Visualizing Google Analytics', Digital Marketing Campaigns' and ERP system's data using BI tools

Kagiampaslidis Michail - Christos

SID: 3305170008

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in e-Business and Digital Marketing

DECEMBER 2019

THESSALONIKI – GREECE

INTERNATIONAL HELLENIC UNIVERSITY

Visualizing Google Analytics', Digital Marketing Campaigns' and ERP system's data using BI tools

Kagiampaslidis Michail - Christos

SID: 3305170008

Supervisor:	Dr. Ioannis Magnisalis
Supervising Committee Members:	Assoc. Prof. Name Surname
	Assist. Prof. Name Surname

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in e-Business and Digital Marketing

DECEMBER 2019

THESSALONIKI – GREECE

Abstract

This dissertation was written as a part of the MSc in E-Business and Digital Marketing at the International Hellenic University by student Kagiampaslidis Michail – Christos under the supervision of Dr. Magnisalis Ioannis.

Data lied inside corporations can provide precious information in order to be used for further analysis. This thesis examines the internal data of a small wholesale and retail sporting goods company and tries to extract useful information from them. The initial goal was to retrieve, incorporate and visualize data coming from Google Analytics, Digital Marketing Campaigns and the ERP system but the magic of the knowledge pursuit led us to conduct a descriptive analysis, using those data which results will be visualized using Power BI tool. Additionally, the results of this analysis motivated us to conduct predictive analysis in order to predict future sales amount.

Looking back at the journey of the MSc studies I would like to thank my supervisor Dr. Ioannis Magnisalis for being the guide of thoughts, my family for being my biggest supporter and my IHU friends for our collaboration. Last but not least I would like to specially thank my beloved Maria.

Kagiampaslidis Michail - Christos

01/12/2019

Contents

1 Introduction

Predicting future is a human desire that has its roots million years ago. The primitives wanted to know in advance the weather, the seasons and if they would be able to find food. Nowadays, people want to know the stock price, the next card coming out of the pack, the lottery numbers and many more uncountable future events.

In order to generate accurate predictions, data could be considered as the most important asset of the procedure. Through years data has changed in terms of its variety, volume and velocity resulting to the term of big data. As data changed its structure, its volume and its speed, data analytics tools became more sophisticated and more powerful in order to manipulate them properly. Predicting the future or more formally conducting predictive analysis could be considered as a step of a wider data analysis which has two more steps, namely descriptive and prescriptive analysis. Descriptive analysis starts with the description of the current and the past state of the data, while predictive analysis incorporates the results of that analysis in order to predict the future state. Once the predictive analysis conducted, prescriptive analysis tries to identify the best course of actions needed to be taken in the future.

Having an understanding of the data analytics ecosystem, this thesis will try to induce the basic terms and techniques in order to conduct a descriptive and a predictive analysis, using a dataset retrieved from Google Analytics, Digital Marketing Campaigns and the ERP system of a small wholesale and retail sporting goods company.

On chapter 2, the terms of data analytics, big data analytics and business analytics is widely defined. Next, the three basic categories of data analytics are presented, paying special attention on the predictive analytics describing the main process and the most common techniques used. Finishing the second chapter, some related work examples in sales prediction are demonstrated.

In the beginning of chapter 3 the thesis scope is explained, and the research's questions are formulated as well. Moving forward, the two main sales forecasting methods are explained and the evaluation method for our predictive analysis will be chosen. In the end, the two main tools, Power BI and Rapidminer that will be used in the descriptive and predictive analysis are presented pointing out the reason of their selection.

The fourth chapter describes the case study implementation. The present case study is divided in two parts, the descriptive and the predictive analysis. First of all, in the descriptive part the data preprocessing is conducted. Moving to the analysis part, two dynamic dashboards are created. Some of the results of the descriptive analysis are used next in the predictive analysis. Once again, a further data preprocessing is conducted, continuing with the feature extraction, the algorithm selection and conducting to the model process constructions.

In chapter 5 the results of both analysis are presented, finishing with chapter 6 in which I make an attempt to evaluate current thesis work and also try to set future work expectations.

2 Literature Review – Theoretical Background

The exponential growth of data generating needs more sophisticated analytics solutions. Every new solution is a result of the basic analytics tools. This chapter tries to define the data analytics landscape, present the main three types of data analytics paying close attention in predictive analytics and finally explain some sales predictions models.

2.1 Definitions of Analytics

Business analytics being a sub-category of data analytics has become a disruptive part of modern digital corporations. In recent years and after the data generation explosion, data analytics have been transformed into big data analytics. The characteristics of big data, namely volume, variety, velocity and veracity (4Vs), are considered responsible for this change. 4Vs not only made fundamental changes in the way that modern businesses operate but created new opportunities regarding their decision-making process.[1][2][3]

Data Analytics

Data analytics is a form of data analysis using the contemporary information communication technology and the Web technology. More specific data analytics is the set of tools and applications used in order to explore, outline, extract knowledge from data. This knowledge is broadly used in order to study, understand and finally make predictions. In other words, data analytics could be described as data-driven learning and decision taking tools.

In general data analytics use mathematics, statistics, engineering and modeling techniques to manipulate current and historical data. Different types of data analytics may include to their process data mining in order to analyze huge datasets, machine learning to reduce the time of the analyzing process or data visualization techniques mainly used from Business Intelligent applications in order to create charts or infographics so as to better describe a situation.[3]

Big Data Analytics

In recent years big data analytics has become an important area of interest among scientists. The advent of Internet of Things and the exponential growth of digital devices led to the generation of enormous amounts of data that are widely known as big data.

As mentioned above, big data dispose four main attributes, the so called 4Vs, namely volume, variety, velocity and veracity. The volume of the data refers to their size, which in many cases is quite big (a.k.a. Terabytes or Petabytes of data). The velocity of the data is the speed of the data accumulation that is at unprecedented rates compared to previous years. The speed of data generation is due to the extended usage of social media and the advent of the smart devices that produce Gigabytes of data every single minute. Big data has a large variety of different sources such as social media, smart devices, enterprise systems and many other digital devices. All those devices produce text, video and audio data. Furthermore, there are data coming from sensors or RFIDs as a result data are different (structured, unstructured, semi-structured) formats based on the generation source. The veracity of the data relates to the accuracy of the data and the trustworthiness of the data source.[4][2]

Big data analytics based on earlier data analytics technologies using statistics, factor analysis, machine learning, data mining and visualization techniques in order to collect, process, analyze and extract results. Big data analytics are divided in descriptive, prescriptive and predictive analytics based on the desirable outcome of the analysis. Descriptive analytics are used in order to describe a phenomenon or to discover a hidden relationship. Prescriptive analytics are used in an attempt to find out the potential effects of a future business decision before the decision is actually made. Predictive analytics are used in order to develop models that are able to estimate future outcomes based on current and historical data.

Business Analytics

Business analytics could be described as an expanded form of data analytics or even better as the data analytics kind of application that performs big data analytics within business. More specifically, business analytics is the science field of exploring already known data so as to identify assumptions and conclusions to assist corporation decision making. Analytics process begins with data collection and continues with data cleaning, processing, visualization and storage. Business analytics performance is widely based in areas such as mathematics, computer programming and statistics in order to develop new

insights and aid better business decisions aiming to a better business performance. Descriptive, prescriptive, and predictive are the main types of business analytics.[1][3]

2.2 Types of Data Analytics

While there is a wide variety of data analytics techniques incorporating the most sophisticated technology that exists, generally they are divided in three main categories as analyzed below.

Descriptive Analytics

Descriptive analytics is the kind of analytics that summarizes and converts raw data in order to "describe" the past and make it intelligible to the user. Past could be mentioned as any point of time that an event happened even it was a minute ago. Descriptive analytics use past or historic data in order to build summaries, generate visual displays or dashboards and create reports answering questions such as "what has happened?" and "what is currently happening?". They have the ability to extract knowledge from past behaviors and to comprehend how they possibly impact future results.

Descriptive analytics in order to generate visualizations, dashboards and reports make use of simple and more intermediate statistics. Some of the statistics are averages, percent changes, mean, median and standard deviation in order to clarify patterns and sequences based on historical data. Descriptive analytics, in a digital marketing paradigm, demonstrate results such as total visitors of a website, conversion rate of an ad, average time on a website, average customer spent in an online shop, average pages per session etc. [7][8][9]

Prescriptive Analytics

Prescriptive analytics, a relatively new field of analytics, could be described as the type of data analytics that through the analysis of raw data tries to identify a variety of potential actions in order to achieve the optimal solution of a situation. Prescriptive analytics in order to guide users to achieve the optimal solution answers questions such as "What should be done?" and "Why should it be done?". In other words, prescriptive analytics are all about providing advices to the users in order to make the optimal decisions.

Prescriptive analytics relies on IT and mathematical techniques to structure and support the process of decision-making. To be more specific, artificial intelligence

techniques such as machine learning operate to make possible the process of the enormous amount of the raw data. In addition, mathematical techniques are involved in this process in order to computationally determine a variety of alternative actions to improve the business performance. Furthermore, prescriptive analytics working in cooperation with predictive analytics is able to recommend the optional solution of the predicted outcomes that predictive analytics has estimated.[8]

Predictive Analytics

Overall predictive analytics is a statistical-oriented science which uses data in order to predict events, trends and behaviors. Predictive analytics is a set of business intelligence (BI) technologies that uses a wide range of statistical techniques, pattern recognition techniques, machine learning and data mining. Predictive analytics collects and analyzes current and past data in order to extract valuable information for the future. It is the process of analyzing huge volumes of data and helps companies to discover meaningful patterns and relationships in order to anticipate future events, behaviors and trends in all business aspects and give them an important assistance in the decision-making process. In other words, predictive analytics answers the question of "What will happen?" and "Why will it happen?".[7][8]

Generally, predictive analytics can be used to any unknown figure which a business needs to predict or track in the future or the present. As in every similar case, the quality of the results is fully dependent on the data quality. The data are included in bigger datasets that may be produced from various heterogenous sources based on historical and current actions. Data can be divided in two categories "structured" and "unstructured" data. Structured data is comprised of clearly defined data types whose format makes them easily analyzable. Unstructured are the data that their format doesn't make them easily analyzable, including textual or non-textual, and human- or machine-generated formats. [7][8]

Predictive analytics are falling under two main categories; the *supervised* and the *unsupervised* method, based on which machine learning technique is used. In supervised method the algorithm is trained with already known historical data in order to generate predictions. On the other hand, in unsupervised method the algorithm explores hidden relationships among variables in given data sets. [7][8]

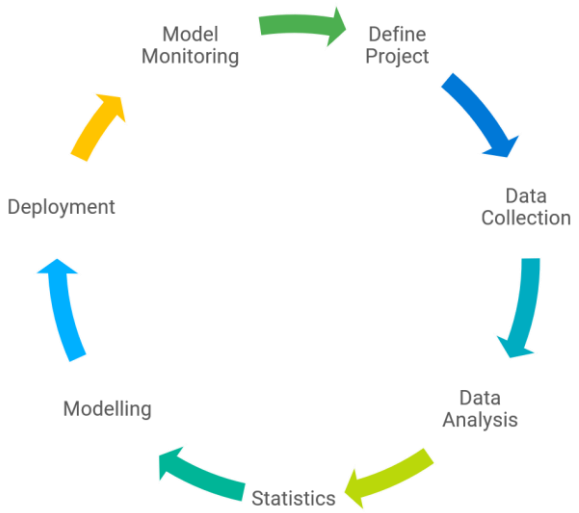
In predictive analytics all the questions and the variables, dependent and independent, are developed by experts in the field of research although the model selection and the relationship are produced driven by data. [7][8]

There is differentiation between predictive analytics and forecasting. In predictive analytics the analysis is done in a more deeply manner, where prediction is based on actual data built on specific patterns. The risk assessment is also included in the above process and now businesses can reduce the failure percentage before they even implement a campaign, a specific tactic or even their comprehensive strategy. [10][11]

Predictive Analytics Process

In order to use predictive analytics, the analyst needs to undergo a seven-step process that is a complete study from project definition to prediction system monitoring. The process starts with the project definition then data are collected. Thereinafter data analysis and statistical analysis take place. Subsequently predictive models are created choosing the best in order to deploy the predictive system and finally monitor the model performance. [10][12][13]

1. **Project definition:** In this stage the goals of the project, the deliverables, the business objectives and also the data sets are identified.
2. **Data collection**: Is the data compilation and preparation of various sources that may be homogenous or heterogenous, in order to build better models for analysis.
3. **Data analysis:** Data analysis is the process of extracting, organizing and modelling raw data in order to obtain useful information that can be applied to formulate conclusions.
4. **Statistics:** Statistical analysis is used to confirm the assumptions and hypothesis after testing them using standard statistical models.
5. **Modelling:** In this step predictive models are created in order to be used in future. From the available options the best could be chosen using multi-model evaluation.
6. **Deployment:** Through the predictive model deployment there is the option to use the analytical results into the everyday decision-making process. In this way the results, the reports and other metrics will be based on the automated model.
7. **Monitoring:** The model needs to be monitored to check the model performance in order to ensure that the desired outcomes are produced.



Picture 1: Predictive Analytics Process

Analytical Techniques

The approaches and techniques to conduct predictive analytics can be divided into regression techniques and machine learning techniques.

Regression Techniques

Regression is a statistical process and a type of predictive modeling techniques that examines the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modelling and finding the casual effect relationship between the variables. Regression analysis is a significant tool for modelling and analyzing data. Some of the predictive models that can be used while performing predictive analytics will be briefly presented subsequently. [10][13][14]

Linear regression model

Linear regression model is the statistical approach for modelling the relationship between a dependent variable and one or more independent variables. The relationship is defined as an equation that predicts the dependent variable as a linear function of the independent variables. In the case of one independent variable, the process is called simple linear regression. In the case of more than one independent variable the process is called multiple linear regression. [13][15]

Discrete choice models

In the multiple linear regression, when the dependent variable is not continuous but categorical or binary then other models are better suited. These models are the discrete choice models and some of them are logistic regression, multinomial logistic regression and probit regression models. Logistic and probit regression models are used when the dependent variable is binary. Logistic regression model is a statistical model that predicts the outcome of a categorical dependent variable. Multinomial logistic regression is an extension of the binary logit model for cases where the dependent variable has more than two categories. The multinomial logistic regression is the appropriate approach especially when the dependent variable categories are not ordered. Probit regression model is the model where the dependent variable can only take two values, for example (man or woman). [10][13][15]

Time series models

Time series models are used in order to predict or forecast the future behavior of variables. These models account for the fact that data points taken over time may have an internal structure that should be accounted for. Using the dynamic path of a variable can optimize results since the predictable feature of the time series can be projected into the future. ARMA (autoregressive moving average) widely used in stationary analysis and ARIMA (autoregressive integrated moving average models) used in non-stationary analysis are the two most common algorithms of time series. ARMA is developed combining autoregressive (AR) and moving-average (MA) models. Recently more sophisticated models such as ARCH (autoregressive conditional heteroskedasticity) and GARCH (generalized autoregressive conditional heteroskedasticity) that commonly used for financial time series are trying to model conditional heteroskedasticity.[21]

Classification and Regression Trees

This is the kind of analysis used to identify the statistical model that maximizes its accuracy for predicting the value of a dependent variable in a dataset which may contain either categorical or continuous variables. These methods create a tree-like structure to map observations of a variable in order to reach conclusions about variables target value. The model that is generated simply classifies the examples in order to conclude to the predicted outcome that can be a class (Classification tree) or a real number (Regression tree).[14]

Machine Learning

In 1950 Alan Turing in his seminal paper “Computing Machinery and Intelligence” first defined the underlying idea of Machine Learning, posing the question "Can machines do what we (as thinking entities) can do?". ML could be basically described as any type of computer program that can “learn” by itself without having to be explicitly programmed by a human.[16]

In 2019 within Artificial Intelligence, ML is the scientific study using algorithms and statistical models that computers employ in order to learn from a sample dataset and improve themselves, having the opportunity to predict future performance measure of a given task.[17][18]

Commonly a ML model, that is the main pillar of a basic ML solution, is constructed using a dataset and ML algorithms. In order the model to be developed it is needed to input a dataset of historic data into the algorithm. To come up with the best ML solution, first of all the problem needs to be defined considering past events and simultaneously taking into account what is needed to be predicted, then the historical data need to be gathered. In the sequent the model is developed and finally ready to be used in order to perform tasks and extract knowledge.[17]

ML techniques can be broadly classified into four categories. The categories are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

Supervised learning’s goal is to generate a function or a pattern using a training dataset that contains already labeled data. The algorithm that supervised learning approach uses is trained by already labeled data that consist the ground truth for the training phase. Training data comprise an input vector **X** and an output vector **Y** of tags or labels. The vector **Y** tags or labels are the characteristics of vector **X**. Every training set could also contain unlabeled vector **X** data. After the training phase, the ML system constructs a function in order to estimate the output of a random X data input. The function can be reconstructed after the comparison of the output and the correct result. The vector **Y** labels can be assigned to each vector **X** entry by a human supervisor or a machine supervisor. Nowadays, more often machines are labeling the **X** vectors but because of the high error rates of the machines human’s supervision is needed in an attempt to decrease errors. However, in many cases there are machines that are completely

reliable in unsupervised labeling. The two categories of supervised algorithms are classification and regression.[19][20]

In contrast with supervised learning methods in unsupervised methods there are no labeled data. In other words, the input **X** data has no vector **Y** tags or labels. The data has no labels due to many reasons, with the main being the unavailability to label them because of their velocity, variety and volume. The ML system analyzes the given data set that has no classified or categorized data in order to find trends, extract hidden patterns and relationships among them. Next, it generates a function used to find out the output for a random input. Clustering or cluster analysis is the main category of unsupervised learning algorithms.[19][20]

Semi-supervised learning is a mix of supervised and unsupervised learning; the given data is a combination of labeled and unlabeled data. This combination of data is used from ML systems in order to construct proper classification models. The semi-supervised approach is constructed in order to take advantage of the limited volume of labeled data combined with the huge amount of unlabeled data. As a result, the semi-supervised algorithm generates more accurate results saving time by not requiring all of the data inputs to be labeled. In semi-supervised classification the goal is to predict classes of future data better than that generated from a classic classification model.[19][20]

The reinforcement learning method algorithm uses reinforcement signals in order to reward the system for its desired operation or the desired outputs. The algorithm operates without training data but lets the system to use observations gathered from the interaction with the environment in order to find out the proper actions and outputs so as to optimize its performance.

Some of the most widely used machine learning techniques are:

Neural networks

Neural networks Are nonlinear sophisticated modeling techniques that are able to model complex functions. Neural networks are used when the nature of the relationship between inputs and output is unknown. The main benefit of neural networks is that they learn the relationship through training. [10]

Multilayer perceptron (MLP)

MLP is a class of feedforward artificial neural network and consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Each node, except for the

input nodes, is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. [10]

Support vector machines

Support vector machines (SVMs) are used in order to determine and exploit complex patterns in data by clustering, classifying and ranking data. SVMs are learning machines that are used to perform binary classifications and regression estimations. In comparison to other machine learning methods SVM is more suitable in solving the problems of high dimension and local minima. SVMs' main idea is to generate a linear classifier with maximal classification margin while minimizing the training error. The key attribute to SVM is the kernel function because of using different kernel functions generated different SVM algorithms.[10]

Naïve Bayes

Naïve Bayes based on Bayes conditional probability rule is used to perform classification problems. It is the simplest form of Bayesian network and assumes that the predictors are statically independent which makes it an effective classification tool. [10]

K-nearest neighbors

The k-NN belongs to the sector of pattern recognition statistical methods. The method does not impose a priori any assumptions about the distribution from which the modeling sample is drawn. It involves a training set with both positive and negative values. The data sample is classified by calculating the distance to the nearest neighboring training case. The performance of k-NN classifier is based on the distance measure used to locate the nearest neighbors, on the decision rule used to produce a classification from the k-nearest neighbors and on the number of neighbors used in order to classify the new data sample.

2.3 Sales Predictions Models

Sales predictive analytics using machine-learning models is widely used in modern business intelligence. Sales forecasting in some cases may not be an easy problem because of the lack of data or missing data or noisy data. In order to generate predictions of sales the problem can be considered as a regression or time series problem. In terms of time series, bibliography can illustrate a wide range of different time series models such as Holt-Winters, ARIMA, SARIMA, SARIMAX, GARCH, etc. Some of the most

common limitations of time series approaches for sales prediction is the lack of long time period historical data, the amount of missing and noisy data and all the external factors which force sales. Practice shows that sales prediction as a regression problem in many cases can generate better results compared with time series methods. Regression approaches' main assumption is that patterns that found in the past research can be repeated in future problems. In order to have more accurate results it is needed to take into account the noise into the sales data and also some edge values.

In order to conduct predictive analytics is vital first of all to conduct descriptive analysis. Predictive analytics can illustrate the sales distribution and can generate many different data visualizations in order to find hidden relations and factors with high impact on sales. In order to do predictive analytics supervised machine learning methods is one of the best solutions with tree-based classifiers being the most popular. Machine learning methods most of the times cannot work with non-sanitary data. During the predictive modeling in many cases more than one predictive models are constructed with a lot of different features, so in many cases it is useful to combine these results to take the best performance. The two main approaches of combining results are bagging and stacking.

In bibliography there are many simple and even more advanced forecasting models based on the above techniques. A small sample of them will be described below.

In [21] the authors created a sales prediction model using ARIMA in order to predict the amount of sales for a tractor company using time series data. Time series problems are considered those using a dataset that consists of values of some entity measured at sequential points of time. In order to manage the huge amount of the sequential data time series analysis has to be automated. ARIMA model is the most widely used model on time series data which are sequential observations generated in equally spaces time periods. ARIMA is generally preferred if the observations of time series are statistically dependent to each other. In order to build a time series model using ARIMA or Box-Jenkins approach, as is also known, Autoregressive (AR) and Moving Average processes are needed. In time series current values depend on its own previous values. In AR process a future variable is presumed to be a linear combination of past observations of the same variable always containing a random error. In MA process the current deviation from mean depends on previous deviations. The MA models use past errors as the explanatory variables. In other words, a MA process is a linear regression of the current values of the time series against the random shocks of one or more previous values. In

the end, in order to model a time series with ARIMA the series needs to be stationary. Stationary series are easier to be modeled because they are not dependent on time and they don't have trend or seasonal effect.

Authors in [22] in order to create a regression predictive model that predicts the sales amount of fresh agricultural products used three different machine learning methods. The models use weather factors in order to optimize the predictions. In the beginning, a baseline prediction model was created based only in previous days sales in lags that had values t-1, t-2, t-3 and t-7. Additionally, some dummy features created based on the position of the day in the week (weekend or not) and the kind of the day (national statutory holiday or not). Furthermore, a weather prediction model was created with the same structure as the baseline model with the difference being on the weather factors added. The weather condition features are the category of weather condition in day t, the category of wind level in day t and the category of air quality index in day t. In order to verify the results three different machine learning algorithms were used. These three are Ridge Regression (RR) based on Linear Regression, Random Forest (RF) based on Tree Model and Support Vector Machines (SVMs) based on the Kernel Method. In order to compare forecasting results between baseline prediction model and weather prediction model Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) were used. Concluding, the results of the weather prediction model under all three algorithms had a significant reduction compared to the baseline prediction model. So, the sales predictions of fresh agricultural products are strongly linked with the weather conditions.

3 Problem Definition and Tools

In this chapter, firstly will be placed the thesis scope and the research questions as well. Next, the general sales forecasting framework will be set presenting the metric that will be used in order to evaluate the predictive models that will be generated through our analysis. Lastly, the two tools that are used in this thesis will be widely explained.

3.1 Thesis Scope – Research Questions

The constant technology evolution lead to radical changes in the business environment. The data growth explosion makes it necessary even for small and medium-sized businesses to enter the digital world. As it is mentioned in the previous chapter there are many tools that are available in order to support this big step. These tools are able to add value on business process firstly through the historical analysis and secondly via the predictive analysis. Going through the big data season it was necessary to get analyzed the term "big data" and to get decomposed the analytics area, illustrating the three main categories of them which are the descriptive, the prescriptive and the predictive analytics. Furthermore, we had a small glance on the most used analytical techniques describing their basic function. Finally, a deeper view on the sales predicting models was presented giving the opportunity to understand the methods that will be used in the next chapters.

Strongly believing in the inestimable value of the business data, this thesis will apply all the literature review research knowledge on a medium-sized business so as to add value on its process. In order to reach the desirable outcomes this thesis will attempt to address the following research questions:

1. How can data from different data sources be linked?
2. What is the added value for a business through descriptive analysis and historical data visualization?
3. How can we generate a predictive model based on internal business data?

3.2 Sales Forecasting

Through decades sales forecasting became a vital process for companies' survival among rivals', suppliers' and customers' continuous pressing forces. Sales forecasting is the process of estimating future sales. In other words, sales forecasts generate predictions,

based on historical data, as to the total amount of future sales in a future period. This period can be daily, weekly, monthly, quarterly, half-annually, or annually. Forecasting sales provides companies with the necessary knowledge in order to make plans and take future intelligent business decisions. The quality of decisions is directly connected to the accuracy of the forecast.

In general, most of the forecasting approaches take into account data generated in the past. Forecast techniques are based on the assumption that what happened in the past will happen again in the future. The above techniques do not take into account potential future variations that may happen in the technology's, marketplace's and costumers' environment. Sales forecasting generation procedure is easier for already established businesses because of the existence of internal historical data. In contrast, newly founded businesses' have to be based on less-verified data including market research data.

The most widely used methods for sales forecasting and consequently for predictive analytics are Regression analysis and Time series analysis.

Table 1: Sales forecasting methods

Method	Description	Algorithm
Regression	Predict the numeric target label of a dataset. The prediction is based on learning from historical data.	Linear regression, logistic regression
Time series	Predict the value of the target value for a future time period based on historical data.	Autoregressive integrated moving average (ARIMA), exponential smoothing, seasonal trend decomposition

Regression being one of the most common predictive analytics techniques is generally based on the function y = f(x), where y is the dependent variable that needs to be predicted and x is the predictor variable that is used in order to predict y. The two most widely used techniques are linear regression for numeric predictions and logistic regression for classification predictions.

Linear regression is one of the oldest and also the most easily to explained method. The central idea is to create a function that will explain and predict the value of the target (dependent) variable based on the already known values of the independent variables. In order to create the function, it is necessary to identify which of the attributes (independent variables) generate the most accurate predictions.

Logistic regression is a statistical model that basically uses a logistic function in order to predict a binary dependent variable. The dependent variable has only two possible values such as "0" and "1". The logistic regression model is created and used to model the probability of a certain class, in our case increase/decrease sales.

On the other hand, there are the time series forecasting approaches that are also among the most popular predictive analytics methods. In contrast with regression methods time series methods have two significant differences. In time series approaches the value of the dependent variable is needed to be predicted but in order to predict it is essential to know how this variable has changed through the time in the past. In regression models the time dimension of the data is not obligatory in order to generate the predictions. Additionally, in time series in many cases there are not available data, or we may not be interested in data for other attributes that could affect the dependent variable. So, in time series forecasting the independent variables are not necessarily in all models but only for multivariate time series.

In our case study's regression problem will be applied several machine learning methods using time series features in order to generate predictions.

The goal of the problem is to predict the daily clear sales for the next four days using data coming from Google Analytics, Digital Marketing Campaigns[1] and the ERP system of a small wholesale and retail sporting goods company. Since the problem involves time component, it will be manipulated as a time-series forecasting problem. Times series can be cast as a supervised learning problem hence various Machine Learning methods can be applied. The basic concept is to build a model, based on previously observed values, able to predict as accurate as possible the amount of sales for the next. The performance will be calculated by the root mean square error (RMSE). RMSE represents the difference between the values predicted by the model and the values observed. The formula is given below:

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

Where $\hat{y}_i$ are the predicted values and y_i the actual values for T different predictions.

[1] It must be pointed out that Digital Marketing Campaigns' data where retrieved through the Google Analytics platform.

3.3 Rapidminer

Developed in 2001 by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer at the Artificial Intelligence Unit of the Technical University of Dortmund YALE (Yet Another Learning Environment) is considered to be the ancestor of RapidMiner. In 2006 the continuous development of the program led to the foundation of a company named Rapid-I by Ingo Mierswa and Ralf Klinkenberg. In the next year the YALE and the company Rapid-I GmbH changed their name to 'RapidMiner'.

Being one of the most widely used data analytics tool RapidMiner provides a unified platform environment for data manipulation, deep learning, machine learning, text mining, predictive analysis and business analytics. Businesses, Universities and labs are only few of the organizations that use it for commercial, educational and research applications. RapidMiner supports each phase of analysis process, including data preprocessing, model generation, optimization, evaluation and results visualization in order to produce the desirable output.

RapidMiner uses a client/server model, written in the Java programming language, with the server able to be offered as a Software, as a Service or on private cloud infrastructures. RapidMiner is an advanced analytical solution providing the ability to users to perform the analysis without needing to write a single line of code. The program provides a graphical user interface in order to perform the analysis. Beginning with the RapidMiner user creates a workflow that is also called *Process* which combines multiple tasks or *Operators*. Each operator is responsible to perform a single function during the process with the output of each operator being the input for the next in the sequence. Moreover, RapidMiner can also be called from other programs or used as an API. Although, RapidMiner contains over than 100 learning schemes for regression, classification, time series and clustering analysis they can be extended using Python and R scripts. [23]

In January 30, 2019 RapidMiner Named a Leader in the Gartner's 2019 Magic Quadrant for Data Science and Machine Learning Platforms for Sixth Consecutive Year. Gartner's report notices that RapidMiner as a leader in data science and ML market provides solid capabilities in depth and breadth across the whole data exploration, model development and operationalization process contributing at the same time exceptional service and support. In 2018 RapidMiner was also named a Leader in the Forrester Research, Inc. report: "The Forrester Wave™: Multimodal Predictive Analytics And

Machine Learning Solutions" for the second year in the row. The report pinpoints that RapidMiner offers "the most productivity-enhancing automated model creation features available" and concludes to the point that RapidMiner is a proper solution for organizations looking to escalate the machine learning practices in their internal environment.[24][25]

In order to implement the analysis in our case study RapidMiner has been chosen. A critical criterion was the 2019 annual Gartner's report that placed for sixth consecutive year RapidMiner among the leaders of data science and machine learning platforms. The other three platforms that are among leaders are KNIME, TIBCO Software and SAS.

RapidMiner remains a leader since 2014 because of the balanced performance of platform usage and data science sophistication. The platform's accessibility makes it a proper solution for data scientists, researchers, students and executives. The openness to open-source code makes it also appealing to experienced scientists too.

KNIME because of its multi-functionality remains among the leaders and for many constitutes the "multi-tool" in Leaders' marketplace. KNIME is a free open-source solution. Its vision and roadmap are among the best in the marketplace.

SAS even if it faces threats from multiple sides such as other big vendors and open-source platforms remains among the Leaders for consecutive years. It still has a strong presence in the market because of its wide range of capabilities.

Last but not least, there is TIBCO Software which was created through the coalescence of Jaspersoft and Spotfire -reporting and BI platform vendors- of Statistica and Alpine Data -descriptive and predictive platform vendors- and of StreamBase Systems -a streaming analytics vendor. In 2019 TIBCO Software joined the leaders' marketplace because of its powerful and intergraded capabilities.

Picture 2: Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms (as of Nov 2018)

3.4 Power BI

The "4th industrial revolution" a term that firstly reported by Klaus Shwab, led to the generation and optimization of the current Business Intelligent (BI) solutions. BI programs have been defined as systems that collect, manipulate and visualize structured data from a wide variety of sources. The enormous amount of data that produced in a daily base by organizations now are easy to get analyzed thanks to BI software. Nowadays, BI solutions provide capabilities such as real-time monitoring and predictive analytics. More and more organizations implement BI tools in their business chain, gaining competitive advantage through the new levels of knowledge that generate.

In order to be considered successful, a BI tool needs to provide highly-speed deployment capabilities. In the age of velocity we live in huge amounts of data are produced every second, so there is the need of an immediate analysis in order to make almost real-time decisions. Additionally, another important factor for a successful BI tool is the ability of data processing. The main objection in data preprocessing is to recognize and separate the relevant and the redundant data. The user has to keep only the important

information for the analysis and reject all the data that produce noise in dataset. The last important attribute of a BI tool is the data visualization techniques that provides. The BI programs are all about visualizations, so the ability of analytical presentations and the dashboards generation for real-time monitoring are some of their most important features.

The current trend of BI tools are programs that are easy to use with a friendly user interface that promote the use by not only IT specialists. The two leaders in the category are Microsoft Power BI and Tableau. For our analysis and visualizations we will use the Microsoft Power BI solution.

Picture 3: Gartner 2019 Magic Quadrant Leader for analytics and business intelligence platform (as of Jan 2019)

Power BI is a cloud-based Microsoft software for data analysis and reporting. It is built to be useful not only for developers and data scientists but also for business analysts and demanding users. Power BI provides a user-friendly interface for creating reports, with a stiff based background of components that help in deeper data analysis.

Power BI Desktop is the main component that assisted by Power Query, DAX, Power BI Service, and Power BI Mobile App, all of them being responsible for

connecting data sources, building reports, performing analytics, sharing reports, creating visualizations etc.

Power BI as the main component of the solution is responsible for the report development. It is a free lightweight tool (≈150MB) that can be easily downloaded, installed and used. Currently Power BI is only available on Windows operating system, but there are ways in order to get installed in other machines.

Power Query Desktop is the tool that is responsible for connecting Power BI to various data sources and preparing the data. It creates the connections, gets the data, gives the opportunity to modify them and finally loads the data into program. DAX and Modeling are the component that do all the calculations and configurations for the user. It also creates the relationships among the tables. DAX (Data Analysis Expression language) itself is the language that is used by the platform in order to perform calculations. Power BI Service is an online service hosted in the Power BI website responsible for publishing and sharing reports. Power BI Mobile App is the alternative way to access, form and browse reports from distance. It is a free application available in the Windows Store or Microsoft, App Store of Apple, or Google Play store.[26]

Even if Power BI dominates the field, Tableau software constitutes a reliable solution for exploring, analyzing and illustrating data in a wide variety of well-structured interactive visualizations. Tableau's popularity is due to its drag-and-drop interface that makes it simple in usage even for users with no programming background. At this time there is a free cloud-based Tableau version for both Windows and Mac that called Tableau Public, but there are also paid versions. The main differences between the two versions of the software are the type of files that can be uploaded, shared and saved as the paid versions offers more flexibility. Tableau software in many cases faces a crucial drawback, that is the necessity of data cleaning and reformation before they can be uploaded. This drawback raises the barriers for many fresh users.[27]

4 Materials and Methods (Case Study Implementation)

In this chapter is going to be explained the dataset used in our case study along with the preprocessing that underwent. Continuing, the whole procedure and the outcome dashboards of the descriptive analysis conducted will be presented. In the last part of the implementation the predictive analysis is widely explained, demonstrating the features extracted, the algorithms used and the final predictive process that was created.

4.1 Dataset

The dataset used in this case study has been retrieved from two different sources. As the thesis research questions describe, the two sources are the ERP system and the Google Analytics account[2] of a sporting goods company that is based in Thessaloniki and sells via three channels, its online store, its physical store and by phone to retail and wholesale customers. The time period of the sample is from 04/01/2018 to 17/10/2019.

ERP stands for Enterprise Resource Planning and refers to the type of software that organizations widely use in their daily activities in order to collect, store, manage and understand the data produced. Accounting and finance, marketing, procurement, sales and human resources are only a few of the departments within an organization that can benefit using such a software. ERP systems are designed around a centralized relational database which collects business information and stores them in tables in order to secure the data interoperability. There are three types of ERP systems based on their implementation: on-premises ERP systems that are installed and run on the local computers on the premises of the organization that uses the software, cloud ERP systems that are stored in clouds outside the organization and used on-demand. The third category consists the hybrid ERP systems that combine the capabilities of the two previous ones.

The case study company uses the on-premises Singular Logic Business ERP version 9.4.0. In order to retrieve the first part of the dataset that is the online store's daily clear sales from the ERP system the next path was followed: administration – information –

[2] As mentioned before (footnote 1) Google Analytics account includes also Digital Marketing Campaigns' data.

OLAP reports – OLAP – sales analysis. We selected the time period for our analysis that is 04/01/2018 – 17/10/2019 and the *sales person* that in our case is the website. Finally, we chose the data that we needed in order to create the first table. In the first column there is the *date*, in the second column there is the *quantity* of the products that was sold in that day and in the last column there is the *clear value* of those product. The dataset does not contain data for the national holidays and the weekends. The outcome extracted in .xlsx counting 460 lines had the format that is displayed in the following table.

Table 2: ERP data extraction sample

Date	Quantity	Clear value
04/01/2018	25	354,0799973
05/01/2018	35	606,8699989
06/01/2018	66	702,3700023

The second part of the dataset is composed of Google Analytics data. Google Analytics (GA) is a web analytics online service, part of the Google Marketing Platform that collects websites' activity of the users. Session duration, pages per session, and bounce rate are only some of the metrics that GA track. Google also supports the integration of Google Ads into GA to track campaigns' performance. GA provide custom dashboards in order to give the opportunity to the users to have an in-depth data view.

GA analytics data are exported by the assistance of Power BI. Using the Power BI Desktop, on the home view the *Get Data* button gives the opportunity to get data from a wide range of sources. The data for this case are retrieved from GA account and more specifically from the E-commerce view of the account. In order to extract all the available values, we did the same process many times because Power BI can only retrieve ten columns per time. The retrieved data are stored in two tables each one containing 655 rows. The first table contains data based on sales via website and the second one contains metrics that describe website's traffic. The two tables have the following format as they are in Power BI Desktop.

Date	Month of Year	Product Revenue	Revenue	Shipping	Tax
Δευτέρα, 1 Ιανουαρίου 2018	201801	18,140000000000001 €	14,630000000000001 €	7 €	3,51 €
Τρίτη, 2 Ιανουαρίου 2018	201801	117,3 €	110,29000000000001 €	11,119999999999999 €	26,469999999999999 €
Τετάρτη, 3 Ιανουαρίου 2018	201801	867,240000000000001 €	699,37 €	3,5 €	158,919999999999999 €
Πέμπτη, 4 Ιανουαρίου 2018	201801	41,450000000000003 €	33,439999999999998 €	3,5 €	8,02 €
Παρασκευή, 5 Ιανουαρίου 2018	201801	0 €	0 €	0 €	0 €

Picture 4: Sales data retrieved from Google Analytics

Date	Month of Year	Entrances	Hits	Bounces	Bounce Rate	ads CTR	ads CPC	ads Cost	ads Clicks	% New Sessions	% Exit	ads CPM	ads Impressions	New Users	Organic Searches	Pages / Session	Pageviews	Unique Pageviews
Δευτέρα, 1 Ιανουαρίου 2018	201801	117	567	64	0,55	0,08	0,2446 €	9,050000000000001 €	37	0,81	0,21	20,6149999999998 €	439	95	64	4,67	546	371
Τρίτη, 2 Ιανουαρίου 2018	201801	273	1314	136	0,50	0,09	0,215 €	9,460000000000001 €	44	0,72	0,21	19,1111 €	495	196	153	4,71	1287	880
Τετάρτη, 3 Ιανουαρίου 2018	201801	268	1627	104	0,39	0,06	0,2273 €	9,32 €	41	0,75	0,17	14,3606 €	649	202	154	5,98	1603	1064
Πέμπτη, 4 Ιανουαρίου 2018	201801	303	1545	139	0,46	0,06	0,2173 €	8,91 €	41	0,69	0,21	13,8569 €	643	208	161	4,78	1449	1025

Picture 5: General data retrieved from Google Analytics

4.2 Descriptive Analysis

Descriptive analysis is an important first step for conducting predictive analysis. Through descriptive analysis basic features of data can easily be described, data distribution can be observed, and outliers can be detected. Graphs and charts are widely used in order to demonstrate the quantitative observations having a more complete view on the dataset.

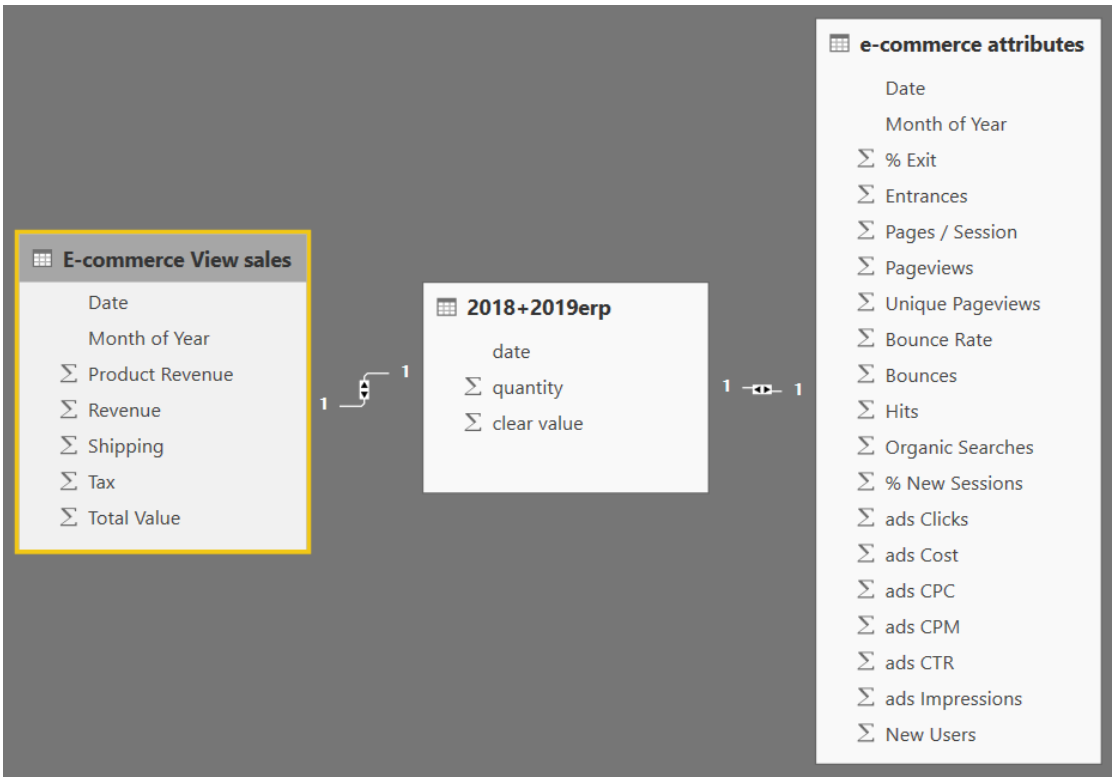
4.2.1 Data Preprocess

In order to perform descriptive analysis Power BI desktop is used. In our analysis three tables have been combined. Two of them are based on Google Analytics data and the third is based on the ERP data. All three tables are presented in chapter 4.1. The table that was retrieved from the ERP system contains the daily clear value and quantity of sales. The first table of Google Analytics contains the daily *product revenue*, *revenue*, *shipping* and *taxes*. The second Google Analytics table contains metrics based on the website's traffic such as the daily *entrances*, *hits*, *bounces*, *bounce rate*, *ads CTR*, *ads CPC*, *ads cost*, *ads clicks*, *ads CPM*, *ads impressions*, *new sessions percentage*, *exits percentage*, *new users*, *organic searches*, *pages per session*, *pageviews* and *unique pageviews*.

Entrances in Google Analytics is the sum of the times users have entered a website. *Hits* are the total amount of interactions between users and a website. *Bounces* are the number of users that left a website without interacting with that, and the *bounce rate* is the percentage of these bounces. The *ads CTR* is calculated by dividing the number of clicks to our ads and the number of impressions of the ads. *Ads cost* is the total cost of the ads and *ads CPC* is the cost per ads click. In the same way *ads CPM* stands for ads cost

per mille or the amount the company pays per one thousand visitors who see the advertisements. *New users* is the number of the new users inserting a website. *Organic searches* is the total amount of the entries in a website after searched a term referring to the website in a search machine. *Pages per session* count the amount of the pages that a single visitor opens in a single session. *Pageviews* is defined as a view of a page on the website in contrast with the *unique pageviews* which are the aggregated pageviews generated by the same user during the same session. A unique pageview represents the number of sessions during which that page was viewed one or more times.

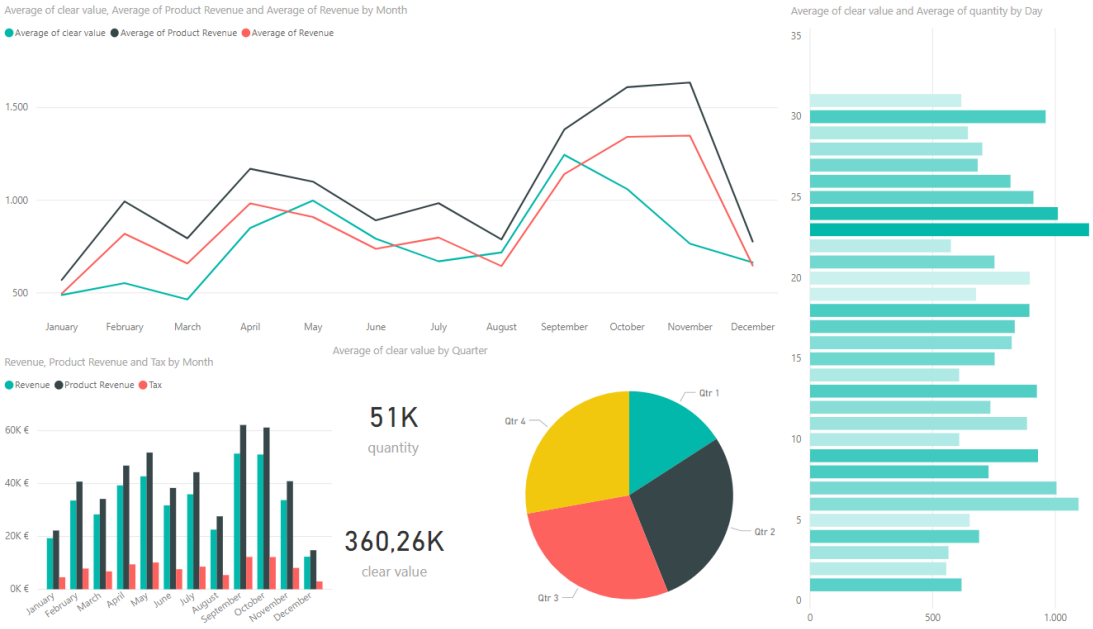
Using Power BI, we created relationships between the three tables above in order to link them. The relationships created using *Manage Relationships* command. Both Google Analytics tables were connected to the ERP system table using *date* columns as the related columns, 1 to 1 cardinality and both directions filtering. The schema created is presented below:



Picture 6: Tables' relationships schema

4.2.2 Implementation

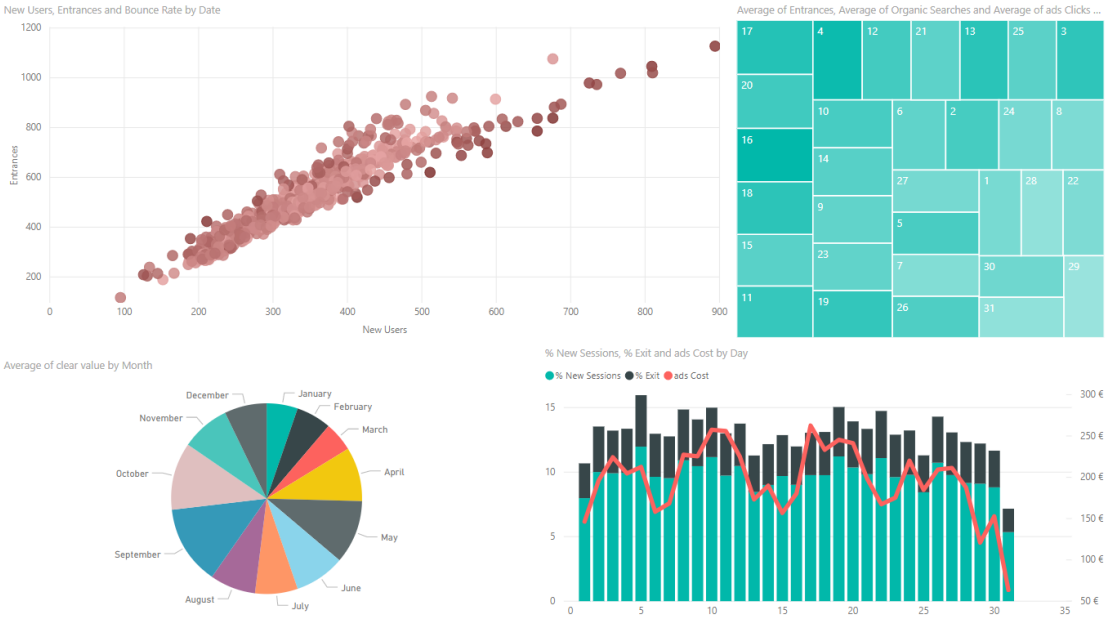
Once data preprocessed and connected, two dynamic dashboards are created in order to illustrate the data relationships. The first dashboard presents relationships based on sales and product quantity metrics from both Google Analytics and ERP data.



Picture 7: Power BI dashboard based on sales data

On the line chart (picture 7) there are presented the daily average of clear sales, product revenue and revenue of each month of a year. On the clustered column chart can be observed the monthly revenue, the product revenue and the taxes. On the two cards are presented the total clear value and the total quantity as they are in ERP dataset. On the pie chart can be observed the average daily clear value based on the quarter the date belongs and also the percentage of the total clear value based on the quarter. On the clustered bar chart are illustrated the total average of clear value and the quantity on each day of a month.

On the second dashboard we used ERP system data and also data from Google Analytics, trying to create relations between clear value and important metrics produced from daily website performance. The dashboard has the following format:



Picture 8: Power BI dashboard based on traffic data

On the scatter chart are displayed the new users and the entrances for each time point in our dataset colored depending on bounce rate of the time point. On the pie chart is presented the monthly average clear value. Moving forward, on the treemap can be observed the daily average entrances colored based on ads clicks. Finally, the line and stacked column chart presents the new sessions percentage, the exits percentage and with the red line is pointed the ads cost for each day of month.

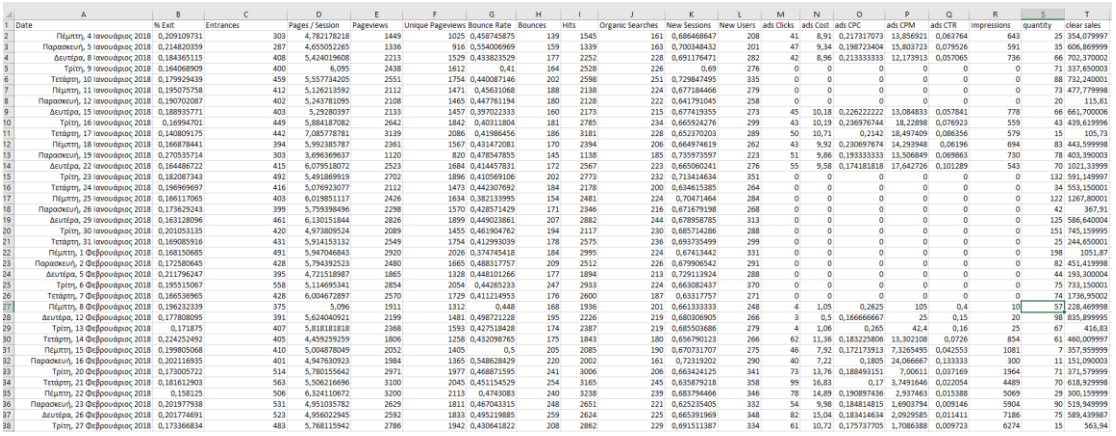
4.3 Predictive Analysis

Once the descriptive analysis conducted its outcomes will be used in the predictive analysis that will be widely explained in sequence. The analysis begins with the preprocessing continuing with the feature extraction and finishing with the model generation.

4.3.1 Data Preprocess

The dataset consists of two files, the first has been retrieved from Google Analytics (including Digital Marketing Campaigns' data) and the other from the ERP system. In order to use them to generate the models and perform predictions a basic preprocessing is performed via Power BI. In Power BI desktop using *Edit Queries* Google Analytics table dropped the *revenue* and *product revenue* columns because of their invalidity that

will be explained in chapter 5.1. In order to merge the *e-commerce attributes* table with the ERP table the *Merge Queries* command was used and also the full outer join type (all rows from both tables) based on the *date* column. On the new table all the rows containing *null* clear sales value were deleted. These nulls were created because on the weekends and National holidays the case study business is closed so there are no sales values. Finally, all the days with negative or extremely high sales are excluded from the dataset as they are considered noise for the models. The last table exported in 2018_2019_final.xlsx file containing 416 rows and 20 columns has the following format.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Date	% Exit	Entrances	Pages / Session	Pageviews	Unique Pageviews	Bounce Rate	Bounces	Hits	Organic Searches	New Sessions	New Users	ads Clicks	ads Cost	ads CPC	ads CPM	ads CTR	Impressions	quantity	clear sales
2	Πέμπτη, 4 Ιανουάριος 2018	0,209109731		4,782178218	1449	1025	0,458745875	139	1545	161	0,686468647	208	41	8,91	0,217317073	13,856921	0,063764	643	25	354,079997
3	Παρασκευή, 5 Ιανουάριος 2018	0,214820359	303	4,655052265	1336	916	0,554006969	159	1339	163	0,700348432	201	47	9,34	0,198723404	15,803723	0,079526	591	35	606,869999
4	Δευτέρα, 8 Ιανουάριος 2018	0,184365115	287	5,424019608	2213	1529	0,433823529	177	2252	228	0,691176471	282	42	8,96	0,213333333	12,173913	0,057065	736	66	702,370002
5	Τρίτη, 9 Ιανουάριος 2018	0,164068909	408	6,095	2438	1612	0,41	164	2528	226	0,69	276	0	0	0	0	0	0	71	337,650003
6	Τετάρτη, 10 Ιανουάριος 2018	0,179929439	459	5,557734205	2551	1754	0,440087146	202	2598	251	0,729847495	335	0	0	0	0	0	0	88	732,240001
7	Πέμπτη, 11 Ιανουάριος 2018	0,195075758	412	5,126213592	2112	1471	0,45631068	188	2138	224	0,677184466	279	0	0	0	0	0	0	73	477,779998
8	Παρασκευή, 12 Ιανουάριος 2018	0,190702087	402	5,243781095	2108	1465	0,447761194	180	2128	222	0,641791045	258	0	0	0	0	0	0	20	115,81
9	Δευτέρα, 15 Ιανουάριος 2018	0,188935771	403	5,29280397	2133	1457	0,397022333	160	2173	215	0,677419355	273	45	10,18	0,226222222	13,084833	0,057841	778	66	661,700006
10	Τρίτη, 16 Ιανουάριος 2018	0,16994701	449	5,884187082	2642	1842	0,40311804	181	2785	234	0,665924276	299	43	10,19	0,236976744	18,22898	0,076923	559	43	439,619996
11	Τετάρτη, 17 Ιανουάριος 2018	0,140809175	442	7,085778781	3139	2086	0,41986456	186	3181	228	0,652370203	289	50	10,71	0,2142	18,497409	0,086356	579	15	105,73
12	Πέμπτη, 18 Ιανουάριος 2018	0,166878441	394	5,992385787	2361	1567	0,431472081	170	2394	206	0,664974619	262	43	9,92	0,230697674	14,293948	0,06196	694	83	443,599998
13	Παρασκευή, 19 Ιανουάριος 2018	0,270535714	303	3,696369637	1120	820	0,478547855	145	1138	185	0,735973597	223	51	9,86	0,193333333	13,506849	0,069863	730	78	403,390003
14	Δευτέρα, 22 Ιανουάριος 2018	0,164486722	415	6,079518072	2523	1684	0,414457831	172	2567	223	0,665060241	276	55	9,58	0,174181818	17,642726	0,101289	543	70	1021,33999
15	Τρίτη, 23 Ιανουάριος 2018	0,182087343	492	5,491869919	2702	1896	0,410569106	202	2773	232	0,713414634	351	0	0	0	0	0	0	132	591,149997
16	Τετάρτη, 24 Ιανουάριος 2018	0,196969697	416	5,076923077	2112	1473	0,442307692	184	2178	200	0,634615385	264	0	0	0	0	0	0	34	553,150001
17	Πέμπτη, 25 Ιανουάριος 2018	0,166117065	403	6,019851117	2426	1634	0,382133995	154	2481	224	0,70471464	284	0	0	0	0	0	0	122	1267,80001
18	Παρασκευή, 26 Ιανουάριος 2018	0,173629243	399	5,759398496	2298	1570	0,428571429	171	2346	216	0,671679198	268	0	0	0	0	0	0	42	367,91
19	Δευτέρα, 29 Ιανουάριος 2018	0,163128096	461	6,130151844	2826	1899	0,449023861	207	2882	244	0,678958785	313	0	0	0	0	0	0	125	586,640004
20	Τρίτη, 30 Ιανουάριος 2018	0,201053135	420	4,973809524	2089	1455	0,461904762	194	2117	230	0,685714286	288	0	0	0	0	0	0	151	745,159995
21	Τετάρτη, 31 Ιανουάριος 2018	0,169085916	431	5,914153132	2549	1754	0,412993039	178	2575	236	0,693735499	299	0	0	0	0	0	0	25	244,650001
22	Πέμπτη, 1 Φεβρουάριος 2018	0,168150685	491	5,947046843	2920	2026	0,374745418	184	2995	224	0,67413442	331	0	0	0	0	0	0	198	1051,87
23	Παρασκευή, 2 Φεβρουάριος 2018	0,172580645	428	5,794392523	2480	1665	0,488317757	209	2512	226	0,679906542	291	0	0	0	0	0	0	82	451,419998
24	Δευτέρα, 5 Φεβρουάριος 2018	0,211796247	395	4,721518987	1865	1328	0,448101266	177	1894	213	0,729113924	288	0	0	0	0	0	0	44	193,300004
25	Τρίτη, 6 Φεβρουάριος 2018	0,195515067	558	5,114695341	2854	2054	0,44265233	247	2933	224	0,663082437	370	0	0	0	0	0	0	75	733,150001
26	Τετάρτη, 7 Φεβρουάριος 2018	0,166536965	428	6,004672897	2570	1729	0,411214953	176	2600	187	0,63317757	271	0	0	0	0	0	0	74	1736,95002
27	Πέμπτη, 8 Φεβρουάριος 2018	0,196232339	375	5,096	1911	1312	0,448	168	1936	201	0,661333333	248	4	1,09	0,2625	105	0,4	10	57	228,469998
28	Δευτέρα, 12 Φεβρουάριος 2018	0,177808095	391	5,624040921	2199	1481	0,498721228	195	2226	219	0,680306905	266	3	0,5	0,166666667	25	0,15	20	98	835,899995
29	Τρίτη, 13 Φεβρουάριος 2018	0,171875	407	5,818181818	2368	1593	0,427518428	174	2387	219	0,685503686	279	4	1,06	0,265	42,4	0,16	25	67	416,83
30	Τετάρτη, 14 Φεβρουάριος 2018	0,224252492	405	4,459259259	1806	1258	0,432098765	175	1843	180	0,656790123	266	62	11,36	0,183225806	13,302108	0,0726	854	61	460,009997
31	Πέμπτη, 15 Φεβρουάριος 2018	0,199805068	410	5,004878049	2052	1405	0,5	205	2085	190	0,670731707	275	46	7,92	0,172173913	7,3265495	0,042553	1081	7	357,959999
32	Παρασκευή, 16 Φεβρουάριος 2018	0,202116935	401	4,947630923	1984	1365	0,548628429	220	2002	161	0,72319202	290	40	7,22	0,1805	24,066667	0,133333	300	11	151,090003
33	Τρίτη, 20 Φεβρουάριος 2018	0,173005722	514	5,780155642	2971	1977	0,468871595	241	3006	206	0,663424125	341	73	13,76	0,188493151	7,00611	0,037169	1964	71	371,579999
34	Τετάρτη, 21 Φεβρουάριος 2018	0,181612903	563	5,506216696	3100	2045	0,451154529	254	3165	245	0,635879218	358	99	16,83	0,17	3,7491646	0,022054	4489	70	618,929998
35	Πέμπτη, 22 Φεβρουάριος 2018	0,158125	506	6,324110672	3200	2113	0,4743083	240	3238	239	0,683794466	346	78	14,89	0,190897436	2,937463	0,015388	5069	29	300,159999
36	Παρασκευή, 23 Φεβρουάριος 2018	0,201977938	531	4,951035782	2629	1811	0,467043315	248	2651	221	0,625235405	332	54	9,98	0,184814815	1,6903794	0,009146	5904	90	519,949999
37	Δευτέρα, 26 Φεβρουάριος 2018	0,201774691	523	4,956022945	2592	1833	0,495219885	259	2624	225	0,665391969	348	82	15,04	0,183414634	2,0929585	0,011411	7186	75	589,439987
38	Τρίτη, 27 Φεβρουάριος 2018	0,173366834	483	5,768115942	2786	1942	0,430641822	208	2862	229	0,691511387	334	61	10,72	0,175737705	1,7086388	0,009723	6274	15	563,94

Picture 9: View of the exported 2018_2019_final.xlsx file

The above dataset used by Rapidminer in order to construct the predictive model.

4.3.3 Feature Extraction

The main goal of feature extraction is to produce and provide firm and ideally simple relationships between input features and the output features for the supervised algorithm to model. In order to generate features, we used all the information that were produced via descriptive analysis described in the previous chapter. The features that were created from the time series dataset and were used in this case study can be divided in three categories based on the type of feature. The categories are:

1. Time-based Features: individual components for each time step observation.
2. Lag Features: values at prior time steps.
3. Window Features: summary of values over a fixed window of prior time steps.

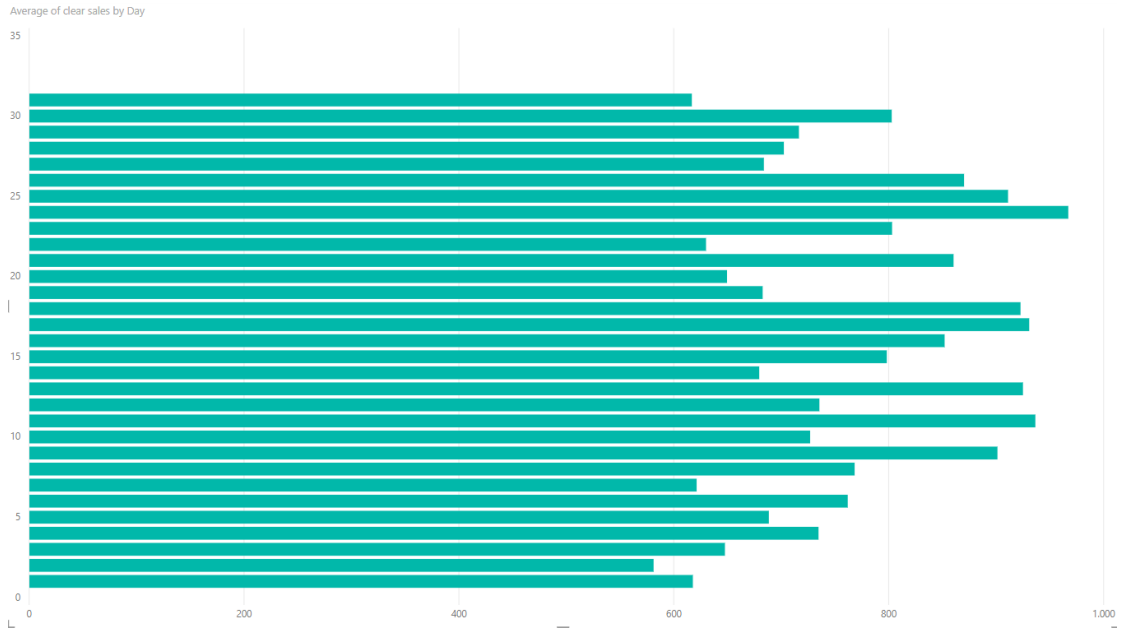
Time-based Features

These features derive from the date of each observation. From the line graph below is observed that the clear sales daily average per month has the following format:

Picture 10: Average of clear sales by month

Based on the line graph above the *month* feature is created. This feature assigns a number from 1 to 12 to each time step observation based on the month that belongs. Each month has its own number from 1 to 12 with 12 belonging to the month with the highest clear sales daily average and down.

In order to create the second Time-based feature, the following bar chart was used:



Picture 11: Average of clear sales by day

The bar chart above demonstrates the average of clear sales by day. In the same way with the previous feature, to each time step observation is assigned a number from 1 to 31 based on the day of the month that belongs to. Each day has its own number from 1 to 31 with 31 belonging to the day with the highest clear sales average by day and down.

Lag Features

These features are used in order to predict the value at time t, given the value at time t-i. In this case study are used the daily sales for the previous day, the same day of previous week and the same day of previous month. Additionally, the t-2 quantity is used as a feature.

Rolling Window Statistics

A step beyond adding simple lagged values is to add a summary of the values at previous time steps. In this kind of features, we calculated the mean values of attributes for previous time steps. In our case study we used the rolling window statistics for the sales, the bounces, unique pageviews and quantity. The different combinations and the time steps are displayed in the table below:

Table 3: Moving Average features

Average value (per day)	Previews time steps (in days)
Average items sold	10
Average bounces	5,10
Average clear sales	2,5
Average unique pageviews	5,10

4.3.4 Algorithms

The time-series forecasting problem of the case study is manipulated as a supervised learning problem, specifically as a regression problem in order to use various machine learning algorithms. Namely, the algorithms used are: Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees and Support Vector Machine.

The Generalized Linear model (GLM) is a step forward from traditional linear model. The algorithm allows for response variables that have error distribution models other than a normal distribution. GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value [28]. In Rapidminer the algorithm fits generalized linear models to the data by maximizing the log-likelihood. The model fitting computation is parallel, extremely fast, and scales extremely well for models with a limited number of predictors with non-zero coefficients.[29]

Deep Learning algorithm in Rapidminer uses H2O and is based on a multi-layer feed-forward artificial neural network that is trained with stochastic gradient descent using back-propagation. A large number of hidden layers consisting of neurons with tanh, rectifier and maxout activation can be comprised in the network. In order to produce high predictive accuracy advanced features such as adaptive learning rate, rate annealing, momentum training, dropout and L1 or L2 regularization are used. Every compute node trains a copy of the general model parameters on its local data asynchronously with multi-threading and provides to the general model through model averaging across the network. In order to perform regression H2O Deep Learning operator is used, predicting the numerical attribute of the dataset. The Deep Learning operator uses the adaptive learning rate option by default. The learning rate based on epsilon and rho parameters

automatically is detected by the algorithm. The hidden layer sizes is the only non-default parameter.[30]

In Rapidminer Studio Core, Support Vector Machine (SVM) Learner is based on the internal Java implementation of the mySVM by Stefan Rueping. This SVM learner provides a fast algorithm with good results for many regression and classification learning tasks. SVM operator supports various kernel types such as dot, radial, polynomial, neural, anova, apechenikov, gaussian combination and multiquadric. The standard SVM takes the input data and predicts for each value in which one among the two possible classes belongs, making the SVM a non-probabilistic binary linear classifier. The given training inputs are spotted as belonging to one of the two categories, as a result the training algorithm builds a model that adds new examples into one category or the other. SVM models try to present examples as points in space, graphed in a way that the examples of the different categories are separated by a clear gap being as distant as possible. The new examples entering the model then are graphed into that same space and predicted to belong to one of the two categories based on the side of the gap they placed.[31]

In a more formal way, SVMs create a hyperplane or a set of hyperplanes in a high or infinite-dimensional space that can be used both for classification and regression. Better separation can be achieved by the hyperplane that has the largest distance to the nearest training data points of any class, because of the fact that the larger the margin the lower the generalization error of the classifier. In many cases even if the problem started in a finite dimensional space, it is probable the sets to being discriminated are not linearly separable in that space. In that way it is proposed that the original finite-dimensional space be is graphed into a much higher-dimensional space, presumably making the separation easier in that space. The mapping used by the SVM schemes is formed to secure the easy computation of the dot products in terms of the variables in the original space, defining them in terms of a kernel function K(x,y) selected to suit the problem.[31]

Decision Trees are tree shaped collection of nodes that try to create a splitting decision on values connection to a class (classification) or an estimation of a numerical target value (regression). On decision trees each node is a splitting rule for one specific attribute. For regression these rules separate the attributes in order to reduce the error as much as possible for the selected parameter criterion. The generation of new nodes stops only when all the criteria are met. In regression problems the estimations for the numerical value are generated by averaging the values in a leaf. The decision tree operator in

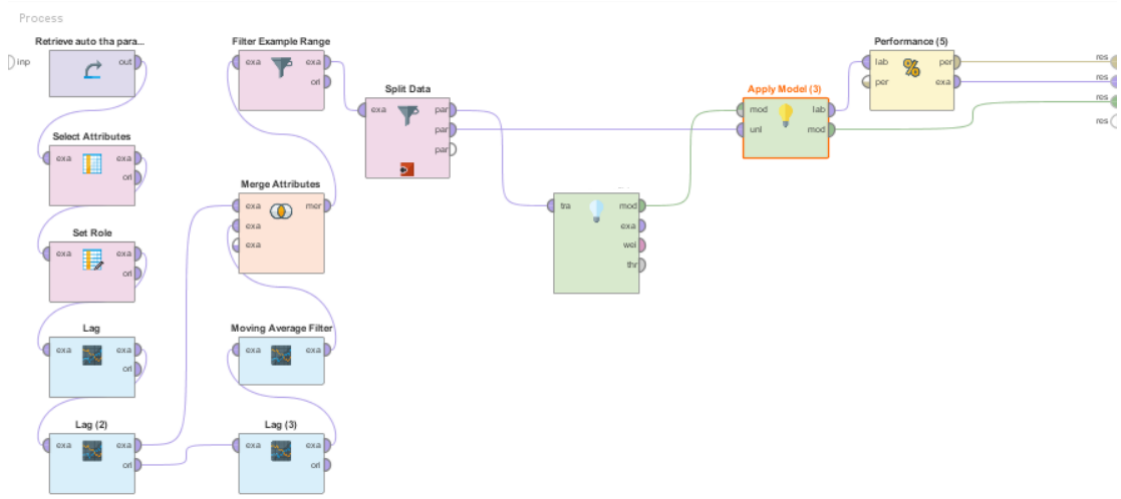
Rapidminer is able to manage both nominal (classification) and numerical (regression) values. Once the model generated, it can be applied to the new examples using an apply operator.[32]

In Rapidminer Random Forest is an aggregation of a number of random trees that can be specified by the number of trees parameter. The input example set trains the trees divided on bootstrapped subsets. As in decision trees, each node of the tree is a splitting rule for one specific attribute. In order to select the splitting rules only one subset of attributes is used and specified with the subset ratio criterion. Following the decision tree procedure in regression the rules separate the attributes to reduce the error as much as possible for the selected parameter criterion and the building of new nodes only stops when the stopping criteria are met. When the new examples applied to the model each random tree generates a prediction for each attribute. The resulting model is a voting model of all created random trees. In order to reduce the complexity of the model in many cases the pruning concept can be used. Pruning replaces sub-trees which provide little predictive power with leaves.[33]

Gradient Boosted Trees algorithm in Rapidminer is a collection of regression tree algorithms. Regression is a forward-learning ensemble method that generates predictive results via gradually improved estimations. Boosting, being a flexible nonlinear regression procedure, contributes into improving the prediction accuracy of trees. Gradient Boosting Trees method generalizes tree boosting in order to minimize the issue of the speed decrease and human interpretability.[34]

4.3.5 Model Process

For our case study prediction model generation Rapidminer is used. In order to construct the prediction model, it is needed to create a process that consists of different kind of operators. In our process data access, blending, transformation, filtering, modeling, scoring and performance operators are used. Based on the operators above we concluded to the final process of our model that has the following structure:



Picture 12: Predictive model process structure

The process begins with the first operator that is a retrieve operator. The retrieve operator retrieves the .xlsx file that was created in the preprocessing, reads it and inputs it in the software. The next operator is the select attributes operator that selects only the attributes that we will use in the process. In our case the attributes selected are the *average_bounces_10*, *average_bounces_5*, *average_unique_pageviews_10*, *average_unique_pageviews_5*, *average_clear_sales_2*, *average_clear_sales_10*, *clear_sales*, *day*, *date*, *month* and *quantity*. The first six are the rolling window features created in excel and will be used as they are. The other will be processed in next steps in order to be used. The next operator is the set role operator that is used in order to assign roles to attributes. *Clear sales* will become a label attribute; in other words the attribute that needs to be predicted, *day* and *month* will become weight attributes because the numbers that contain are weights based on the average clear sales. Moving forward there are three lag operators that assign t-i steps values to the t step. The first lag operator creates three features that will be used and they namely are the previous day sales, the same day of previous week sales and the same day in previous month sales. The next lag operator assigns the previous day's sold quantity in the current day and overwrites the current value in order to use it in next step. The last lag operator overwrites the t-2 quantity in the current day quantity. The next operator is a moving average filter operator that creates the last feature. The operator uses the output of the lag operator and creates the *average_quantity_10* feature.

Having all the features been created, a merge operator is used in order to merge them. Next step is the filtering examples step in which the filtering operator keeps only the attributes containing a value and discards all the empty ones. Subsequently the split operator is used in order to split the data in the training and the test set. The training set partition consists of the 99% percent of the dataset (384 examples) and the test set consists of the remaining 1% (4 examples). The split is performed using the linear sampling because of the problem's nature. Linear sampling simply divides the example set into partitions without changing the order of the examples. Next is the algorithm operator. In our case we used six different algorithms in order to find which trains the model in the most efficient way. Apply operator is the next operator used. The apply operator applies the trained model to the test set in order to create predictions for unseen data. The last operator of the process is the performance operator. In the present case is a regression performance operator used for the statistical performance evaluation of our regression task and delivers a list of performance criteria values of the regression task. In the present case the only criterion is the root mean square error.

5 Results

Designing the case study many problems were observed that led us to modify the initial plan. Beginning with the data collection, the retrieval of the Google Analytics data was the main problem. We came across the first obstacle when we tried to use products' SKUs. In many products the website used different SKUs than the ERP system; as a result we couldn't create a relationship between the datasets. Another problem we faced is that all the Google Ads campaigns concern product categories and not products separately. More specifically, the landing page of Ads is the page of a product category and not of a product itself. So many Google Ads data could not be used in our analysis. Moving forward, another obstacle we came across is that in Ecommerce view on Google Analytics all the data concerning the *medium* used by visitors to come to the case study's website where *nulls*. This problem cannot be explained once the *medium* information is available in Google Analytics but not able to be transferred in Power BI.

The following chapters, 5.1 and 5.2, explain the results of the descriptive and the predictive analysis conducted in this case study.

5.1 Descriptive Analysis Results

Descriptive analysis conducted with the assistance of Power BI concludes into two dynamic dashboards that illustrate the relationships between case study's data. Decomposing the visualizations, valuable insights are mined and used in the next step, predictive analysis.

Beginning with the first dashboard (picture 8), it was created based on sales data from both Google Analytics and ERP. The first and the most important observation presented in the line chart is the difference that exists between *clear value*, *product revenue* and *revenue.* Based on a literature research, I came to the conclusion that first of all the difference between *product revenue* and *revenue* results from the fact that the two metrics come from different part of code. The *product revenue* metric contains the taxes, the *revenue* metric that does not. The taxes' percentage is different based on the orders' geographical district, this is why the difference between the two metrics cannot be calculated exactly. This can be easily observed on the clustered column chart. Moving forward, on the same graph there is also a huge difference between *clear value* and the

revenue. The first one presents the actual daily clear sales based on the ERP system, while the *revenue* presents the daily clear sales based on the Google Analytics. Based on the knowledge coming from our experience working for the case study's company, this difference is a sequence of the cancelation of many of the orders that were placed through website. The cancelations were not be applied via website so they cannot be counted and demonstrated on the *revenue* metric.

Taking into consideration all the above on the remaining of our analysis we use only the sales values coming from ERP system dataset. On the pie chart (picture 8) is presented the daily average of *clear sales* value for each quarter of the year. From the pie chart can easily be observed that the average daily *clear sales* are decreased in the second, the third and the fourth quarter compared to the first. The *clear sales* value is presented also as percentage of the total yearly sales. On the clustered bar-chart the average value of *clear sales* per day of month is illustrated and colorized based on the average sales *quantity* for each day. From the clustered bar-chart it is clear that the last and the first day of the month have the lowest average *clear sales*. Additionaly, there are some peak days mainly in the middle of the month with the day with the most average *clear sales* being the 23rd of the month. Furthermore, looking at the colors of the bars we can observe that the more products the e-shop sold the more the sales value increased. This cannot be counted as a rule but drives to the conclusion that most of the products are of low value. Both pie chart and clustered bar chart results are used on the predictive analysis as attributes of the model. On the dashboard there are also two cards showing the total quantity and clear value of the products sold dynamically based on the date selection.

On the second dashboard (picture 9), we begin with the scatter chart that presents the *new users* and the *entrances* quantity for each available date colorized based on the *bounce rate* of the date. Through this chart can easily be observed that the more the new users the more the entrances and also the higher the bounce rate. On the pie chart the daily average *clear sales* for each month is presented. September, October and May are the months with the highest average. On the other hand, December, January and February are those with the lowest. The tree map presents the average *entrances* and *organic searches* quantity colorized based on the average of the *ads clicks*. On this graph it is detected that the *entrances* quantity depends more on the *organic searches* and less on the *ads clicks*. Taking into consideration the last observation, I presume that the company has a strong brand name with a lot of loyal customers. Finishing with the line and stacked column

chart that presents the *exit rate*, the *new sessions* percentage and the *ads cost* of each day of a month, one can see that the higher the *ads cost* the higher the *exit rate*. This sounds reasonable but it is also an alert for checking the quality of the active Ads.

The value of the two dynamic dashboards cannot only be extracted from a single chart but from the interaction between all the charts contained in them, as the one dynamically changes the other.

5.2 Predictive Analysis Results

Predictive analysis is conducted using Rapidminer. The goal of the analysis was to predict the sales amount for the next four days using historical data. Since the dataset involves the time component the problem defined as a time series forecasting problem. In this case study the problem tried to get solved as a supervised learning problem using various Machine Learning methods. In order to predict as accurate as it was possible the future sales values, the next six widely known algorithms are used: Generalized Linear, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees and Support Vector Machine algorithm. Considering that the value expected is a continuous quantity, the problem is transformed into a regression task. The evaluation of the performance is calculated by the root mean square error (RMSE). As mentioned in 3.2 RMSE represents the difference between the values predicted by the model and the values observed. The formula is given below:

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

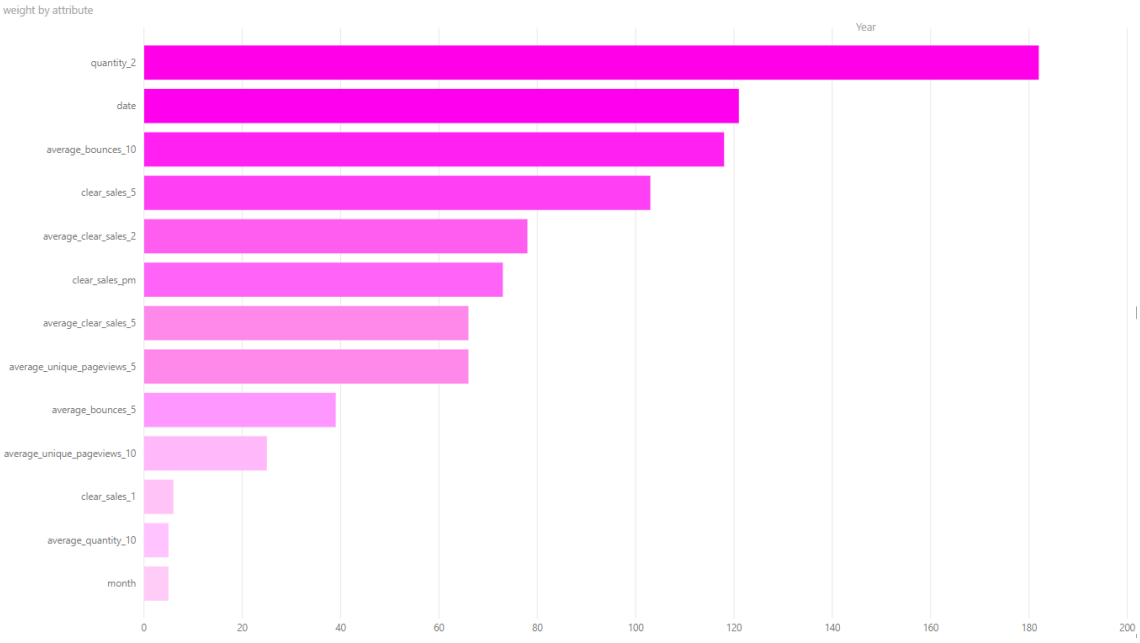
The predictive model is created using the Rapidminer operators. The operators used belong to data access, blending, transformation, filtering, modeling, scoring and performance categories. The created model was used six times with the six different algorithms and the RMSE evaluation method. The performance of each algorithm is summarized in the table below.

Table 4: Algorithm' RMSE results

Algorithm	RMSE
Generalized Linear Algorithm	476.526
Deep Learning	459.630
Decision Tree	**217.233**
Random Forest	392.384
Gradient Boosted Trees	551.532

Support Vector Machine	597.795

As it can be observed, Decision Tree outperformed the other five algorithms with next being the Random Forest algorithm and the one with the less accurate results being the Support Vector Machine algorithm. So, in the next bar-chart is presented the importance of the features according to the Decision Tree algorithm based on the weight of each one.



Picture 13: Attributes' weight

Considering the results displayed at the bar chart above the most important attribute on our model is the *quantity_2* which corresponds to the quantity of the products sold at the time point t-2. The next three important attributes with narrow deviation between them are the *date*, the *average_bounces_10* and the *clear_sales_5*. The first concerns the date of the transaction, the second is the moving average of the bounces for the previous ten days of the transaction and the third is the clear sales value on the same day of the previous week of the transaction. The two least important attributes for the created model, having the same weight are the *month* and the average_*quantity_10*, with the first presenting the month which the day of the transaction belongs and the latter presenting the moving average of the sold quantity for the ten previous days.

6 Conclusions and Future Work

In this thesis data analytics and more specifically descriptive and predictive analytics are discussed. Data analytics is not a new term but a constantly changing science field. Descriptive and predictive analysis are both widely used in most of organizations on their internal and external data. The conducted study of this thesis focuses on data analysis in the frame of a small wholesale and retail sporting goods company and specifically uses data retrieved from ERP system and Google Analytics.

The initial goal of this study, as the title indicates, was simply to visualize Google Analytics', Digital Marketing Campaigns' and ERP system's data using BI tools. However, in order for the research to be thorough I decided to expand its scope including also a predictive analysis.

Generally, through the study it is indicated that the internal data of any organization can generate precious knowledge that can be used in future decision taking. Notably, the descriptive analysis gave us information about the validity of the data collected via Google Analytics in comparison with ERP system data. In addition, it demonstrated the seasonality of the sales during a year and also suggested the most and the less busy days in terms of sales. Furthermore, using website's traffic data that determine relations between metrics such as bounces, entrances, ads cost etc. relative information were extracted. Specifically, in our case it was shown that the most products sold via website seem to have low value. Also, it became clear that the increase of the new users raises the bounce rate. Furthermore, the data indicate that the company has a strong brand name with a lot of loyal customers. Finishing, according to the study the quality of the Digital Marketing Campaigns' ads should be checked.

As far as predictive analysis is concerned it is crucial to be said that we used many of the results of the descriptive analysis. In this phase I forecasted as accurate as possible next four days sales amount using historical data collected through Google Analytics and ERP system, proving that those data can improve the accuracy of the results. Through the analysis a model was constructed, on which six different algorithms have been applied and evaluated using RMSE, with the Decision Tree being the best fitted in our problem. The accuracy of the predictions generated is promising and also encouraging for future work on the field.

The accuracy of the predictions is directly connected to the volume and the quality of the data used. In our case study the small volume of the available data affected the prediction results. Furthermore, the non-proper application of the Google Analytics on the website and also the non-proper use of the products' SKUs set up limitations on the data retrieval. Taking into consideration all the above I strongly believe that increasing the volume and improving the quality of the data used in the case study the results will become more accurate.

Concluding, I strongly believe that organizations need to be more conscious when implementing tools such as Google Analytics in order to be able to extract the maximum value of their internal data. Additionally, in my opinion it is vital for companies to use a unified database integrating data coming from various sources using products' SKUs as a primary key. As a researcher finishing this study, I am determined to do my very best in order to improve the predictions accuracy of the current case study and strongly motivated to keep studying the field.

Bibliography

[1] Shukla, S. (2016). *Study of big data analytics landscape: considerations for market entry of an E-commerce analytics vendor*. MSc Thesis. Massachusetts Institute of Technology. Available at: https://dspace.mit.edu/handle/1721.1/104515.

[2] Elragal, A. and Klischewski, R. (2017). Theory-driven or process-driven prediction? Epistemological challenges of big data analytics. *Journal of Big Data* 4(19). Available at: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0079-2#citeas.

[3] Sun, Z., Strang, K. and Firmin, S. (2017). Business Analytics-Based Enterprise Information Systems. *Journal of Computer Information Systems* 57(2), 169-178. Available at: https://www.tandfonline.com/doi/abs/10.1080/08874417.2016.1183977.

[4] Jeble, Sh., Kumari, S. and Patil, Y. (2016). Role of big data and predictive analytics. *International Journal of Automatation and Logistics* 2(4), 307-331. Available at: https://www.researchgate.net/publication/309809606_Role_of_big_data_and_predictive_analytics.

[5] Klisarova-Belcheva, S., Ilieva, G. and Yankova, T. (2017). *Trakia Journal of Sciences* 15(1), 298-304. Available at: http://tru.uni-sz.bg/tsj/TJS_Suppl.1_Vol.15_2017/53.pdf.

[6] Kaur Saggi, M. and Jain, S. (2018). A survey towards an integration of big data analytics to big insights for value-creation. *Information Processing and Management* 54(5), 758-790. Available at: https://www.researchgate.net/publication/323047768_A_survey_towards_an_integration_of_big_data_analytics_to_big_insights_for_value-creation.

[7] Canon Moreno, J. (2017). *Loading data analytics transformations*. MBA Thesis. Massachusetts Institute of Technology. Available at: https://dspace.mit.edu/bitstream/handle/1721.1/111472/1003321999-MIT.pdf?sequence=1&isAllowed=y.

[8] Huisman, D. (2015). To what extent do predictive, descriptive and prescriptive supply chain analytics affect organizational performance?. In: *5th IBA Bachelor Thesis Conference*. Enschede: University of Twente. Available at: https://essay.utwente.nl/67423/1/Huisman_BA_MB.pdf.

[9] Sedkaoui, S. (2018). How data analytics is changing entrepreneurial opportunities?. *International Journal of Innovation Science* 10(2), 274-294. Available at: https://www.emerald.com/insight/content/doi/10.1108/IJIS-09-2017-0092/full/html.

[10] Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. 1st edn. Hoboken-New Jersey: Wiley.

[11] Attaran, M. and Attaran, Sh. (2018). Opportunities and challenges of implementing predictive analytics for competitive advantage. *International Journal of Business Intelligence Research* 9(2), 64-91. Available at:

https://www.researchgate.net/profile/Mohsen_Attaran/publication/325934828_Opportunities_and_Challenges_of_Implementing_Predictive_Analytics_for_Competitive_Advantage/links/5b5f46c3aca272a2d67558ef/Opportunities-and-Challenges-of-Implementing-Predictive-Analytics-for-Competitive-Advantage.pdf.

[12] Fhyzics Business Consultants. *Process and its Applications.* Available at: http://fhyzics.com/predictive-analytics-process-and-its-applications.html.

[13] Predictive Analytics Today Research. *What is Predictive Analytics?*. Available at: https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/.

[14] Analytics Vidhya. *7 Regression Techniques you should know!*. Available at: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/.

[15] Freedman, D. (2009). Statistical models: *Theory and practice*. 2nd edn. Cambridge: Cambridge University Press.

[16] Wehle. H.D. (2017). Machine learning, deep learning and AI: What's the difference?. Available at: https://www.researchgate.net/publication/318900216_Machine_Learning_Deep_Learning_and_AI_What%27s_the_Difference.

[17] Uchechukwu,N.G. (2016). *Comparison of cloud machine learning services*. MA Thesis. Universitet i Stavanger. Available at: https://uis.brage.unit.no/uis-xmlui/bitstream/handle/11250/2413901/Nketah_Gabriel.pdf?sequence=1&isAllowed=y.

[18] Emerson, S., Kennedy, R., O'Shea, L. and O'Brien, J. (2019). Trends and applications of machine learning in quantitative finance. In: *8th International Conference on Economics and Finance Research (ICEFR 2019)*. Lyon. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3397005.

[19] Mohammed, M., Khan, M. and Bashier, E. (2016). Decision trees, in: *Machine Learning Algorithms and Applications*. 1st edn. Boca Raton: CRC Press.

[20] James, C. (2018). *Artificial intelligence in marketing*. MA Thesis. Arizona State University. Available at: http://www.jamescannella.com/wp-content/uploads/2018/04/Cannella_J_Spring_2018.pdf.

[21] Shakti, S., Hassan, M., Zhenning, Y., Caytiles, R. and Lyenger N. (2017). Annual automobile sales prediction using ARIMA model. *International Journal of Hybrid Information Technology* 10(6), 13-22. Available at: https://www.researchgate.net/profile/N_Ch_Sriman_Narayana_Iyenger/publication/319061523_Annual_Automobile_Sales_Prediction_Using_ARIMA_Model/links/5b2899110f7e9b332a31eaba/Annual-Automobile-Sales-Prediction-Using-ARIMA-Model.pdf.

[22] Wang, X. Lin, D., Fan, W. and Wang, T. (2018). Research on sales forecast of fresh produce considering weather factors. In: *Proceedings of the 18th International Conference on Electronic Business (ICEB)*. Guilin.

[23] Norris, D. (2013). *RapidMiner – a potential game changer*. Available at: https://www.bloorresearch.com/2013/11/rapidminer-a-potential-game-changer/.

[24] Idoine, C., Krensky, P., Brethenoux, E. and Linden, A. (2019). *Magic Quadrant for Data Science and Machine Learning Platforms*. Available at: https://www.gartner.com/doc/reprints?id=1-65WC0O1&ct=190128&st=sb.

[25] Gualtieri, M., Carlsson, K., Sridharan, S. Perdoni, R. and Yunus, A. (2018). *The Forrester Wave™: Multimodal predictive analytics and machine learning solutions, Q3 2018*. Available at: https://www.forrester.com/report/The+Forrester+Wave+Multimodal+Predictive+Analytics+And+Machine+Learning+Solutions+Q3+2018/-/E-RES141374#.

[26] Rad, R. (2018). *Pro power BI architecture: Sharing, security and deployment options for microsoft power BI solutions.* Berkeley: A press.

[27] Deardorff, A. (2016). Tableau (version 9.1). 837 North 34th Street, Suite 200, Seattle, WA 98103: Tableau. info.@tableau.com; http://www.tableau.com; free and paid versions available [Review of Tableau (version 9.1)]. *Journal of the Medical Library Association* 104(2), 182-183. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4816475/.

[28] Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* 135(3), 370-384. Available at: https://www.jstor.org/stable/2344614?seq=1#metadata_info_tab_contents.

[29] Rapidminer. *Generalized linear model (H20).* Available at: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/functions/generalized_linear_model.html.

[30] Rapidminer. *Deep learning (H20).* Available at: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/neural_nets/deep_learning.html.

[31] Rapidminer. *Support Vector Machine* (RapidMiner studio core). Available at: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/support_vector_machines/support_vector_machine.html.

[32] Rapidminer. *Decision tree (Concurrency).* Available at: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_decision_tree.html.

[33] Rapidminer. *Random forest (Synopsis).* Available at: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_random_forest.html.

Pictures

Picture 1: https://www.log-hub.com/predictive-analytics/

Picture 2: https://www.kdnuggets.com/2019/02/gartner-2019-mq-data-science-machine-learning-changes.html

Picture 3: https://info.microsoft.com/ww-landing-gartner-mq-bi-analytics-2019.html?LCID=EN-US