. 1, pp. 72–79, 2008.

[7] J. Kogan, C. Nicholas, and M. Teboulle, "Grouping multidimensional data", in. Springer, Berlin, Heidelberg, 2006, ch. Berkhin P. A Survey of Clustering Data Mining Techniques, pp. 25–71.

[8] A. Jain, "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[9] A. B. Ayed, M. B. Halima, and A. Alimi, "Survey on clustering methods: Towards fuzzy clustering for big data", in *6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, Tunis, Tunisia, 2014, pp. 331–336.

[10] E. W. Forgy, "Cluster analysis of multivariate data: efficiency vs interpretability of classifications", *Biometrics*, vol. 21, pp. 768–769, 1965.

[11] K. Kalti and M. Mahjoub, "Image Segmentation by Gaussian Mixture Models and Modified FCM Algorithm", *The International Arab Journal of Information Technology*, vol. 11, no. 1, 2014.

[12] A. B. Ayed, M. B. Halima, and A. Alimi, "Adaptive fuzzy exponent cluster ensemble system based feature selection and spectral clustering", in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017.

[13] K. Mupparaju and et al., "A Comparative Study of Machine Learning Frameworks for Demand Forecasting", 2018.

[14] L. Zhang, A. Stoffel, M. Behrisch, and et al., "Visual analytics for the big data era — a comparative review of state-of-the-art commercial systems", in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012.

[15] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, *Mastering The Information Age - Solving Problems with Visual Analytics*. Jan. 2010, ISBN: 978-3-905673-77-7.

[16] J. R. Harger and P. J. Crossno, "Comparison of open-source visual analytics toolkits", in *Visualization and Data Analysis*, 2012.

[17] J. Richardson, R. Sallam, K. Schlegel, A. Kronz, and J. Sun. (). Magic quadrant for analytics and business intelligence platforms, [Online]. Available: `https://www.gartner.com/doc/reprints?id=1-1XYUYQ3I&ct=191219&st=sb`. (accessed: 11.02.2020).

[18] A.Abela. (). Chart suggestions - a thought starter, [Online]. Available: `https://extremepresentation.com/tools/`. (accessed: 10.12.2010).

[19] T. Fukui, "A systems approach to big data technology applied to supply chain", in *2016 IEEE International Conference on Big Data (Big Data)*, 2016.

[20] Z. Pirani, A. Marewar, Z. Bhavnagarwala, and M. Kamble, "Analysis and optimization of online sales of products", in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017, pp. 1–5.

[21] M. Singh, B. Ghutla, R. L. Jnr, A. F. S. Mohammed, and M. A. Rashid, "Walmart's sales data analysis - a big data analytics perspective", in *2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, 2017, pp. 114–119.

[22] S. Shikha. (). Types of retailers, [Online]. Available: http://www.yourarticleli com/marketing/distribution-channels/types-of-retailers/99723.

[23] A. S. Harsoor, A. Patil, and M. Tech, "Forecast of sales of walmart store using big data applications", *International Journal of Research in Engineering and Technology*, vol. 04, pp. 51–59, 2015.

[24] S. M. Varghese, "Leveraging map reduce with hadoop for weather data analytics", 2015.

[25] P. Chouksey and A. S. Chauhan, "A review of weather data analytics using big data", *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 6, pp. 365–368, 2017.

[26] V. Chauhan and M. Sharma, "A review: Mapreduce and spark for big data analytics", Jun. 2016.

[27] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, 2012, ISBN: 9781491957660. [Online]. Available: https://www.amazon.com/Python-Data-Analysis-Wrangling-IPython-dp-1491957662/dp/1491957662/ref=mt_paperback?_encoding=UTF8&me=&qid=.

[28] (). Cross-validation, [Online]. Available: http://www.long-short.pro/post/kross-validatsiya-cross-validation-304. (accessed: 17.06.2013).

[29] S. Korolev. (). Open machine learning course. topic 7. unsupervised learning: Pca and clustering, [Online]. Available: https://medium.com/open-machine-learning-course/open-machine-learning-course-topic-7-unsupervised-learning-pca-and-clustering-db7879568417. (accessed: 26.03.2018).

[30] X. Wu, V. Kumar, J. Ross Quinlan, and et al., "Top 10 algorithms in data mining", *Knowledge Information Systems*, vol. 14, pp. 1–37, 2008. DOI: https://doi.org/10.1007/s10115-007-0114-2.

[31] X. Xu, L. Tang, and V. Rangan, "Hitting your number or not? a robust & intelligent sales forecast system", in *2017 IEEE International Conference on Big Data (Big Data)*, 2017.

[32] A. Bloomenthal. (). Coefficient of determination, [Online]. Available: https://www.investopedia.com/terms/c/coefficient-of-determination.asp. (accessed: 26.03.2020).

[33] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results", *Journal of Transportation Engineering*, vol. 129, no. 6, pp. 664–672, 2003.

[34] M. Abdellatief, E. Shaaban, and K. Abu-Raya, "Egyptian case study-sales forecasting model for automotive section", Dec. 2019, pp. 1–6. DOI: 10.1109/SmartNets48225.2019.9069751.

[35] M. Galarnyk. (). Understanding boxplots, [Online]. Available: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51. (accessed: 12.09.2018).

[36] Z. Li, C. Shang, and Q. Shen, "Fuzzy-clustering embedded regression for predicting student academic performance", in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2016, pp. 344–351.

[37] J. Sukup. (). When k-means clustering fails: Alternatives for segmenting noisy data, [Online]. Available: https://www.datascience.com/blog/k-means-alternatives. (accessed: 19.02.2018).

[38] P. M. Arsad and N. Buniyamin, "Prediction of engineering students' academic performance using artificial neural network and linear regression: A comparison", in *IEEE 5th Conference on Engineering Education (ICEED)*, 2013.

[39] A. Agarwal. (). Polynomial regression, [Online]. Available: https://towardsdat
com/polynomial-regression-bbe8b9d97491. (accessed: 09.10.2018).

[40] Y. Kashnitsky. (). Open machine learning course. topic 4. linear classification and regression, [Online]. Available: https://medium.com/open-machine-learning-course/open-machine-learning-course-topic-4-linear-classification-and-regression-44a41b9b5220. (accessed: 26.02.2018).

[41] G. Cybenko, "Approximations by superpositions of sigmoidal functions", *Mathematics of Control, Signals, and Systems*, vol. 2, no. 4, pp. 303–314, 1989.

[42] K. Hornik, "Approximation Capabilities of Multilayer Feedforward Networks", *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.

[43] G. E. Hinton, S. Osindero, and Y. W. Teh, "A Fast Learning Algorithm for Deep Belief Nets", *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[44] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python", in *9th Python in Science Conference*, 2010.

[45] S. Date. (). Testing for normality using skewness and kurtosis, [Online]. Available: https://towardsdatascience.com/testing-for-normality-using-skewness-and-kurtosis-afd61be860. (accessed: 22.11.2019).

[46] D. Sergeev. (). Open machine learning course. topic 9. part 1. time series analysis in python, [Online]. Available: https://medium.com/open-machine-learning-course/open-machine-learning-course-topic-9-time-series-analysis-in-python-a270cb05e0b3. (accessed: 10.04.2018).

[47] A. Keshvani. (). Using aic to test arima models, [Online]. Available: https://coolstatsblog.com/2013/08/14/using-aic-to-test-arima-models-2. (accessed: 14.08.2013).

[48] A. Manokhina. (). Open machine learning course. topic 10. gradient boosting, [Online]. Available: https://medium.com/open-machine-learning-course/open-machine-learning-course-topic-10-gradient-boosting-c751538131ac. (accessed: 16.04.2018).

[49] S. Lowe. (). Stratified validation splits for regression problems, [Online]. Available: https://scottclowe.com/2016-03-19-stratified-regression-partitions/. (accessed: 19.03.2016).

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[51] A. Luz. (). Why you should be plotting learning curves in your next machine learning project, [Online]. Available: https://medium.com/@datalesdatales/why-you-should-be-plotting-learning-curves-in-your-next-machine-learning-project-221bae60c53. (accessed: 06.11.2017).

[52] J. Heaton. (). The number of hidden layers, [Online]. Available: https://www.heatonresearch.com/2017/06/01/hidden-layers.html. (accessed: 01.06.2017).