OULU UNIVERSITY OF
APPLIED SCIENCES

Bogdanova Mariia

FINTECH UNDERWRITING USING MACHINE LEARNING

FINTECH UNDERWRITING USING MACHINE LEARNING

Mariia Bogdanova
Bachelor's Thesis
Autumn 2019
Degree Programme in Information Technology
Oulu University of Applied Sciences

ABSTRACT

Oulu University of Applied Sciences
Degree Programme in Information Technology

Author: Bogdanova Mariia
Title of the bachelor's thesis: Fintech Underwriting using Machine Learning
Supervisor: Lasse Haverinen
Term and year of completion: fall semester 2019 Number of pages: 84

This thesis aimed to research and develop a FinTech Underwriting solution and to prove that FinTech Underwriting solution using machine learning is possible to create. This thesis was commissioned by the company WeBuust Oy.

The main objective was to prove that it is possible to create a rating system using machine learning that will check the potential of the startups to become successful so that the potential investor can determine if the startup is worth investing into and to prototype and implement the system like this. The company is aiming to develop a platform that will connect the investors and startups, and this FinTech Underwriting system will be implemented within the services of the company as a feature. Supervised and unsupervised machine learning algorithms are used to prove that developing this system is possible and to achieve the best results and the best approach is selected.

As the result of this study it was proven that it is possible to develop FinTech Underwriting solution using machine learning. The best implementation approaches were selected.

Keywords: FinTech, Machine Learning, MATLAB, Python, Startup

PREFACE

This thesis was written in Turin, Italy and Oulu, Finland from fall 2018 till fall 2019 semesters. It was commissioned by the company WeBuust Oy.

It was written with the support of the thesis supervisor Lasse Haverinen and the company representatives Iikka Meriläinen, CTO of WeBuust, and Janne Lauanne, CEO of WeBuust.

I would like to say thank you to everyone who supported me during my research.

Oulu, Finland, October 14th 2019
Mariia Bogdanova

CONTENTS

1 INTRODUCTION

1.1 Goals and Objectives of the work

The idea of this thesis comes from the need for the rating estimation system of startups by the company WeBuust Oy (figure 1).

The main idea of this thesis is developing a system that is based on machine learning. The main functionality of this system must be to compute and estimate how successful the startup has the potential to be so that the investors can decide if they are interested in investing into this project or not. The main reason for this is the fact that this decision is made based on the estimation if the project that received investment would be able to return the funds.

FIGURE 1. WeBuust Oy logo

The company WeBuust Oy is interested in including this system into its services. The company aims to create a platform for communication between startups and investors, for both sides to benefit from this connection. The rating system will potentially be included in the platform as a feature. The prototype and implementation developed during this thesis will be a solid base for this feature. However, the actual feature will include more correlations with the business plan of the startup and the financial abilities of the investor. This feature will motivate the startup to look for a beneficial business idea and it will give the investor an overview of what to expect.

The main objective of this thesis was to create a proof of concept that developing the FinTech Underwriting system using machine learning is possible. One of the goals is to investigate how to implement this system and to find the best approach for the development process. The proof of concept needs to include the investigation if it is possible to create the mathematical relational model that afterward will make the rating decision autonomously. The machine learning system needs to rate the startups based on their success potential. That can be achieved by developing a prototype and potentially implementing this system, which is the secondary objective.

2 FINTECH

2.1 Definition

FinTech, which is short for Financial Technology, is the technical area that is based on implementing the modern, cutting edge technologies to the traditional methods that are used in financial services. FinTech is currently a fast-growing and well-developing area that aims to improve activities and operations in the finance sector (Schüffel, 2016). Many companies are operating in this area nowadays (figure 2).

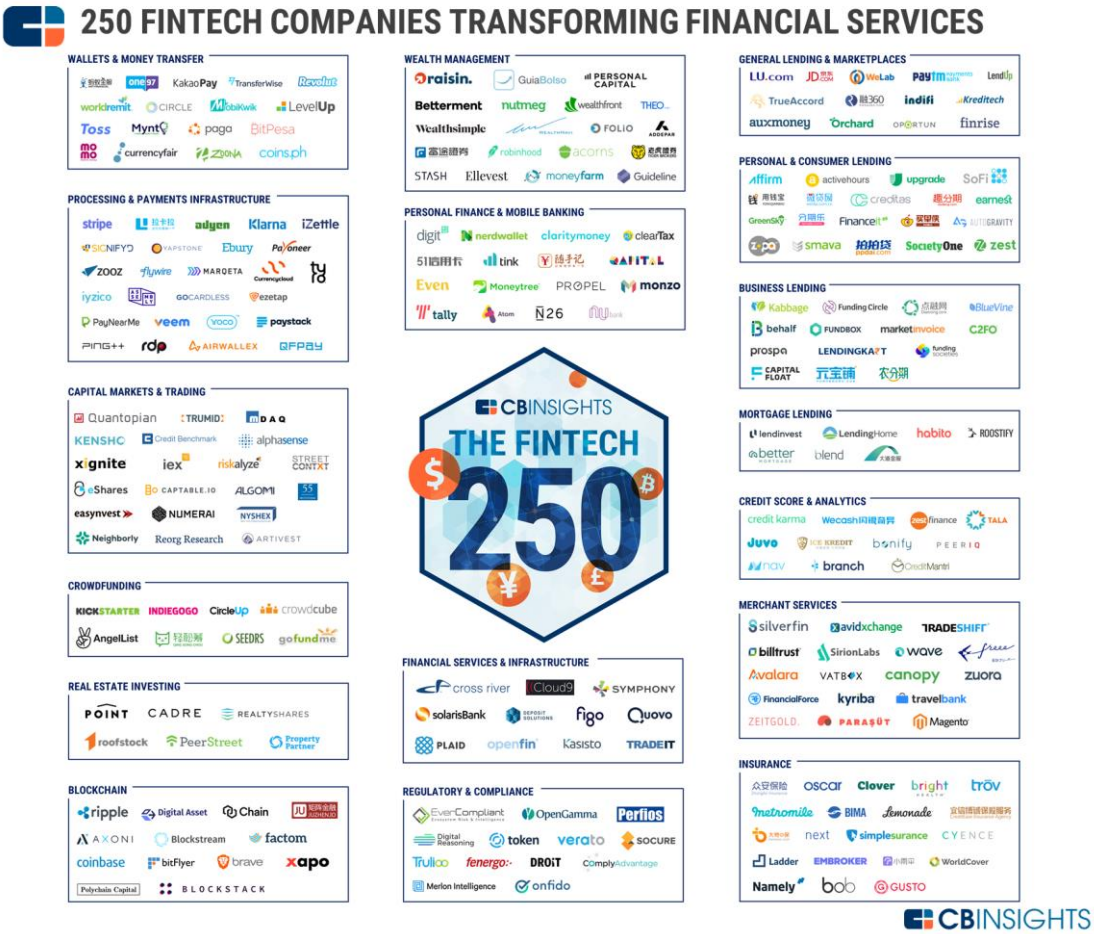

FIGURE 2. 250 FinTech companies transforming financial services by CB Insights

FinTech was formed as an area and a term during the 21st century, however, it has been functioning and shaping the financial sector for a long time. In the 1950s the credit cards went into the common use, in the 1960s cashiers were replaced by ATMs. FinTech has shaped financial services as they are known nowadays, and this industry is still moving forward in many areas, automating and digitalizing them (Zigurat, 2017). Further, there is a description of the popular FinTech areas at the end of the year 2018 and the beginning of the year 2019 (Sullival, 2018).

2.2 Fast-developing FinTech areas

Mobile payments

The payment system has been transforming through the history of the financial industry. Between 2016 and 2021 the number of payments conducted using mobile devices will increase tenfold. The amounts transferred using mobile devices reached $275 billion, and it is expected to increase tenfold between 2016 and 2021 (Scott-Briggs, 2018). This industry depends heavily on research and development in many technical areas, for example, biometrics and security.

InsureTech

InsureTech is short for insurance technologies, and it is an area that is digitalizing insurance (Hargrave, 2019). Insurance has always been an important part of the financial industry. Nowadays FinTech allows to make it more convenient and simple using the calculations pay-per-mile for car insurance rate, on-demand protection and other technologies to satisfy the need of the customer to be insured and safe (Baumann, 2018).

Blockchain

Blockchain is the technology that lies behind cryptocurrencies (Narayanan, Bonneau, Felten, Miller & Goldfeder, 2016). It provides peer-to-peer communication and data transfer. This technology is safe and fast, and it allows to avoid the information to be altered by the means of the cryptographic hash. Many banks are currently developing blockchain-based solutions. However, money-related matters are not the only implementation of the blockchain. Nowadays it is very successfully used for such technologies as Smart Contracts, distributed ledger, and video games (Ovenden, 2017).

Biometrics

Biometrics is the technology that is implemented in various FinTech fields as one of the digitalization tools. It is used to prove the identity of the person, authenticate, ensure security and prevent possible fraud. It is utilized in banking applications, security matters and the protection of personal data. The development and research in this area is very important and moves many other industries forward (Goel, 2015).

Data analytics

The data analysis became a major topic in the financial industry because it allows managing, sorting, comparing and predicting circumstances and identifying the causes of the changes both in the whole financial sector and in the separate site of the specific company (Finance Train, 2018). Collecting and analyzing data makes such actions as, for example, investing and underwriting better reasoned and it allows to get more profit in the result.

2.3 Underwriting as an area of FinTech

Underwriting is a background check of the person or company before the checker takes the financial risk. Underwriting is used before applying for the mortgage, car loan or other finance concerning matters that involve risk and a certain level of trust from another party. The term originated in London, where the oldest operating insurance company in the world used to base. It was called “Lloyd’s of London” (Wertheimer, 2006) (figure 3).

FIGURE 3. Lloyd's underwriting room (Wertheimer, 2006)

The term underwriting was introduced from the word underwriter. Underwriter was the person who signed the contract under the line on the bottom. By this action, they would ensure the contract with their guarantee and this way they would be responsible in case of an accident (Herbon, 2012). In the modern days, the meaning has transformed but the idea remains the same: the underwriting process helps in making a decision that involves risk. It is used for checking the outcome of the action and ensuring the best result.

FinTech has made major changes to the underwriting, combining it with the innovative technologies and transforming it into an automated tool that does not include manual work of the person (Eishawa, 2017).

2.4 Machine Learning as FinTech technology

Machine learning is the scientific field that is exploring the abilities of machines to be autonomous and make decisions based on the specific algorithm. It is a very popular technology which is used in a wide variety of areas, starting from medicine to the financial sector. It is already playing a great role in the FinTech services and continues to grow and expand. Further, there is a description of the areas of FinTech that machine learning is implemented in (Goel, 2015).

2.4.1 Machine Learning in FinTech areas

Algorithmic trading

Algorithmic trading is a technology used for making the financial decision into an automated process and increasing trades. At the moment 73% of all the trading performed is done by the machines, using the predictive nature of the machine learning technology. Many of these actions cannot be replicated manually. Machine learning can analyze large datasets, and that allows to predict the variety of market and stock market changes, avoiding risks, identifying opportunities and predicting changes.

Fraud prevention and detection

The approach to fraud prevention and detection based on machine learning has been gaining popularity. The technology is used for finding forbidden activities, reducing the number of verification measures, anomaly detecting and dealing with other matters related to the security (Feedzai, 2017). This area is involved in banking and insurance-related concerns.

Customer support

Machine learning is used for increasing the satisfaction of the customer support experience. Chatbots is one of the examples of this technology being implemented. They are popular nowadays and increase the amount of good reviews (Clayton, 2019). Personalization of this experience can also be achieved by using machine learning.

2.5 FinTech Underwriting using Machine Learning

Nowadays underwriting is executed using human manual work and analytical skills. This approach, even though it has been tested for many decades, might be inaccurate. Using machine learning for the FinTech underwriting makes it possible to create a model that would estimate the risks interconnected by the specific financial activities, for example giving a loan, investing or giving a grant. It allows it to be done automatically, without including human work into the equation.

Automating this process comes with many advantages. Machine learning makes the whole process faster, solving the issue within milliseconds, seconds or, in the worst-case scenario, minutes. On the other side, the human would take hours of studying background information and making assumptions. Machine learning makes underwriting more precise because it is only focusing on the sections with significant information, while the human action, which is imperfect in many situations, can be unreliable. Machine learning would also expand the access to the credit for the larger population and by lowering the cost of the background check it might aid to reduce the default rate (Do, 2017).

3 MACHINE LEARNING

3.1 History and definition

The term “Machine Learning” dates back to 1959 when it was first used by Arthur Samuel while he was working at IBM (Kohavi & Provost, 1998) (figure 4).

FIGURE 4. IBM logo

Arthur Samuel (05.12.1901 - 29.07.1990) was one of the first developers and researchers into computer gaming and Artificial Intelligence in the USA, and during his career, he impacted the early development and definition of machine learning as a separate area (McCarthy & Feigenbaum, 1990) (figure 5).

FIGURE 5. Arthur Samuel at IBM (McCarthy, Feigenbaum & Feigenbaum, 1990)

The official definition phrased by Arthur Samuel states that machine learning is: “Field of study that gives computers the ability to learn without being explicitly programmed“ (Ng, 2017).

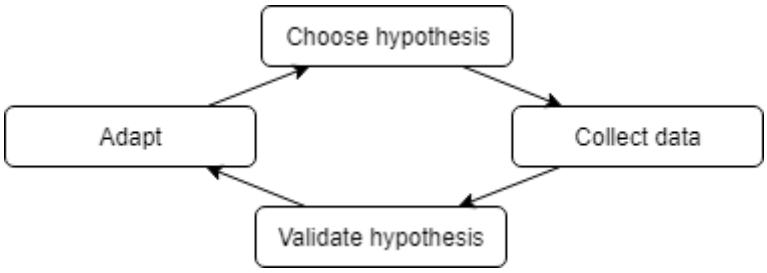

FIGURE 6. The main processes in machine learning (Jung, 2018)

Another definition is presented by Tom Mitchell (born August 9, 1951), an American computer scientist: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”. This definition is visualized in figure 6 (Jung, 2018).

Machine learning is an application of Artificial Intelligence, and it developed from the studies in this area when the researchers faced the need and became interested in having machines learn from the previously collected data (Ng, 2017).

The theory of machine learning and the development and design of the algorithms are included into its own subfield of Artificial Intelligence, called a computational learning theory. It concentrates on getting a learner (a machine) to generalize the data received from previous experiences in order to be able to make accurate decisions when receiving new unseen data (Association for Computational Learning).

3.2 Methods and approaches

There are many approaches to machine learning. The main ones can be defined as supervised learning, unsupervised learning and reinforcement learning (Ng, 2017) (figure 7). Further, there is a description of these approaches, their workflow and main differences.

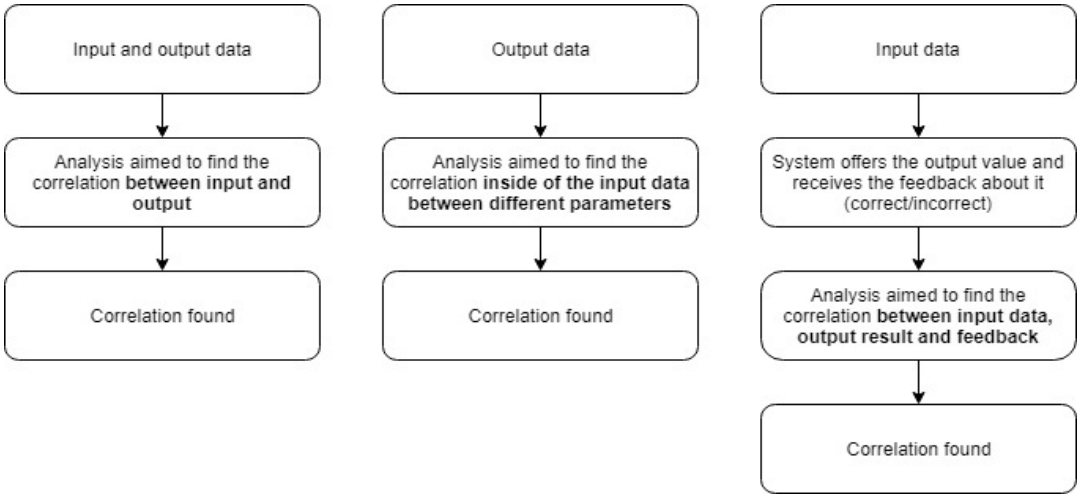

FIGURE 7. Main differences between machine learning techniques

3.2.1 Supervised learning

Supervised learning is a type of machine learning algorithm where the model is trained based on the input data with the known output. The model creates a mathematical equation that defines the relation between the input and the output provided in the training set. This way, after being trained, the model can predict the output based on the input, using the mathematical equation that the model was trained on (Rouse, 2016). The main examples of these algorithms are classification and regression. Classification is used for classifying or categorizing data based on features and output. The difference is demonstrated in figure 8. Regression is based on the estimation of relations between the input features and the output (Ng, 2017).

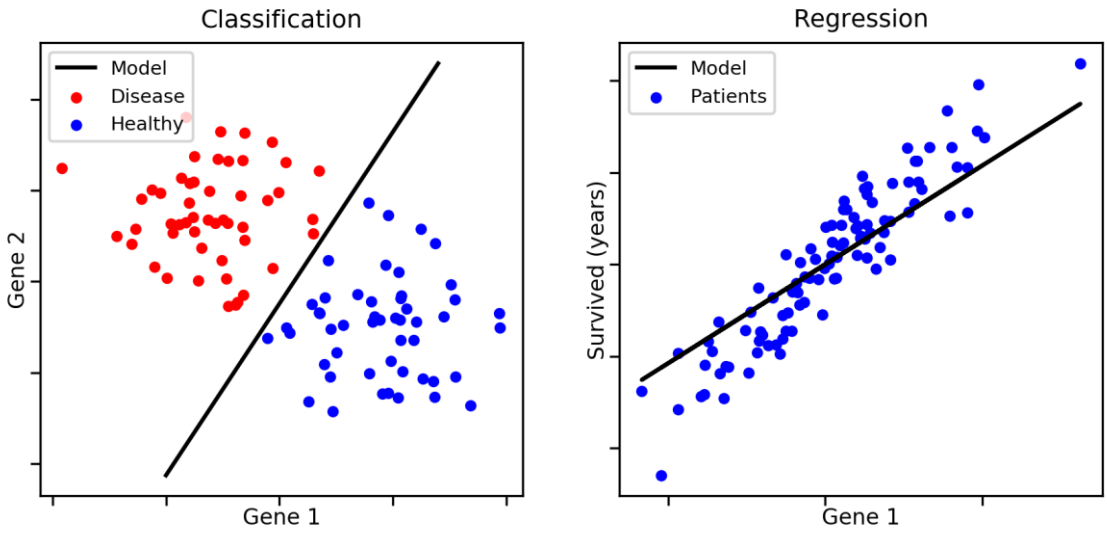

FIGURE 8. Comparison of classification and regression (Ram, 2018)

3.2.1.1 Neural networks

Neural networks are an example of an advanced supervised learning algorithm. It is modelled in the way the human brain functions, and modern research in Artificial Intelligence and deep learning

is based on this concept (DeMuro, 2018). An artificial neural network is a system of algorithms and it is inspired by and simulates the neural network of the human brain. It replicates the thinking process of a person for the complicated actions to gain the ability to be programmed. It is achieved by the sensory data being represented in the machine-readable format and analyzed. This technology is used as a part of deep learning mechanisms and it is implemented in many areas, for example, computer vision, speech recognition, medical imaging, translations. Similarly to the human brain, the neural network system consists of neurons (separate nodes) and the connections that pass signals between them. The neurons process them and transfer them further. The signal that is being passed is represented by the real number, so the neurons perform computational actions on the signal until the result is obtained. There can be many layers of the neurons between the input and output.

3.2.1.2 Training technique

The model is trained using supervised machine learning by first receiving the input and the output data from the dataset. So the model has the data about both conditions of the event, which are called features or parameters and the outcome. The problem is the following: it is unknown how valuable and weighty each of those features are. It means that one feature can affect the outcome significantly, while the other one can have a zero influence. There are many approaches to calculating the coefficients, for example, the normal equation that uses the derivative. Another option is a gradient descent with the cost function that uses minor changes in values in the desired direction for as long as it is needed until the goal is reached. It is the most common and popular approach (Ng, 2017).

3.2.2 Unsupervised learning

Unsupervised learning is a type of machine learning algorithm that is using only the input features for training, without any data about what the output value can be. There are two types of

unsupervised learning algorithms: based on clustering and not based on clustering, which is the classification of the data points into separate groups (clusters) (figure 9). The algorithms that are not based on clustering are called “Cocktail party algorithms” (Ng, 2017).

3.2.2.1 Training technique

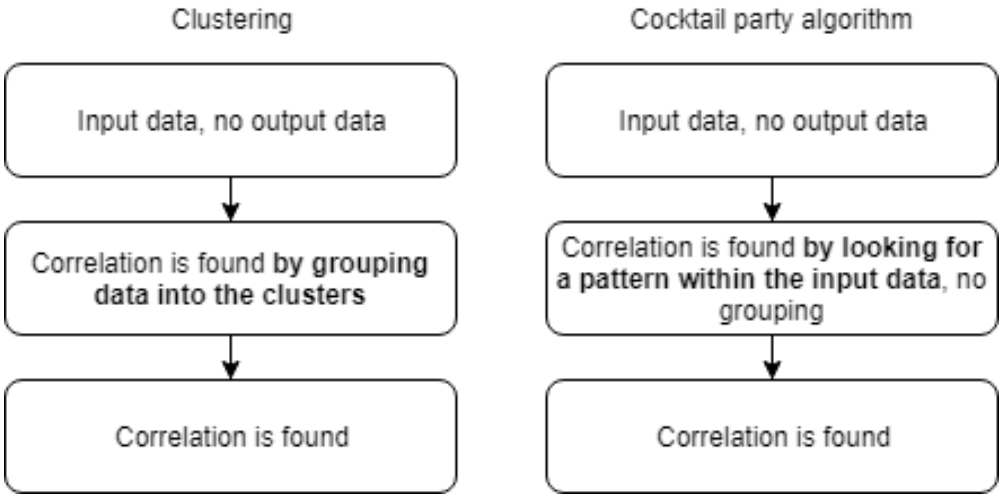

FIGURE 9. Main differences between unsupervised learning algorithms

Clustering is a set of algorithms which create a mathematical equation that groups the data into sectors, which are called clusters. Those clusters are created based on some feature or group of features (figure 10).

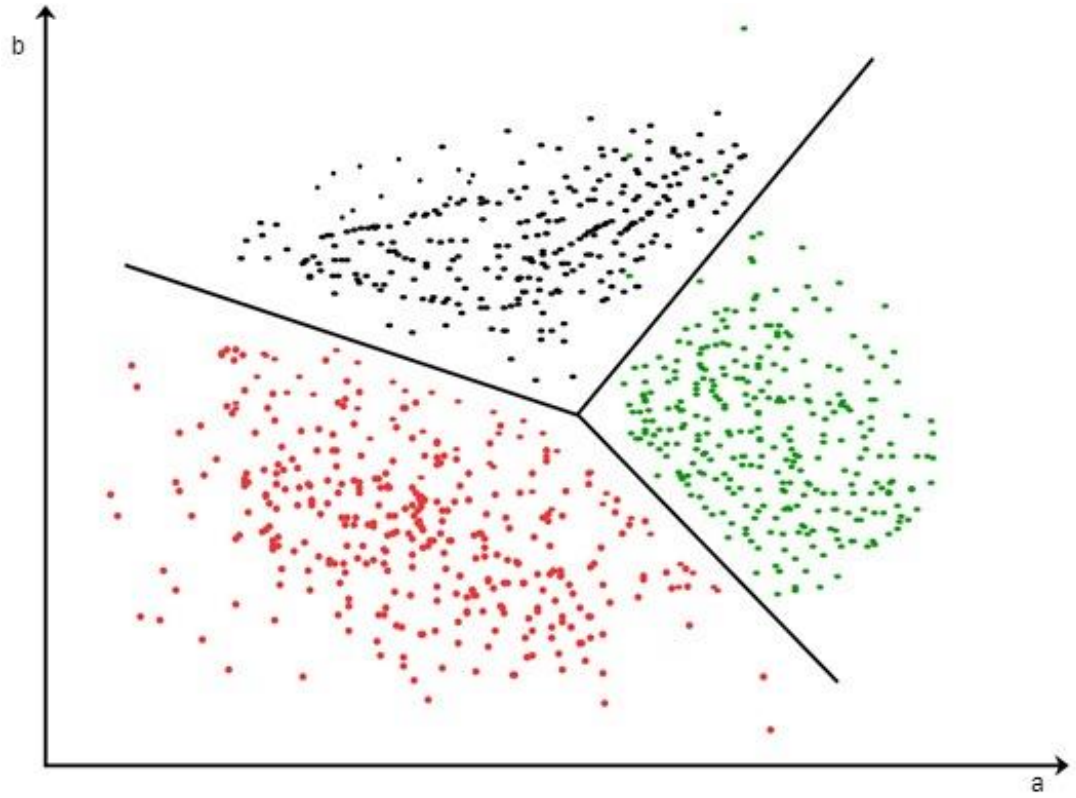

FIGURE 10. Clustering example, where axes a and b are features (Surya, 2017)

Many algorithms are based on this technique, for example, K-Means, Mean-Shift and Density-Based Spatial Clustering of Applications with Noise, but all of them rely on classifying the dataset based on features, not the outcome (Huneycutt, & Seif, 2018).

Non-clustering algorithms are based on the “cocktail party problem” when different techniques are used to differentiate sounds of specific voices, music or other noise in the chaos of the party. Usually, during such chaos, human brain functions in that way that an individual can still differentiate one sound they are concentrating on, for example, the voice of the person they are communicating with. Similarly to that, non-clustering algorithms are based on finding the hidden mathematical concept of the relationships between input values without taking the output into an account, and then making predictions or guesses based on that (Ng, 2017) (figure 11).

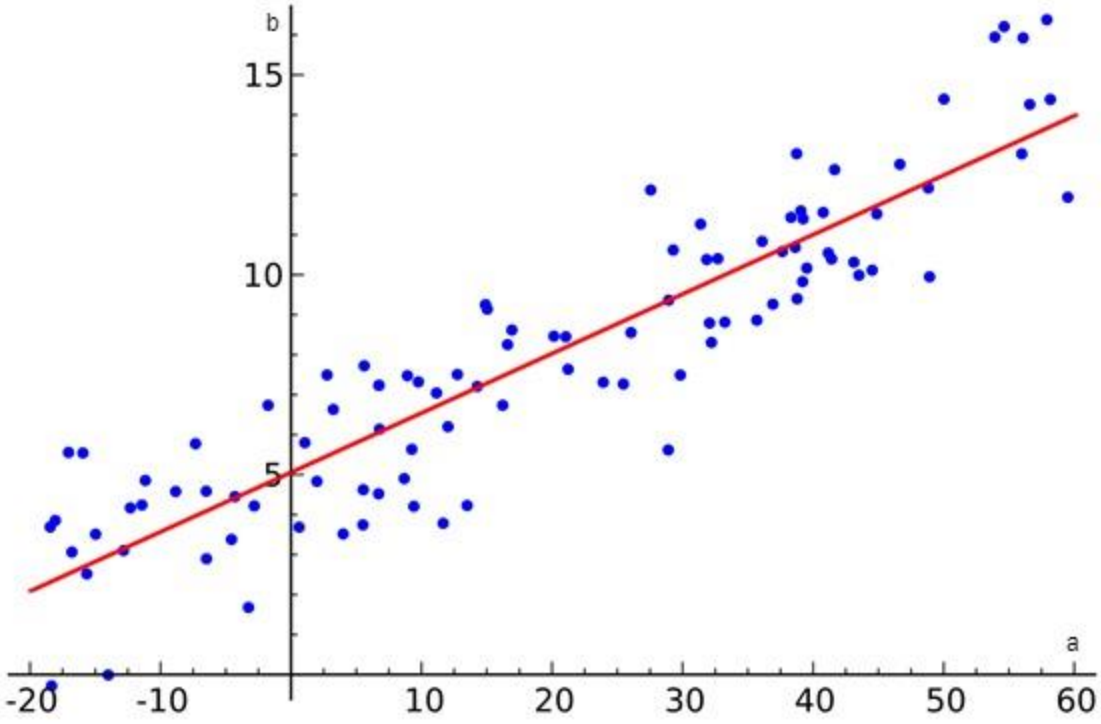

FIGURE 11. Example of the cocktail party algorithm, where axes a and b are features (Swaminathan, 2018)

3.2.3 Reinforcement learning

Reinforcement learning is a different approach to machine learning. It is related to simple approval or disapproval actions. The training technique is based on the system making a decision and providing an output from the input data. After that, the output is checked and rated as correct or incorrect. From that, the system learns about the correlations and changes that need to be made to give a correct output next time (van Otterlo, Wiering, 2012).

3.3 Relation of machine learning to the other areas

Machine Learning is the field that is based on many other areas and related to them as well as uses its tools. Historically, it came from Computer Science and Artificial Intelligence but it is more related to mathematics and statistics at its core. The relation to some of the other fields is specified in figure 12 (University of Helsinki, 2018).

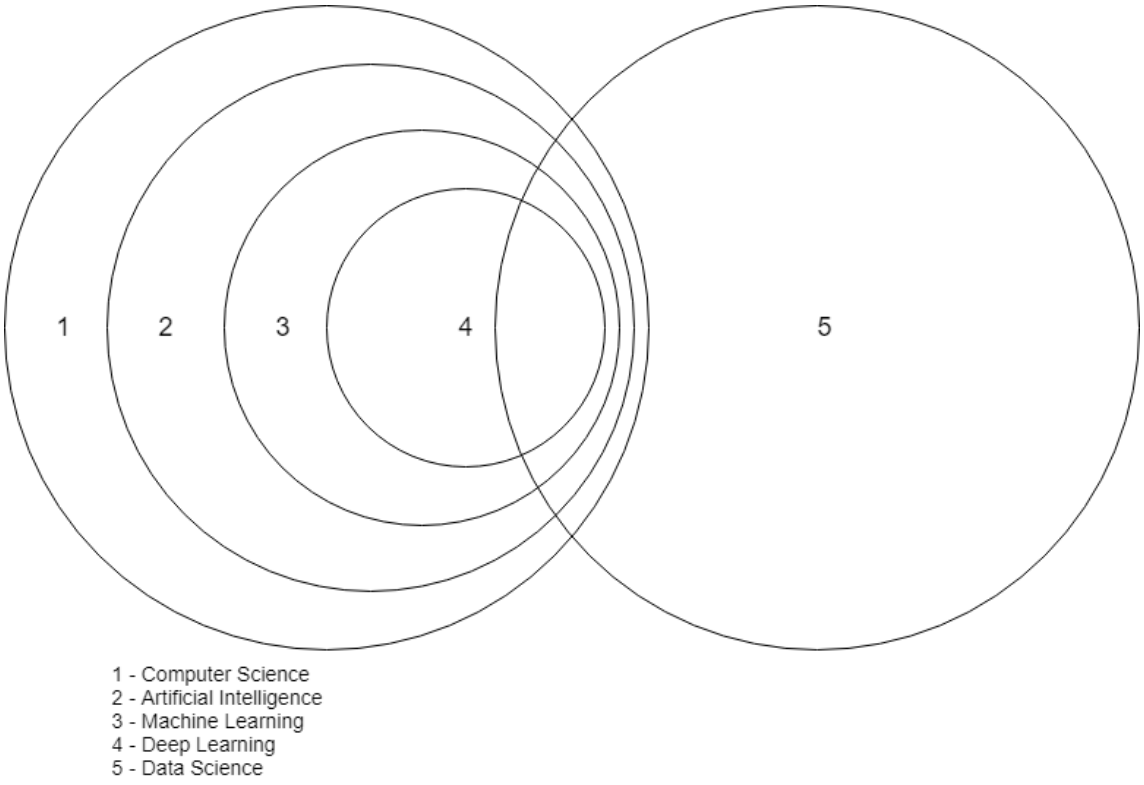

FIGURE 12. Different areas related to machine learning (University of Helsinki, 2018).

3.3.1 Mathematics

Mathematics is the science about numbers, their correlations, the laws that define them and that allow them to find their values. Mathematics studies quantity, structure, space, and change of the objects (Mura, 1993). Mathematics is implemented in almost every area in the world and it is used in a variety of related subjects: statistics, physics, chemistry, and many others. Mathematics includes many fields that each explore different subjects and aim for different achievements. For example, algebra studies functions and number sets, geometry - space and shapes, and calculus - change. Those three are the most well-known areas of mathematics, but however, there are more of them. For example, discrete mathematics, linear algebra and statistics (Devlin, 1996).

3.3.1.1 Mathematics for Machine Learning

As field machine learning lies on the intersection of statistics, computer science and algorithmic studies, therefore when thinking about the development of any machine learning system the mathematical side should be considered and calculated. Machine learning requires the mathematical toolset that includes such areas as linear algebra, multivariable calculus, statistics, and probability theory. Machine learning utilizes the complicated mathematical schemas to find the correlations inside the data and between the features or between features and result because of the algorithmic nature of the subject. Because of that requirement, the implementation process utilizes the tools provided by discrete and integral calculus, complicated areas of algebra and some geometrical tools (Parbhakar, 2018). As the dataset includes strictly numeric values, for its management the techniques from 3D-algebra are applied, for example, matrices and vectors and operations with them. All these tools are also used by related areas, such as data science, deep learning and Artificial Intelligence (Ng, 2017).

3.3.2 Statistics and probability theory

Statistics is a branch of mathematics. It explores data and how to collect, analyze, interpret and represent it. The most important part is that statistics studies the relations in the data (Romeijn, 2014). Probability is another branch of mathematics that is connected with statistics. It studies probability, which is a mathematical measurement of the likelihood of the occurrence of events on the scale from 0 to 1 (Webster's Revised Unabridged Dictionary).

3.3.2.1 Statistic and probability theory for machine learning

Machine learning heavily relies on statistics as the subject that gives it tools to manipulate data and create algorithms and correlations. Probability theory techniques and the ability to predict events based on its likelihood are another important parts of machine learning at a core (Radke, 2017).

3.3.3 Computer Science

Computer Science is a general area that studies the processes of interaction with the data in the form of algorithms and further programs. It is a very wide theoretical and practical area, which includes many disciplines, including, for example, data manipulation and computational coding theory, computer systems and applications, software engineering and hardware.

As an area computer science has been around since antiquity, even though it was represented in the form that might not be recognizable for the modern world, for example, abacus (figure 13) and basic algorithms (Ifrah, 2001).

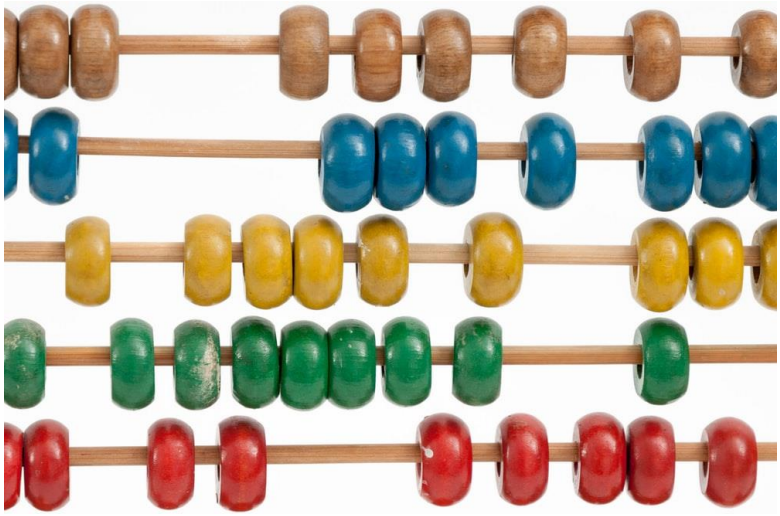

FIGURE 13. Modern abacus

Computer Science as the world knows it now was originated when the first mechanical calculator was invented in 1623 by Wilhelm Schickard (figure 14).

FIGURE 14. Mechanical calculator created by Wilhelm Schickard (Ifrah, 2001)

Artificial Intelligence is one of many computer science fields and it was originated in 1956, at a workshop at Dartmount College (McCorduck, 2004 & Crevier, 1993).

3.3.4 Artificial Intelligence

Artificial Intelligence is a Computer Science area that concentrates on researching and developing systems that can successfully perform the tasks that require human expertise. This definition may include artistic and creative functions (University of Helsinki, 2018). The replication can be done by machines mimicking actions and functions that a human can perform without consideration.

Artificial Intelligence has a large philosophical context behind it. First of all, the exact definition of it is still not properly set because many questions need to be answered. There have been attempts to define the Intelligence since antiquity by mathematicians and philosophers, and that makes it complicated for the Artificial Intelligence area to replicate in such unstable conditions. For example, it is doubtful that any system based on Artificial Intelligence will ever be able to make decisions on its own. In this case, the decision-making processed will be programmed, and if it is, can the system be considered intelligent because it is just following commands? Precisely replicating the thinking process is the main challenge of this area. There is an old joke definition that says the following: “AI is cool things that computers can’t do“ (University of Helsinki, 2018). The issue with this statement is that besides being unofficial, it limits the area in a way that if anything is implemented, it stops relating to Artificial Intelligence and it becomes impossible to make any progress. Thus, it is impossible to use this definition.

The concepts that Artificial Intelligence relates to are autonomy and adaptivity: an ability to function without the guidance and an ability to adapt to the ongoing circumstances. Those are the main goals that the systems aim to achieve.

Artificial Intelligence gave a new perspective on all the questions about intelligence even though it is agreed that the requirements for intelligence for humans and the machines would be different.

Artificial Intelligence is originally the area that Machine Learning has developed from as a tool to support it, used to recreate the ability of a human to see connections and correlations in the data, analyze and predict the outcome.

The last but not the least, Artificial Intelligence carries a lot of legacy from its interpretation, which was created by science fiction novels and movies, so it is mistakenly classified as an area that might cause danger to world peace in the future. It is important to remember that AI is a highly scientific area that can only exist under the control of the developers and researchers. It is impossible to reach full autonomy and absolute learnability. AI methods more relate to profoundly automated systems than to intelligent autonomic programs (University of Helsinki, 2018).

3.3.5 Data Science

Data Science is the informatics discipline that studies the issues related to the data analysis, formatting and representation of the data in the digital form. This field uses different scientific and algorithmic methods and processes and powerful hardware approaches to achieve the goals. Data Science can be defined as a field on the edge between statistics, data analysis, and machine learning, which uses the tools and approaches of those areas (Leek, 2013). Data Science and statistics are tied together, but even though they have a lot in common, they are very different areas at their core as well. Data Science bases on statistics, which helps to find the correlations in the data and make the data analysis more efficient. Unlike the aforementioned areas, Data Science is a new modern field which was established and defined as the world knows it within the past 30 years.

3.3.6 Deep Learning

Deep learning is a machine learning technique which enables more advanced features and algorithms that allow using visual and audial data as the dataset. Deep learning supports the manipulations, analysis, and recognition of this data. Because of that, the implementations like self-driving cars, image recognition and sound analysis applications exist. This tool brings traditional machine learning techniques to step closer to the more advanced Artificial Intelligence applications (Schmidhuber, 2015).

3.4 Dataset for machine learning

Machine learning requires the numeric data that can be analyzed by the machine in the computational format. The missing values and incorrectly formatted fields are unacceptable and need to be handled and modified. The data that is not presented in the numerical format can be categorized. For example, if the dataset includes the information about users, “1” can refer to the user if the user is female and “0” if the user is male. In some cases, the meaning and weight of the categorized field needs to be considered. The final step in the preparation of the dataset is feature scaling. The values of the features are scaled depending on the average value in the column. For example, if the average age of users in the dataset is 30, then the user whose age is 31 is marked as if their age is 1 and the user whole age is 20 is marked as if their age is -10.

After all the preparations have been completed the data needs to be divided into two major sections: a training set and a testing set in the approximate ratio of 70% for training and 30% for testing (Ng, 2017). In this way the dataset is ready to be used for the analysis and implementation of the machine learning system.

3.5 Implementations

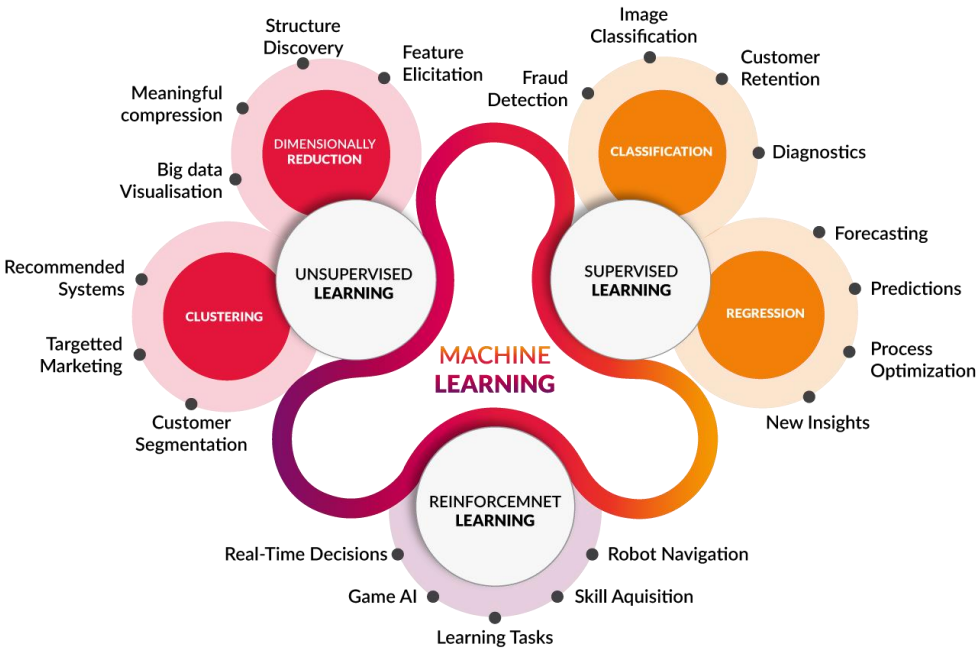

FIGURE 15. Implementation of machine learning using different machine learning techniques (Dwivedi, 2018)

Nowadays, machine learning, even though it is still a developing technology, is implemented in many areas of everyday life. Different techniques are used for these implementations (figure 15) Further there are examples of such areas.

Virtual assistants

They use machine learning algorithms on the requests of the users in order to give better answers and make better predictions. Examples: Siri, Alexa, Google Now.

Commuting

Machine learning algorithms are used for selecting the best route based on, for example, previous data about road conditions in the area. Examples: Google Maps, Yandex Maps, Yandex Transport.

Social media

Social media widely use machine learning in some features. Examples: “People you may know”, “Face recognition”.

Filtering

Filtering functionality can be implemented using a supervised learning classification algorithm. Examples: basic spam filter.

Machine learning tools are also implemented as the part of Artificial Intelligence and Deep Learning-based solutions (Daffodil Software, 2017).

3.6 Ethics

Ethical issues of machine learning are connected with the ethical issues of Artificial Intelligence (University of Helsinki, 2018). One of the most significant ethical concerns in machine learning is based on analyzing the data that was collected but is not correct from the ethical point of view, for example, it is racist, sexist or nationalist. As any machine learning algorithm relies on the data behind it, this can cause an incorrect outcome. The situation with the Amazon AI recruiting tool can be used as an example. It has been favouring male applicants over the female ones as the data, which was gathered, came from the resumes sent mostly by male applicants (Dastin, 2018).

Another ethical problem that concerns machine learning and Artificial Intelligence is that they are replacing human actions and thus, people will lose jobs. In case it goes too far, it can lead to a crisis. For example, many jobs that require an analysis of the data can be replaced by autonomous machines as technology improves.

Artificial Intelligence causes a significant amount of precaution attached to it as it is considered dangerous for humanity in case the system goes out of control (University of Helsinki, 2018).

3.7 Tools

Machine learning implementation requires many specific tools and techniques. There is a wide range of different tools that serve many purposes, for example, scientific evaluation of the data, matrix manipulation, mathematical computations and others. Further, there is a description of the most popular ones (Yegulalp, 2017).

3.7.1 Software for prototyping

FIGURE 16: Matlab logo

Matlab (figure 16) is a software environment which enables computational actions and manipulations with matrices. It runs algorithms, as well as plots graphics and creates the connection of the computational function to the user interface. This is the tool that is widely used for prototyping machine learning solutions. It utilizes the MATLAB language.

FIGURE 17: Octave logo

Octave (figure 17) is s software which has similar functionality to MATLAB. It uses the language that is compatible with the MATLAB language and bears no significant difference. Octave is used for the same purposes, and it has a benefit of being free of charge.

3.7.2 Software for implementing

3.7.2.1 Python and Python libraries

Python is a high-level, minimalistic and multitask programming language which aims for increasing productivity and readability of the code. It is widely used in the machine learning area alongside with R. Python includes many scientific, mathematical and plotting packages which are created specifically for the data science purposes.

FIGURE 18. Anaconda logo

Anaconda (figure 18) is an open source distribution of R and Python for a scientific research. It performs the actions of the packet manager for the libraries.

FIGURE 19. SciPy logo

SciPy (figure 19) is Python ecosystem that includes Python libraries that are widely used for data science: NumPy, Matplotlib, Sympy, Pandas, IPython. NumPy is used for scientific computing. Matplotlib enables 2D plotting. Sympy provides symbolic mathematics capabilities. Pandas library provides data analysis and data structure tools. IPython is an interactive computing tool.

3.7.2.2 R

FIGURE 20. R logo and command line

R (figure 20) is a programming language which enables statistical and mathematical computations. R has both a command line and a graphical interface. Anaconda, which was earlier described in detail, can be used both with R and Python. Besides that, R includes numerous packages and libraries for various mathematical and statistical goals (R FAQ, 2018).

4 THEORETICAL BASE OF THE IMPLEMENTATION

4.1 Task

During summer 2018, the company WeBuust Oy assigned the task. The assignment required to create a proof-of-concept that it is possible to use machine learning as a solution for FinTech Underwriting in terms of the services of the company, so to estimate how likely it is that the investment into a particular startup will pay off. It was decided to develop a prototype of the solution at first. If a prototype would be created successfully, it could be continued with the implementation.

4.2 Planning

The research into this topic was started from defining the main machine learning algorithms to use. It was decided to try to approach the problem from two different points of view: using a linear regression supervised learning algorithm and a K-Means clustering unsupervised learning algorithm. Further, there is an explanation of approach and the description of these algorithms.

4.2.1 Explanation of the approach

It was decided to approach the task from both supervised and unsupervised learning.

Supervised learning algorithms, which were discussed, were linear regression and classification. The main reason for that was that from the task and the data it is visible that the values in the dataset need to have a mathematical correlation for the system to able to function accurately. The weight and value of each feature needs to be defined. Later the decision was made that the algorithm used needs to be a linear regression algorithm because the classification feature was

considered unnecessary for supervised learning. It was considered more important to find the linear correlation.

The unsupervised learning approach, which was selected, was clustering to classify the data into segments. It is reasonable to do it in this way in the unsupervised learning section because of the lack of output and the fact that in this case the segmentation would be done based on the features solely, unlike in the classification, where the output would also be considered. The K-Means algorithm was used for Clustering because it is simple to implement and the proof of concept task does not require complicity.

Reinforcement learning was not considered as one of the implementation options because of the lack of relevant data.

4.2.2 Linear regression

The term “linear regression” originates from statistics. It is defined as linear modelling between the variable and the result. In machine learning, linear regression is one of the most well-known and popular supervised learning algorithms, and it plays an important role as it is relatively simple and easily understandable. To understand linear regression, it is first necessary to understand regression itself. Regression is the set of techniques applied for calculating or estimating the relations between several variables that are used in the model. This type of analysis is very popular for such tasks as prediction and forecasting, and these issues and problems are popular in the field of machine learning as well (Freedman, 2009).

4.2.2.1 Cost function with gradient descent

One of the implementations paths for linear regression is using the cost function and gradient descent to find coefficients of parameters.

In the simplified version linear regression is a type of regression technique where there is only one feature and one output. These variables create a linear relations model on the basic linear equation. It is shown in formula 1.

FORMULA 1

$h_0(x) = \theta_0 + \theta_1 x$, where

Θ_0, Θ_1 = coefficients

x = the input

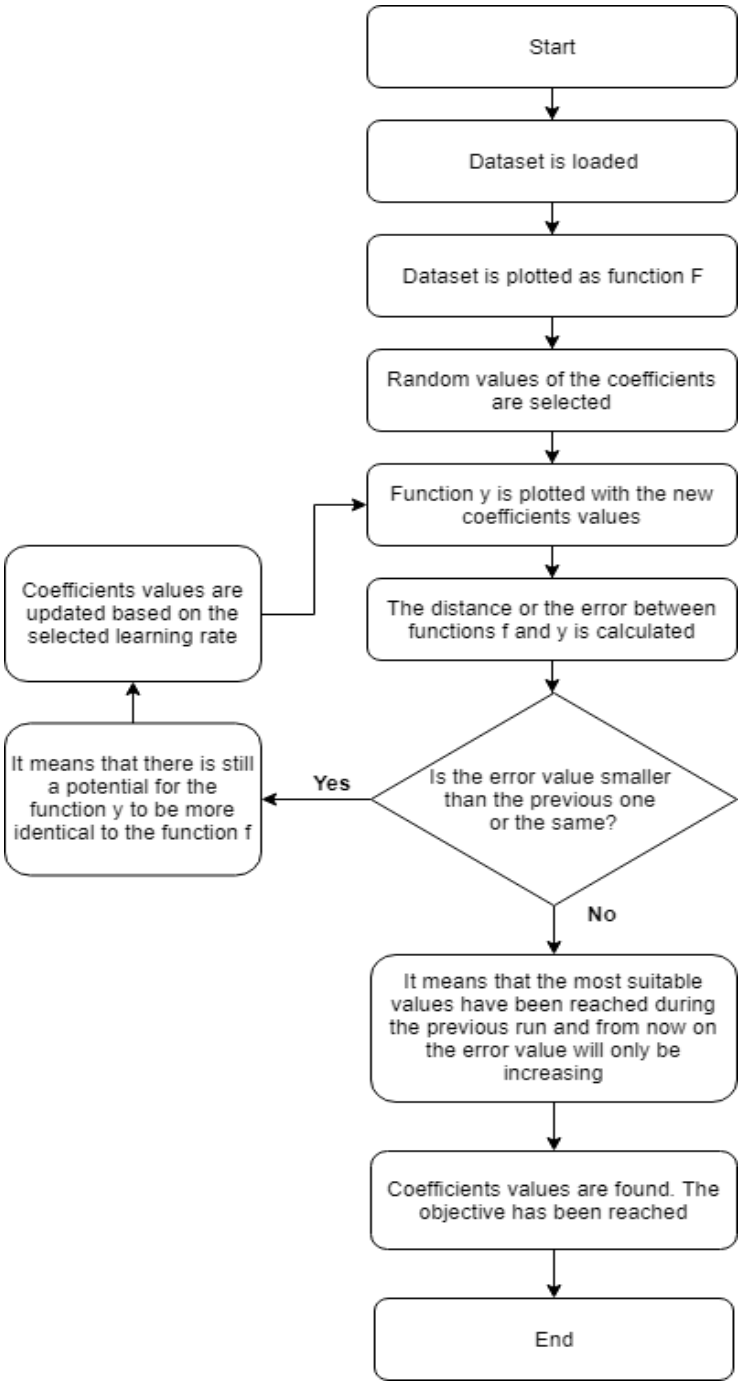

FIGURE 21. Linear regression using cost function and gradient descent flow chart

In the figure 21 the linear regression flow chart shows the main idea of how linear regression using cost function and gradient descent works. To calculate the coefficients, the cost function is applied.

Minimization and Cost function is the algorithm used for calculating the error between the function y and the points in the dataset so that the best values for the coefficients can be selected. The aim is to find the best option for the values for the coefficients that the graph of the function would fit the data points of the dataset with as small error as possible (Jung, 2018).

Minimization and Cost function calculation is shown in formula 2.

FORMULA 2

$J(\Theta_0, \Theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_0\left(x^{(i)}\right) - y^{(i)}\right)^2$, where

Θ_0, Θ_1 = coefficients

x = input

y = output

m = number of samples

$h_0(x) = \Theta_0 + \Theta_1 x$ is the hypothesis

The goal is to find the smallest value of $J(\Theta_0, \Theta_1)$ as possible, which means that the error is as low as it is possible. In order to update the values of Θ_0 and Θ_1, the gradient descent formula is used. This equation is described in the formula 3.

FORMULA 3

$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_0\left(x^{(i)}\right) - y^{(i)}\right)$, where

$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_0\left(x^{(i)}\right) - y^{(i)}\right) \cdot x^{(i)}$, where

α = learning rate

Θ_0, Θ_1 = coefficients

x = input

y = output

m = number of samples

$h_0(x) = \Theta_0 + \Theta_1 x$ is the hypothesis

The gradient descent starts the calculation with random values of θ_0 and θ_1. After that, the values update using the selected learning rate, which is the numerical value of one step of the update. The learning rate is selected based on the intuition of the developer and it is possible to change the learning rate during the implementation. At high value of the learning rate can lead to the lowest error value not being found. At low learning rate causes the training to run for a very long time. The values update for as long as it is needed to find the smallest value of the cost function. Reaching that would mean that the most suiting values of θ_0 and θ_1 are found. In more complicated linear regression algorithms, there can be many features used with one output. In this case the same procedure is performed on all the parameters, in search of coefficients (Ng, 2017).

4.2.2.2 Normal equation

The normal equation is the alternative of the cost function and gradient descent methods. It is used if the number of the parameters is not high but the dataset is rather large. With this formula it is not

needed to select the learning rate and the step, which means that this method lacks updating the values of the variables. The normal equation formula is shown in formula 4.

FORMULA 4

$\theta = (X^T X)^{-1} X^T y$, where

Θ = hypothesis parameters that define it the best

X = input

Y = output.

The main idea is that $J(\Theta_0, \Theta_1)$ is minimized by taking its derivatives with respect to $\Theta j's$ and setting them to equal zero. This is the basic algebraic way to determine optimal theta (Ng, 2017).

4.2.3 Using linear regression regarding the problem

Regarding the assignment, it was decided to use the linear regression algorithm because of the linear structure of the dataset and the mathematical relationship between the values. For example, it was considered that there would be linear relations between how much revenue the startup got in the specific amount of time. The category it is related to how large of investment is still required to be received and the rate of how successful the startup has the potential to become.

4.2.4 K-Means Clustering

The term “clustering” or “cluster analysis” refers to the set of unsupervised learning algorithms used for grouping the objects or the data points. They are classified as belonging to the specific group based on the features or the valuables given in the dataset (figure 22).

The clustering analysis was originally a term used in anthropology, originally introduced by Driver and Kroeber in 1932. K-Means Clustering was first used by James MacQueen in 1967, even though the idea belongs to the mathematician and educator Hugo Steinhaus and it goes back to 1957 (Driver & Kroeber, 1932).

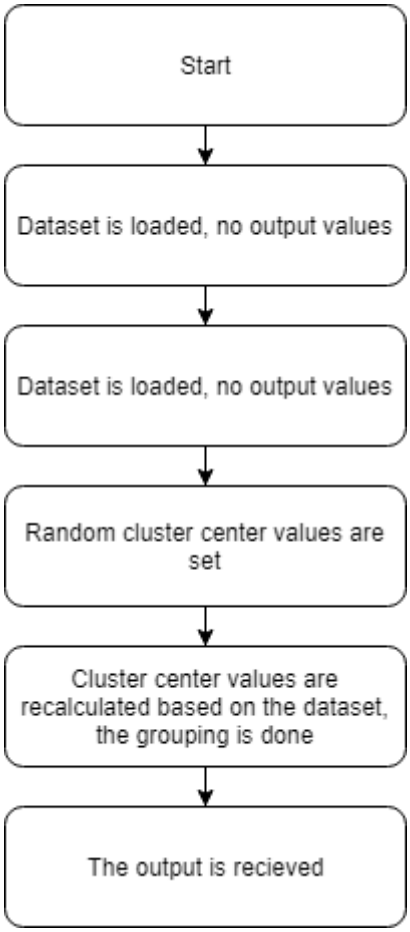

FIGURE 22. Flow chart about the nature of clustering

Because of the nature of unsupervised learning, this algorithm does not require the output value.

There are many clustering analysis algorithms. K-Means Clustering is one of the most popular and well-known approaches. The implementation of it goes the following way.

First, the required number of the clusters, which means groups or classes, is selected. The center points of the clusters are randomly assigned or selected. It means that at this stage the dataset is divided into random clusters. After that, the distance of the vector starting at each data point and going to each group center is calculated. The data point is assigned to belong to the cluster that is located closest to the data center (Ng, 2017). Based on the outcomes of the previous steps, the centers of clusters are recalculated. For that, the mean of all the vectors is computed. This method computes the optimal placement of the group centers and divides the dataset into the groups based on the values of the features (Ng, 2017).

4.2.5 Using K-Means Clustering regarding the problem

It was decided to use K-Means clustering for this problem case to test how the unsupervised learning approach would function in these circumstances and for this scenario. K-Means was selected as the clustering algorithm because of its popularity, simplicity and the good documentation and community around it.

5 IMPLEMENTATION

5.1 Dataset

It was decided to use the dataset obtained using the data.world service. This website includes various ready-made datasets used for machine learning and data science purposes. From the service, the .csv document was received with the statistics of Kickstarter platform, which is the crowdfunding platform for startups in different areas. The dataset is portrayed in table 1.

TABLE 1. Vertical representation of the original dataset

Column name	*Example of data 1*	*Example of data 2*
ID	1000002330	1000003930
Name	The Songs of Adelaide & Abullah	Greeting From Earth: ZGAC Arts Capsule For ET
Category	Poetry	Narrative Film
Main_category	Publishing	Film & Video
Currency	GBP	USD
Deadline	09/10/2015	01/11/2017
Goal	1,000.00	30,000.00
Launched	11/08/2015 12.12	02/09/2017 4.43
Pledged	0.00	2,421.00
State	failed	failed
Backers	0	15
Country	GB	US
USD pledged	0.00	100.00
USD pledged real	0.00	2,421.00
USD goal real	1,533.95	30,000.00
success_rate	0.00	1.13

The column “success_rate” includes the output of the system. In the dataset the values of the “success_rate” are varying from 1 to 100 and this way the success of the startups is defined and scaled.

The statistics of Kickstarter include the data about 378,662 projects. The features the data describes are the name and ID of the startup, its category, main category, currency and country, deadline, goal, when it was launched, how much it pledged (in the local currency and USD), state (failed, canceled, successful or live), how many backers participated and the success rate.

	A	B	C	D	E	F	G	H
1	ID	main_category_code	deadline	launched	state_category	usd_pledged_real	usd_goal_real	usd_investments_required
2	1	12	09/10/2015	11/08/2015	1	0.00	1533.95	1533.95
3	2	14	01/11/2017	02/09/2017	1	2421.00	30000.00	27579.00
4	3	14	26/02/2013	12/01/2013	1	220.00	45000.00	44780.00
5	4	15	16/04/2012	17/03/2012	1	1.00	5000.00	4999.00
6	5	14	29/08/2015	04/07/2015	0	1283.00	19500.00	18217.00
7	6	5	01/04/2016	26/02/2016	2	52375.00	50000.00	-2375.00

I	J	K	L	M	N
usd_investments_required_real	usd_over_the_plan	percentage_total_accomplished	percentage_over_the_plan	percentage_accomplished	success_rate
1533.95	0.00	0.00	0.00	0.00	0.00
27579.00	0.00	8.07	0.00	8.07	1.13
44780.00	0.00	0.49	0.00	0.49	0.07
4999.00	0.00	0.02	0.00	0.02	0.00
18217.00	0.00	6.58	0.00	6.58	0.00
0.00	2375.00	104.75	4.75	100.00	10.48

FIGURE 23. Dataset screenshot

For the project, the dataset was modified (figure 23). The fields mane, currency, how much it pledged in local currency, deadline and when it was launched were removed as unnecessary. Non-numerical features, such as category and state, were switched to numerical by assigning the number to each value, based on the following characteristics: for the category feature, success statistics of the Kickstarter categories were used, marking the least successful as 1 and the most successful as 15, the state successful and live were marked as 2 as it is the best outcome, failed was marked as 1 as it is a negative outcome, cancelled was marked as 0 as it is the lack of outcome. The new version of the dataset is portrayed in table 2.

TABLE 2. a vertical representation of the modified dataset

Column name	*Example of data 1*	*Example of data 2*
ID	1	2
Main_category_code	12	14
Deadline	09/10/2015	01/11/2017
Launched	11/08/2015	02/09/2017
State_category	1	1
USD_pledged_real	0.00	2421.00
USD_goal_real	1,533.95	30,000.00
USD_investments_required	1,533.95	27,579.00
USD_over_the_plan	0.00	0.00
Percentage_accomplished	0.00	8.07
Percentage_over_the_plan	0.00	0.00
Success_rate	0.00	1.13

Several lines were removed from the dataset due to an extreme value, which means that the statistics about one project was removed. In the end, the data about 378,500 startups was utilized. Finally, the dataset has been divided into two parts: a training dataset containing approximately 30% of the original data and a testing dataset, containing approximately 70% of the original data.

5.2 Structure of the project

The project was divided into two parts: prototyping and implementation. The prototyping stage was held first, followed up by the implementation stage.

The prototyping part consists of two directories. One is for the prototype of linear regression, the other is for the clustering prototype. Each of them include .m extension files and the .csv dataset modified specifically for this case. The plan is shown more precisely in figure 24.

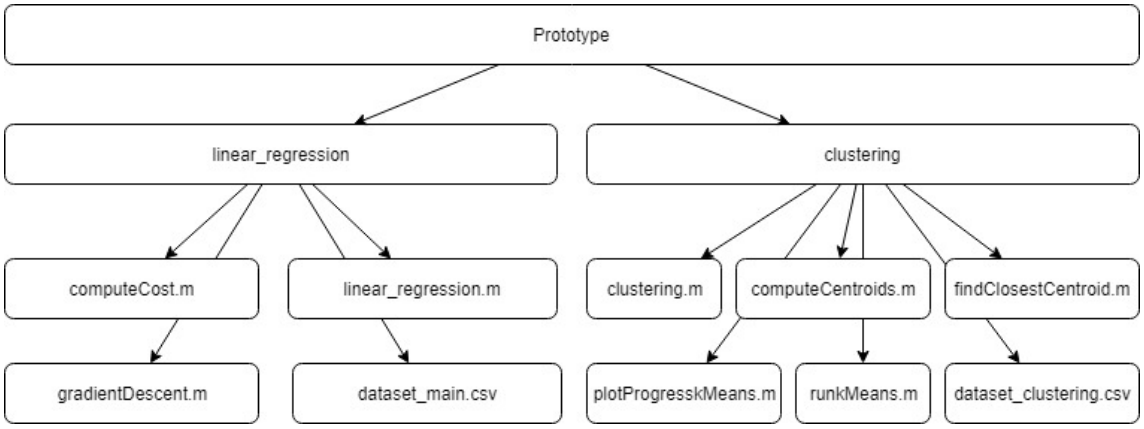

FIGURE 24. Structure of the prototype stage of the project

Similar to the prototyping part, the implementation part is divided into two sections: linear regression and clustering. Both of them include .csv dataset modified specifically for this case. Besides that, they consist of .py extension files that include the full implementation of the algorithm. The plan is shown more precisely in figure 25.

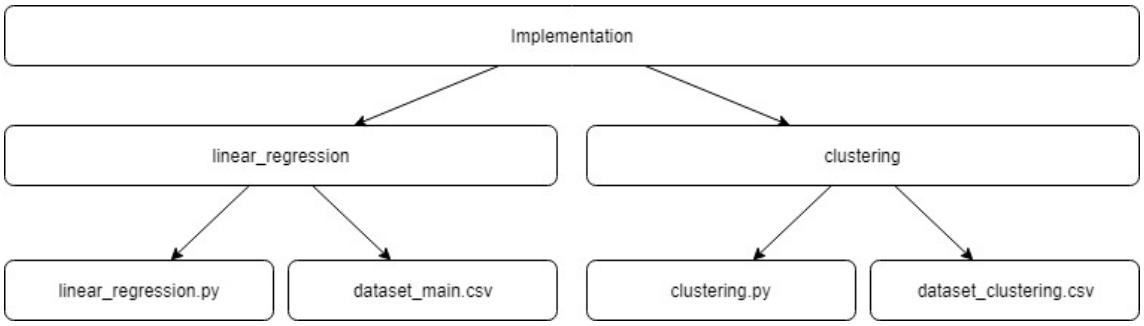

FIGURE 25. Structure of the implementation stage of the project

5.3 Prototyping

The planning phase was followed by the prototyping phase. The prototype was created using mathematical and scientific prototyping software to plan the implementation and structure of the steps needed for the successful result. The software utilized for this stage was Octave.

5.3.1 Linear regression prototype

First, the data is loaded from the dataset as it is shown in figure 26.

```
pkg load io
A = csvread('dataset_main.csv');
```

FIGURE 26. Loading data in linear_regression.m

Then the row of ones is added at the beginning of each line as it is shown in figure 27.

```
y = A(:, 12);
X = [ones(m, 1), A(:,:)];
```

FIGURE 27. Added the vector in linear_regression.m

For the prototype of the linear regression, the normal equation was used in this specific case as it is shown in figure 28.

In this specific case using a normal equation suited the dataset because of a small number of features and a large number of projects presented in the statistics material. On the other hand, in case the cost function and gradient descent were used and with such a large number of inputs, training the dataset would be too slow and possibly faulty.

```
X = [ones(m, 1), A(:,:)];
theta = pinv(X' * X) * (X' * y);
```

FIGURE 28. Normal equation in linear_regression.m

After that, the result is plotted as the linear regression graph that is shown in figure 29.

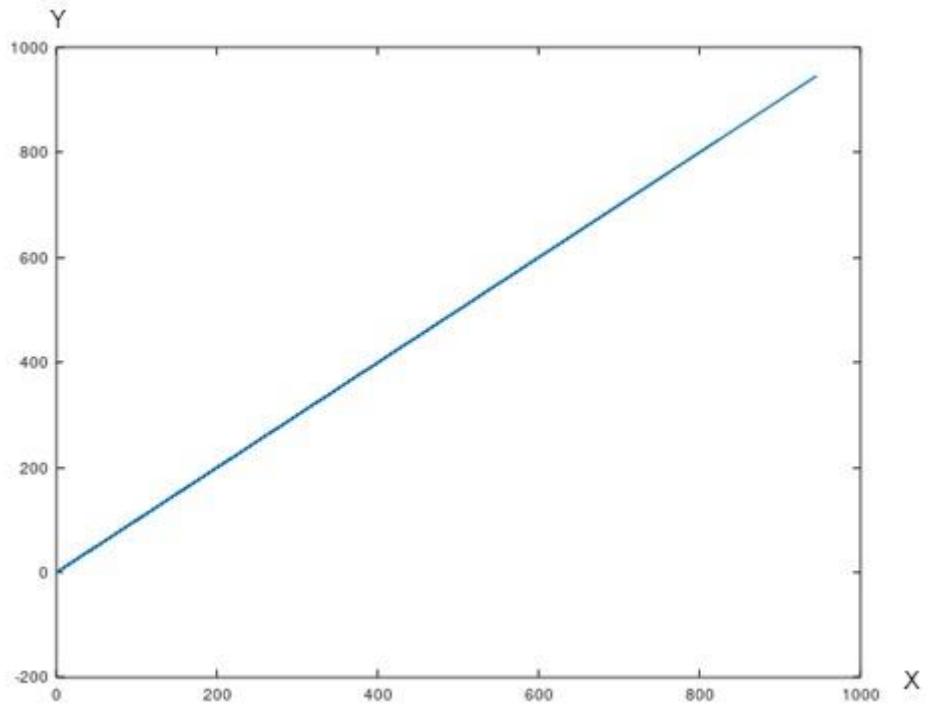

FIGURE 29. The result of linear regression prototype, where X is a collectible variable to input (features) and Y is output (success rate).

5.3.2 K-Means Clustering prototype

First, the data is loaded from the dataset as it is shown in figure 30.

```
pkg load io;
A = csvread('dataset_clustering.csv');
```

FIGURE 30. Loading data in clustering.m

Then the amount of the centroids is selected and the random lines of the dataset are set as initial centroids. It is shown in figure 31.

```
K = 10; % 10 Centroids
initial_centroids = [1 12 1 0 1533.95 1533.95 1533.95 0 0 0 0;
     30000 11 2 1825 1000 -825 0 825 182.5 82.5 100;
     60000 6 1 10 1700 1690 1690 0 0.59 0 0.59;
     90000  5   2   20427   7500    -
12927  0   12927   272.36  172.36  100;
     120000 2   1   85.67   14277.76    14192.09    14192.09    0   0.6 0
   0.6;
     150000 9   1   1   350000  349999  349999  0   0   0   0;
     180000 9   0   2433.33 159401.63   156968.3    156968.3    0   1.53
   0   1.53;
     210000 11  1   210 1000    790 790 0   21  0   21;
     240000 5   1   650 4500    3850    3850    0   14.44   0   14.44;
     280000 6   2   9560.5  9000    -
560.5  0   560.5   106.23  6.23    100];
```

FIGURE 31. Centroids are selected in clustering.m

The K-Means Clustering algorithm is used for re-calculating the centroid centers and defining the groups of data points that are closest to it as it is shown in images 32, 33 and 34.

```
idx = findClosestCentroids(A, initial_centroids);
centroids = computeCentroids(A, idx, K);
```

FIGURE 32. Centroids are re-calculated in clustering.m

```
function idx = findClosestCentroids(A, initial_centroids)
K = size(initial_centroids, 1);
idx = zeros(size(A,1), 1);
m = size(A,1);
for i = 1:m
    distance_array = zeros(1,K);
    for j = 1:K
        distance_array(1,j) = sqrt(sum(power((A(i,:)-
initial_centroids(j,:)),2)));
    end
    [~, d_idx] = min(distance_array);
    idx(i,1) = d_idx;
end
end
```

FIGURE 33. Re-calculation function in findClosestCentroids.m

```
function centroids = computeCentroids(A, idx, K)
[m n] = size(A);
centroids = zeros(K, n);
for k=1:K
   centroids(k, :) = mean(A(idx==k, :));
end
end
```

FIGURE 34. Centroids computation function in computeCentroids.m

In the end, the data is plotted as the scattered graph as it is shown in the figure 35. Unfortunately, Octave lacks 3D plotting tools, which makes results non-visual (figure 36).

```
[centroids, idx] = runkMeans(A, initial_centroids, max_iters, true);

plot(centroids(:,1), centroids(:,2), centroids(:,3), centroids(:,4),
centroids(:,5), centroids(:,6), centroids(:,7), centroids(:,8),
centroids(:,9), centroids(:,10));
```

FIGURE 35. Data is plotted in clustering.m

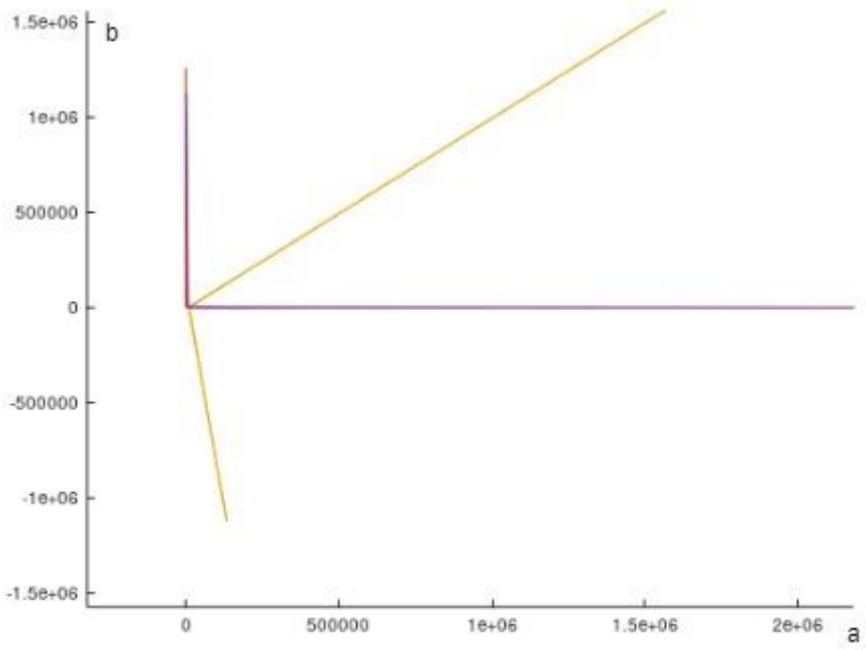

FIGURE 36. The clustering result plot, where a and b are parameters

5.4 Prototype to implementation

As it was possible to identify the mathematical concept and receive coherent results, the prototype was developed successfully. As the prototype was developed successfully and it can be used as a solid base for the further stages, the project was continued with the implementation of the system. It can be achieved by following the same development steps as in the prototype but using relevant tools and technologies.

5.5 Implementation

The project was implemented using the Python programming language and its libraries and Anaconda as a package and library manager. The packages Numpy, Pandas, Matplotib were utilized. It followed the steps in the prototype accurately. The only major difference was adding the rating feature.

5.5.1 Linear regression implementation

First, the imports are included and the data is loaded from the dataset as it is shown in figure 37.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from numpy import genfromtxt

A = genfromtxt('dataset_main.csv', delimiter=',')

y = A[:, 11] #output row
```

FIGURE 37. Packages are imported and the data is loaded in linear_regression.py

As was mentioned beforehand, during the prototyping stage, it was decided to use a normal equation as an equivalent to the cost function and gradient descent because of the various advantages in this scenario and with the current dataset.

```
def normalEquation(A, y):
    m = int(np.size(A[:, 1]))
    theta = []

    bias_vector = np.ones((m, 1))

    X = np.append(bias_vector, A, axis=1)

    X_transpose = np.transpose(X)

    theta = np.linalg.inv(X_transpose.dot(X))
    theta = theta.dot(X_transpose)
    theta = theta.dot(y)

    plt.plot(y, X*theta)
    plt.ylabel('success_rate')
    plt.show()

    return theta, X

p = normalEquation(A, y)

print(p)
```

FIGURE 38. The normal equation function in linear_regression.py

After that, the result was plotted as the linear regression graph (figure 38). It is visually hard to understand because it is a 12-dimensional space and the plotting library supports only 2D (figure 39).

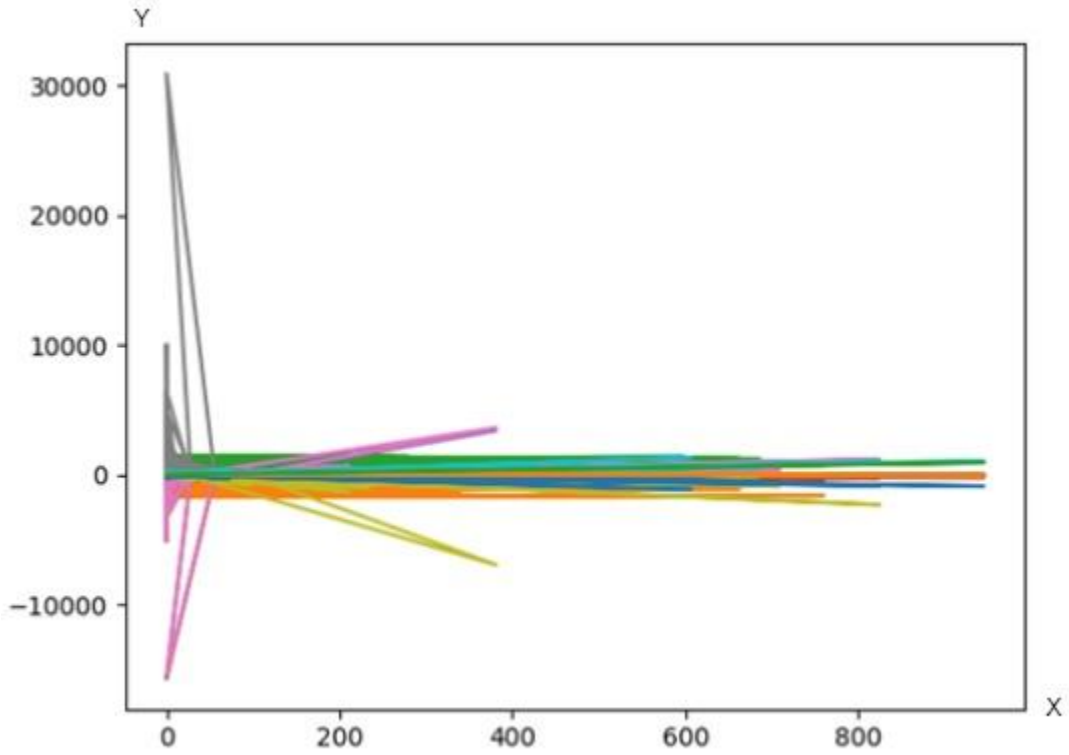

FIGURE 39. Linear regression, result. X is the collectible variable for input (features) and Y is output (success rate).

After the coefficients of the features are found and the model is trained, the functionality of rating potential success of a startup is added as it is shown in figure 40.

```
print("Enter the line number:")
selected_line = int(input())

selected_value = A[selected_line-1, :]
selected_value = np.reshape(selected_value, (1, 12))

selected_value = selected_value[:, 11]

max_value = max(y)
min_value = min(y)
section = (max_value - min_value)/100

i = 1
selected_section = 0

while selected_section == 0:
    value = min_value + section * i

    if value >= selected_value[0]:
        selected_section = i
    else:
        i += 1

print("Line number ", selected_line, " belongs to the section number ",
selected_section)
```

FIGURE 40. Rating functionality in linear_regression.py

The number of the line can be selected by the user, which means that some particular startup is selected, and the program rates this startup on the scale from 1 to 10 based on the data features of this line and where the data point is placed on the plane.

```
Enter the line number:
2000
Line number  2000  belongs to the section number  3
```

FIGURE 41. The result of the rating feature

After the coefficients of the features are found and the model is trained, the functionality of rating potential success of a startup is added as it is shown in figure 41.

5.5.2 K-Means Clustering implementation

First, the imports are included and data is loaded from the dataset as it is shown in figure 42.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from numpy import genfromtxt
from sklearn.cluster import KMeans # import KMeans
from scipy.spatial import distance_matrix

A = genfromtxt('dataset_clustering.csv', delimiter=',')
```

FIGURE 42. The packages are imported and the dataset is loaded in clustering.py

Then the amount of the centroids is selected and the random lines of the dataset are set as initial centroids as it is shown in figure 43.

```
K = 10 # 10 Centroids

initial_centroids = [[1, 12, 1, 0, 1533.95, 1533.95, 1533.95, 0, 0, 0,
0],
[2, 14, 1, 2421, 30000, 27579, 27579, 0, 8.07, 0, 8.07],
[3, 14, 1, 220, 45000, 44780, 44780, 0, 0.49, 0, 0.49],
[4, 15, 1, 1, 5000, 4999, 4999, 0, 0.02, 0, 0.02],
[5, 14, 0, 1283, 19500, 18217, 18217, 0, 6.58, 0, 6.58],
[6, 5, 2, 52375, 50000, -2375, 0, 2375, 104.75, 4.75, 100],
[7, 5, 2, 1205, 1000, -205, 0, 205, 120.5, 20.5, 100],
[8, 5, 1, 453, 25000, 24547, 24547, 0, 1.81, 0, 1.81],
[9, 10, 0, 8233, 125000, 116767, 116767, 0, 6.59, 0, 6.59],
[10, 14, 0, 6240.57, 65000, 58759.43, 58759.43, 0, 9.6, 0, 9.6]]
```

FIGURE 43. Centroids are selected in clustering.py

The K-Means Clustering algorithm is used for re-calculating the centroid centers and defining the groups of data points that are closest to it as it is shown in figure 44.

```
kmeans = KMeans(K) # create kmeans object
kmeans.fit(A) # fit kmeans object to data
print(kmeans.cluster_centers_) # print location of clusters learned by
kmeans object
y_km = kmeans.fit_predict(A) # save new clusters for chart
```

FIGURE 44. The K-Means Clustering algorithm is implemented n clustering.py

In the end, the data is plotted as the scattered graph as it is shown in figure 45.

```
plt.scatter(A[y_km ==0,0], A[y_km == 0,1], s=100, c='red')
plt.scatter(A[y_km ==1,0], A[y_km == 1,1], s=100, c='black')
plt.scatter(A[y_km ==2,0], A[y_km == 2,1], s=100, c='blue')
plt.scatter(A[y_km ==3,0], A[y_km == 3,1], s=100, c='cyan')
plt.scatter(A[y_km ==4,0], A[y_km == 4,1], s=100, c='grey')
plt.scatter(A[y_km ==5,0], A[y_km == 5,1], s=100, c='gold')
plt.scatter(A[y_km ==6,0], A[y_km == 6,1], s=100, c='lavender')
plt.scatter(A[y_km ==7,0], A[y_km == 7,1], s=100, c='navy')
plt.scatter(A[y_km ==8,0], A[y_km == 8,1], s=100, c='salmon')
plt.scatter(A[y_km ==9,0], A[y_km == 9,1], s=100, c='tomato')

plt.show()
```

FIGURE 45. The graph is plotted in clustering.py

It is visually hard to understand because it is a 12-dimensional space and the plotting library supports only 2D (figure 46).

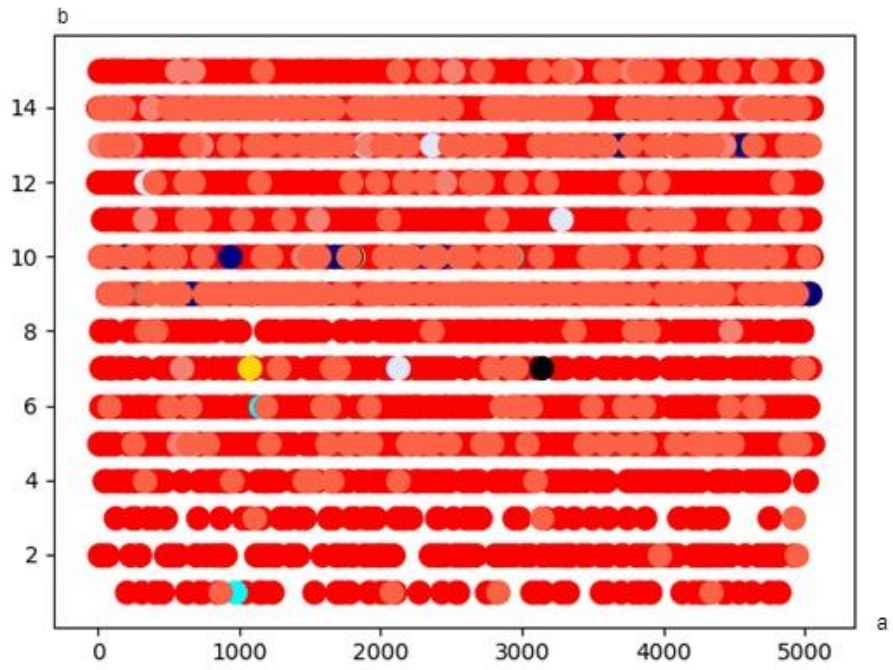

FIGURE 46. The clustering result, where a and b are parameters

After the coefficients of the features are found and the model is trained, the functionality of rating the potential success of the project is added.

```
print("Enter the line number:")
selected_line = int(input())

selected_value = A[selected_line-1, :]
selected_value = np.reshape(selected_value, (-1, 1))
selected_value = np.reshape(selected_value, (1, 11))

print(selected_value)

i = 0
small_value = 100000

for i in range (0, K):
    target_cluster = np.reshape(kmeans.cluster_centers_[i], (-1, 1))
    target_cluster = np.reshape(target_cluster, (1, 11))
    distance = distance_matrix(target_cluster, selected_value)
    if distance < small_value:
        small_value = distance
        cluster = i

print("Line number ", selected_line, " belongs to the cluster number ",
cluster)
```

FIGURE 47. The rating functionality in clustering.py

The number of the line can be selected by the user, which means that some particular startup is selected, and the program rates this startup on the scale from 1 to 10 based on the data features of this line and which cluster center the data point is located closest to (figures 47 and 48).

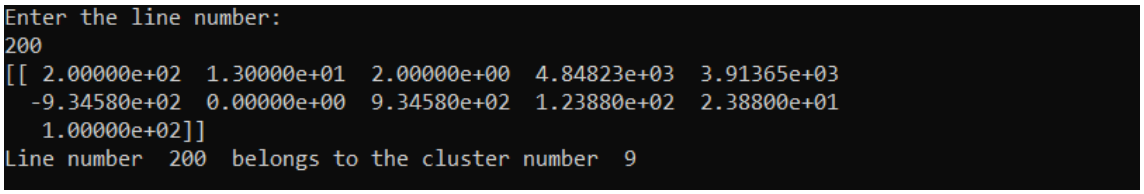

FIGURE 48. The rating feature screenshot

This way the success of the startup can be computed and rated, to define if the investment into this project will be refunded.

6 RESULTS

6.1 Presentation of results

The results have been achieved and it is possible to rate startups based on the collected data about it, thus the background check feature has been implemented.

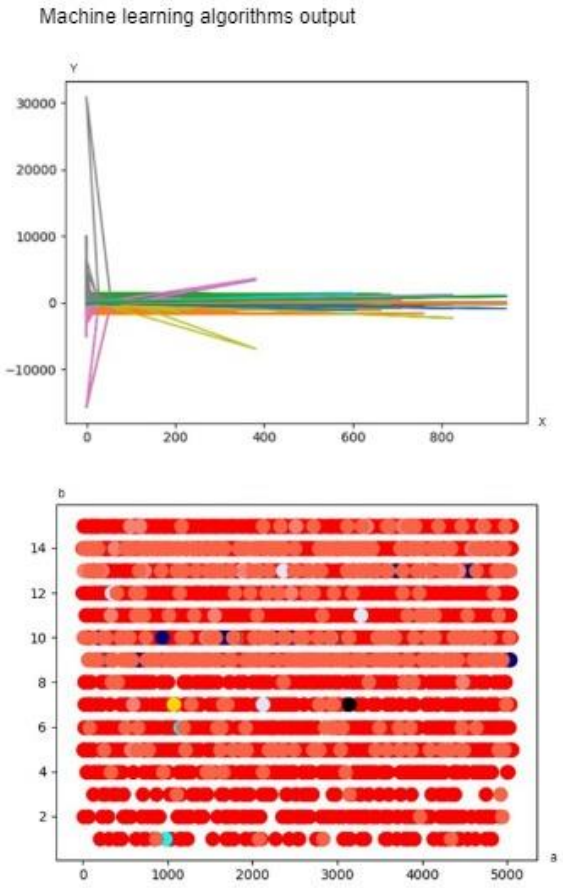

FIGURE 49. The output of machine learning algorithms. 1) Linear regression. 2) Clustering. X, a, b are features, Y is an output

As it is shown in figure 49, the first result of the system is the output of the machine learning algorithm. It is based on the mathematical correlations inside the data and it can be represented in a visual format. The linear regression system results in the weights or coefficients of each feature. From these values, the necessary computation for receiving the desired output can be obtained. This way it can be considered that the training of such a system resulted in an mathematical action, which can be performed on the features to receive the output that was given in the dataset originally.

The calculations of the relations inside the data are not performing any action. It means that originally they do not enable the rating feature. This feature is based on the results and output of the machine learning algorithm and it uses the calculated correlations to achieve the objectives.

To make the representation of the results more explicit, it can be specified that in this particular case the machine learning algorithm can be used and referred to the background check feature. It checks the data and finds the correlations. The rating is done further with the rating feature.

In figure 50 it is demonstrated that the rating feature can be enabled by the selection of the number of the line, which relates to specific startup.

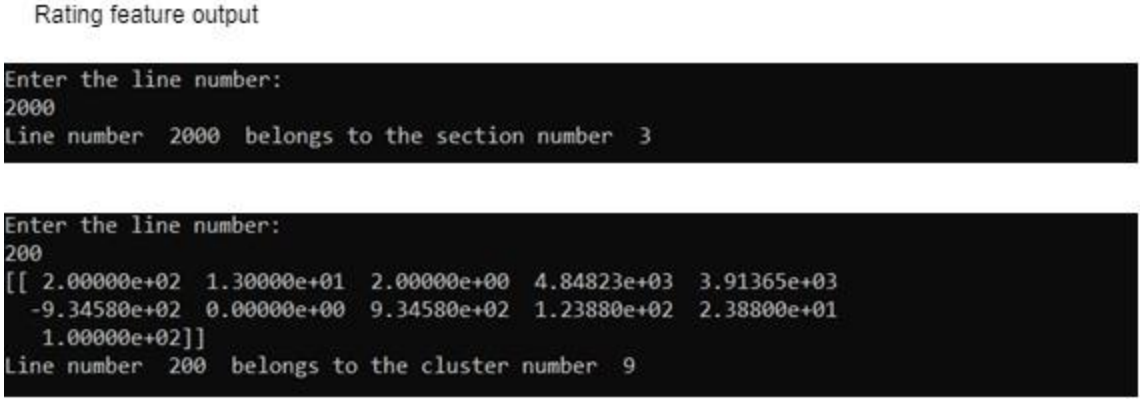

FIGURE 50. The output of the rating feature. 1) Linear regression. 2) Clustering

In the case with linear regression after the correlation between features and output is found, the system produces the numeric value of the section that the specific startup belongs to. In this

case, the rating result is based on the correlation between the features. That correlation is divided into 10 sections, each of them represents a specific rating value. The visual representation of correlation can be found in figures 35, 45 and 55 (figure 35 gives a more visually understandable idea of the linear relations between features).

In the case with the clustering algorithm after the division to the clusters has been completed, the systems output a numeric value of the cluster that the specific startup belongs to. 10 is set as the number of clusters. The visual representation of correlation can be found in figures 42, 52 and 55. Figure 42 demonstrates the division between clusters, while figure 52 describes the clusters themselves.

For both cases, the representation differences are caused by the alterations between plotting libraries.

Currently, only the backend of the rating feature is available. The client user interface is not developed yet. However, it can be implemented as part of the existing services of the company or the separate service.

As a result, the company received the further deliverables of the project: the source code, the demonstration, and the presentation.

6.2 Testing

After the implementation was completed, the testing has been performed on the system. It was performed in order to determine if the result of the system is correct and to confirm if the system functions properly. For the testing purposes, the testing dataset was utilized. It contains 70% of the original dataset.

6.2.1 Linear regression algorithm testing

During the testing of linear regression algorithm, the testing result is based on the percentage of the output values that fit into the mathematical correlation that has been found during the training process. It is done via a basic mathematical computation because as the result of the training process, the mathematical correlation has been received. After the computation, the result that the system produces is compared with the one originally given as an output in the testing dataset.

Figure 51 demonstrates that the testing dataset is divided into two copies. The first original copy includes both input and output data from the dataset (as the linear regression dataset contains both input and output). The second copy includes the same input data, but the output data is obtained using a mathematical calculation, which was determined from the training linear regression system. After that, the obtained result is compared with its counterpart from the original dataset.

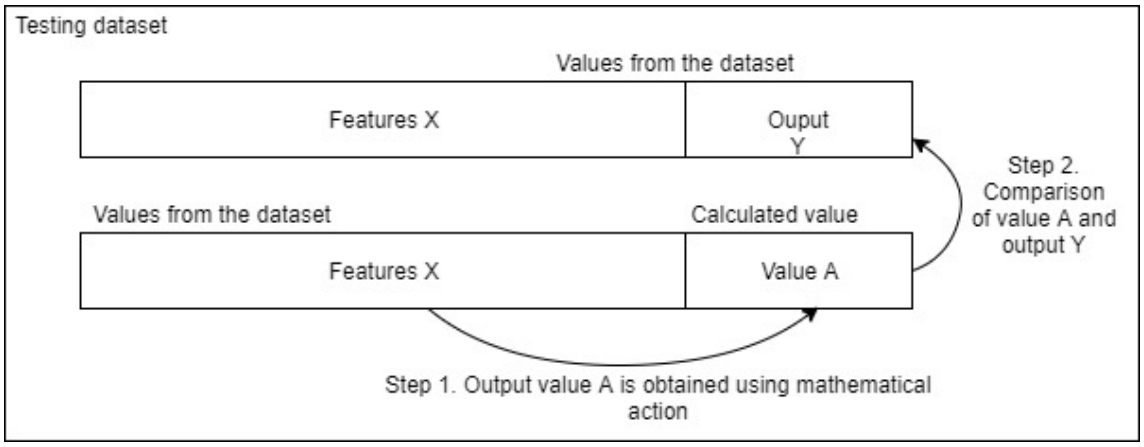

FIGURE 51. The linear regression testing flow, where features X are the input values for the system, output Y is an output value from the dataset, value A is an output value produced by the system

TABLE 3. The vertical representation of example of testing data

Feature name	*Example data 1*	*Example data 2*
Name	Road to the Shire	Spiral Electric Skylab Recording
ID	83	84
main_category_code	14	15
deadline	14.2.2012	26.2.2015
launched	11.1.2012	4.2.2015
state_category	2	2
usd_pledged_real	4,045.00	1540.00
usd_goal_real	4,000.00	500.00
usd_investments_required	-45,00	-1,040.00
usd_investments_required_real	0,00	0.00
usd_over_the_plan	45,00	1040.00
percentage_total_accomplished	101.13	308.00
percentage_over_the_plan	1,13	208.00
percentage_accomplished	100.00	100.00
success_rate	28.32	92.40

This testing example uses the data from the testing dataset that is shown in table 3. The column "success_rate" includes the output values of the system, varying from 1 to 100.

During the testing, both successful and unsuccessful results were received. An example of the testing success with the data from the example data 2 from table 3 is demonstrated in the figure 52. There it is visible that the result of a mathematical action to the data equals the output from the first copy of the dataset. This means that the system has produced a correct value.

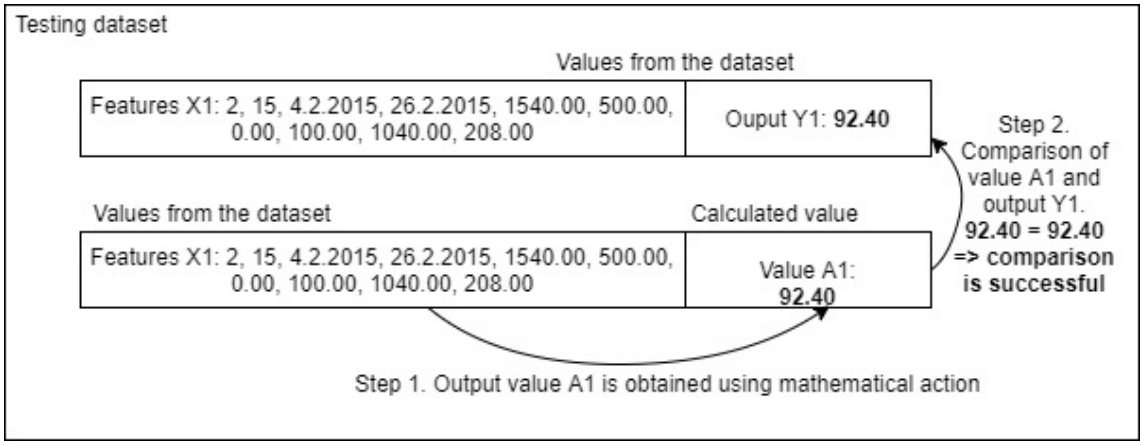

FIGURE 52. The linear regression testing success example using example data 2 from table 3. Features X1 are the input values for the system, output Y1 is an output value from the dataset, value A1 is an output value produced by the system

An example of the testing failure with the data from the example data 2 from table 3 is demonstrated in figure 53. There it is visible that the result of a mathematical action to the data does not equal the output from the first copy of the dataset. This means that the system has produced an incorrect value.

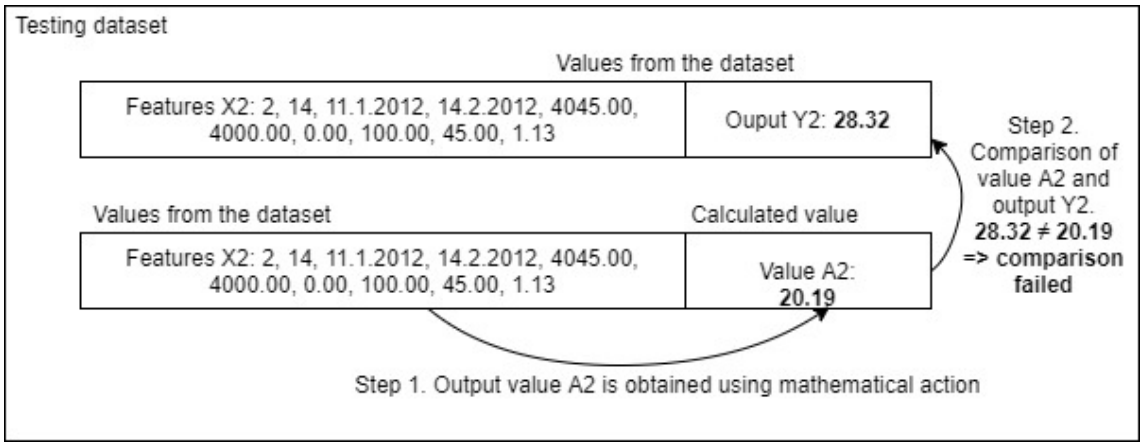

FIGURE 53. The linear regression testing failure example using example data 1 from table 3. Features X2 are the input values for the system, output Y2 is an output value from the dataset, value A2 is an output value produced by the system

For linear regression, during different re-computation on the full testing dataset, the testing result varied approximately from 60% to 85%. This means that is was from 60% to 85% of the successful comparisons.

6.2.2 Clustering algorithm testing

For the clustering algorithm testing the remaining data of the testing, the dataset is proceeded with the same computation as the training data. Consequentially, the testing data either fits into the already existing cluster data (figure 54) or it does not fit into the cluster data (figure 55). In the K-Means clustering algorithm it is determined by comparison of distances to the selected cluster center and centers of other clusters.

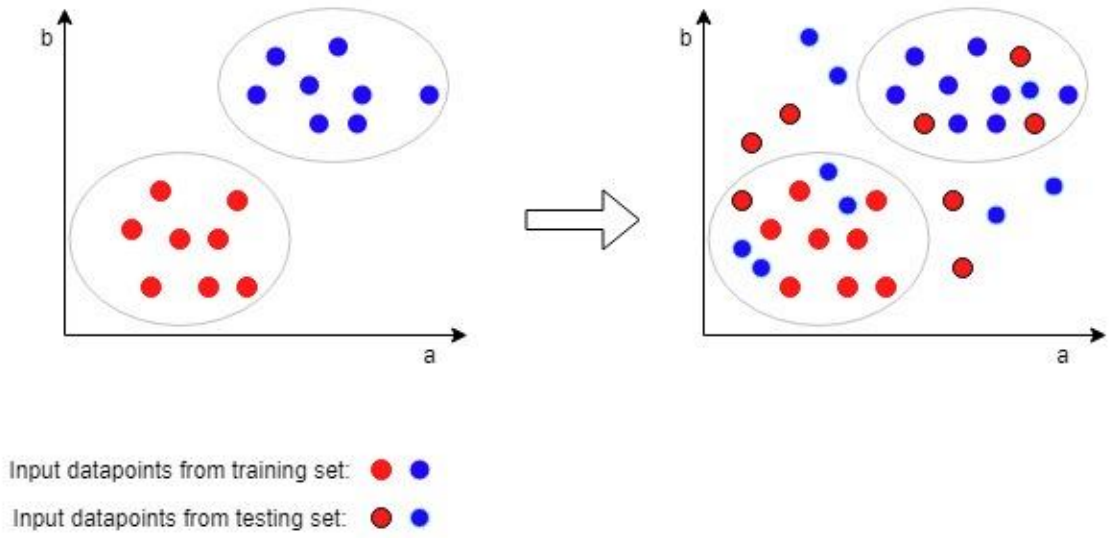

FIGURE 54. The clustering algorithm failure example, where a and b are features

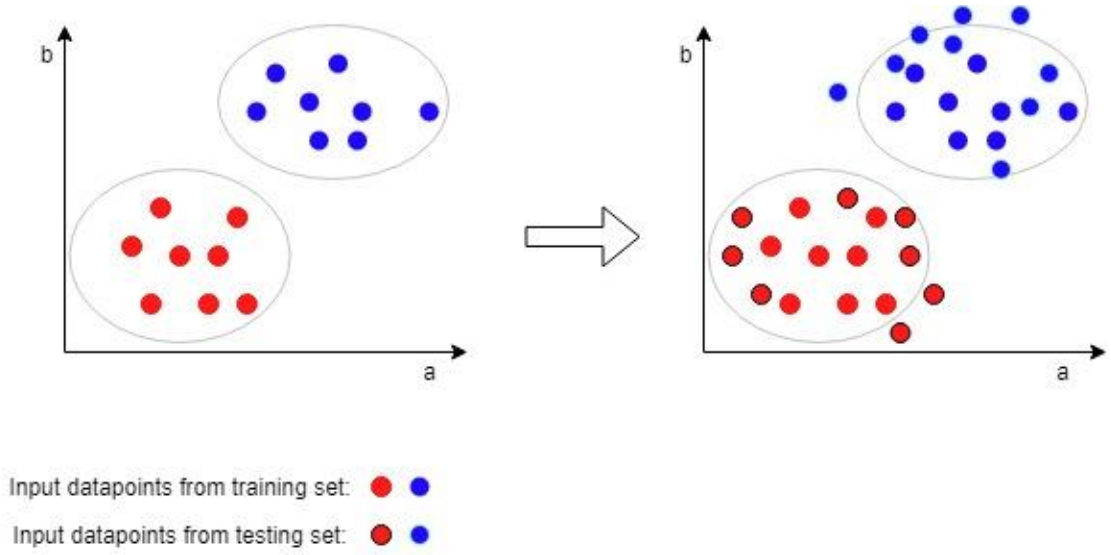

FIGURE 55. The clustering algorithm success example, where a and b are features

For the clustering algorithm during a different re-computation, the testing result varied in the larger range, from 40% to approximately 85%. This is acceptable but more concerning. This means that it was from 40% to 85% of the successful comparisons.

6.2.3 Testing result

The models are considered to fit into the dataset and the success rating feature gives correct and coherent results in most of the cases. 60% to 85% and 40% to 85% is considered a good result as most of the data points fit. A higher fitting success is not suggested as it can cause confusing and wrong results from the mathematical model overfitting, which is too close modeling of the data (figure 56).

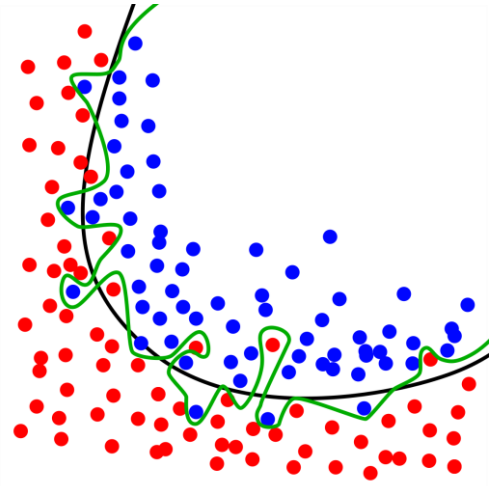

FIGURE 56. An overfitting example. A green line corresponds to overfitted model, a black line corresponds to the optimal model

Based on that, the implementation can be considered successful. This way the potential success of the startup can be computed and rated to define if the investment into this project will be refunded.

6.3 Analysis

6.3.1 Analysis of linear regression implementation results

During the implementation, it was decided that the linear regression algorithm is a better choice for the current solution. From the implementation results, it can be seen that the mathematical correlation between the values was found successfully. However, the biggest issue with this approach is that supervised learning requires the dataset that includes the result or the outcome for the model training. In case if this dataset is obtained and correctly used, this approach can be considered working well, but in the current case the example dataset used was inaccurate.

Nevertheless, linear regression has shown itself as a more complicated system to implement but a simpler one to understand and it was decided that it is more reliable. It makes basic linear computations that create a solid system and basis. In the future, if more work is put into with the system, it can be transferred to the actual dataset of the company.

6.3.2 Analysis of clustering implementation results

K-Means Clustering implementation was considered successful, but it has some disadvantages.

The main reason is that the K-Means Clustering algorithm has not proved itself as the system that is reliable enough to make computations on the dataset because it seems to be making doubtful decisions related to the rating feature. For example, the size of clusters would change significantly with each re-computation and the clusters that each dataset point belongs to did not have any relation to each other, they were randomly ordered.

However, the great advantage of this approach is that with unsupervised learning algorithms it is not required to have an outcome of the model as the system finds correlations inside of the data by itself. In this specific case that might be a good approach as the required dataset is lacking.

6.3.3 Analysis result

From the analysis it can be determined that the best approach for the implementation is the linear regression algorithm as it is more applicable. It also fits this specific case better. The same results can also be concluded from the testing. Using linear regression for this case provides more accurate results.

6.4 Applicability

6.4.1 Current case

The developed system applies to the current case with the available dataset unconditionally. This means that with the dataset that was used it is possible to make the background check of the startups and rate them according to the results of the background check. This system can be implemented as a feature in the services of the company.

However, in case if in the further situation a different dataset is used the system requires transformation that is mostly related to the changing number of features, different output columns or other data modifications. There can be other changes related to the changes in the idea and the objectives, but in general, if the use case does not change very much, not many transformations will be required on this side.

6.4.2 General usage

The general usage of this system is a more interesting case. To begin with, as was mentioned earlier any change in the dataset would require a change in the system.

The specific system that was developed during this research can be utilized not only for performing the background check and rating of the startups, but also for people, larger companies and other cases that require these functions. The key features are evaluated depending on the content of the dataset. Different possible use cases include background check before offering a position or a study place, before giving a loan or before performing other actions that require underwriting. Further, this system can be applied using different feature data and more background information.

Another important thing is that if the quality dataset is provided, machine learning can be used as a very powerful tool as it is mainly directed by the statistical data. Thus, the system in its raw format can be implemented for any feature based on the respectful machine learning algorithms.

6.4.3 Advantages of FinTech underwriting using machine learning over traditional underwriting

As it was previously mentioned in 2.4 FinTech Underwriting using Machine Learning, using FinTech technologies for underwriting comes with many advantages.

In general, the underwriting is an expensive procedure. This is caused by the existing methodology in this area, such as human work and banking transactions and confirmations. Because of that in many cases for a small to medium-sized business, it is unlikely to receive a smaller loan. $100,000 loans can cost to the loan-giving authority up to $1,000,000, but the profit received from this transaction is many times less. Therefore, small to medium-sized businesses might struggle from this issue and not be able to receive the financial support that they require to grow and develop.

The traditional underwriting heavily relies on the following areas:

- 35% Payment History
- 30% Revolving Utilization
- 15% Credit History Length
- 10% Types of Credit Used
- 10% Inquiry Count (Troy Do, 2017)

Different systems may use other classifications, such as for example, the FICO score.

Machine learning enables the use of more features for the underwriting, which are traditionally not utilized for the background check. Such features include social network data, geolocation, articles and videos, bill payments and others. The automated underwriting system also allows to bring the default rate down. Thus, the FinTech approaches in the underwriting enable such possibility as loan for small to medium-sized businesses and startups with a lower default rate, lower price of the transaction and larger target group (Do, 2017).

In the particular case of WeBuust Oy services, the traditional underwriting would create unnecessary spending both from the side of the company and clients because the clients are mostly startups and smaller businesses. From this point of view, the implementation of such system as part of the services of the company is essential.

6.5 Issues

6.5.1 Dataset

The current dataset does not relate well to the specific task case. However, it produces coherent rating results and the system finds the correlation from it correctly. Further, a different dataset needs to be used, which means that the systems requires minor changes before that.

6.6 Further development

6.6.1 Different approaches

Even though the implementation and the research have been performed successfully, there is still space for future research, using different approaches and going more in-depth with more advanced technologies. There are many machine learning algorithms and many of them would be suitable for this specific case. Linear classification is one of the algorithms that would combine a solid supervised learning approach and the need for the task to classify the startups. It would allow dividing the data into groups. This approach can be tested and it can be determined if it is more accurate and stable.

6.6.2 More advanced technologies

There are several more advanced technologies that can make this system better and more accurate. They can be implemented in case the system develops further and grows in size and features. This way that would be a suitable support for the services of the company.

The first idea would be to implement the system using deep learning functionality. It would allow adding such features for the background check as video, voice and image processing. This would enable checking the information more in-depth and also reviewing the logo, news articles, and videos, and background information about specific people in the startup.

The other approach would be to use more complicated and advanced tools of neural networks and Artificial Intelligence. In this case, it would enable even a deeper analysis which can use web scraping to check the contents of articles, analyze competitors and more advanced features.

7 CONCLUSION

By researching, developing the prototype and implementing the system, it has been proven that it is possible to use machine learning for FinTech Underwriting in terms of services of WeBuust Oy. The rating system was researched as well and implemented successfully. The best development approach was selected and other potential options were defined. The system is ready to be implemented as a part of the services of WeBuust Oy.

The results were reported to the company and the project was considered as successful.

REFERENCES

Association for Computational Learning. Date of retrieval: 30.04.2019. http://www.learningtheory.org/

Baumann, Neal. 2014. The fintech evolution in insurance. Date of retrieval 23.02.2019. https://www2.deloitte.com/global/en/pages/financial-services/articles/fintech-revolution-in-insurance.html

Clayton, Timothy. 2019. AI and Machine Learning in Fintech. Five Areas Which Artificial Intelligence Will Change For Good. Date of retrieval 30.03.2019. https://www.netguru.com/blog/ai-and-machine-learning-in-fintech.-five-areas-which-artificial-intelligence-will-change-for-good

D., Anastasia. 2018. 7 Exciting Uses of Machine Learning in FinTech. Date of retrieval 19.02.2019. https://rubygarage.org/blog/machine-learning-in-fintech

Daffodil Software. 2017. 9 Applications of Machine Learning from Day-to-Day Life. Date of retrieval 21.02.2019. https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Dastin, Jeffrey. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Date of retrieval: 07.09.2019. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

DeMuro, Jonas. 2018. What is a neural network? Cognitive science applied to computer learning theory. Date of retrieval 01.03.2019. https://www.techradar.com/news/what-is-a-neural-network

Devlin, Keith. 1996. Mathematics: The Science of Patterns: The Search for Order in Life, Mind and the Universe. Date of retrieval 05.02.2019. https://en.wikipedia.org/wiki/Special:BookSources/978-0-7167-5047-5

Do, Troy. 2017. Machine Learning in Credit Underwriting. Date of retrieval 18.02.2019. https://medium.com/engineer-infinite-value-from-finite-resources/machine-learning-in-credit-underwriting-aaa3a7d10dd5

Driver and Kroeber. 1932. Quantitative Expression of Cultural Relationships. Date of retrieval 15.01.2019. http://dpg.lib.berkeley.edu/webdb/anthpubs/search?all=&volume=31&journal=1&item=5

Dwivedi, Divyansh. 2018. Machine Learning for beginners. Date of retrieval 18.10.2018. https://towardsdatascience.com/machine-learning-for-beginners-d247a9420dab

Eishawa, Ehmad. 2017. How FinTech Has Changed Underwriting Forever. Date of retrieval 19.02.2019. https://www.lendio.com/blog/small-business-insights/fintech-changed-underwriting-forever/

Feedzai. 2017. Machine Learning and AI for fraud prevention: a primer. Date of retrieval 23.02.2019. https://feedzai.com/resources/machine-learning-ai-fraud-prevention-primer/

Finance Train. How Data Science is Used in Fintech (Financial Technologies). Date of retrieval 02.03.2019. https://financetrain.com/how-data-science-is-used-in-fintech/

Freedman, David. 2009. Statistical Models: Theory and Practice. Cambridge University Press.

Goel, Amit. 2015. 7 Trends in Biometric Technology as It Applies to FinTech. Date of retrieval 23.02.2019. https://gomedici.com/7-trends-in-biometric-technology-as-it-applies-to-fintech

Hargrave, Marshall. 2019. Insurtech. Date of retrieval 23.02.2019. https://www.investopedia.com/terms/i/insurtech.asp

Herbon, Julian. 2012. Where the term ‘underwriting’ comes from. Date of retrieval 21.02.2019. https://thebasispoint.com/where-the-term-underwriting-comes-from/

Huneycutt, Jake. 2018. An Introduction to Clustering Algorithms in Python. Date of retrieval 01.03.2019. https://towardsdatascience.com/an-introduction-to-clustering-algorithms-in-python-123438574097

Ifrah, Georges. 2001. The Universal History of Computing: From the Abacus to the Quantum Computer. Date of retrieval 10.03.2019.

Jung, Alexander. 2018. Machine learning: basic principles. Date of retrieval 10.03.2019. https://arxiv.org/abs/1805.05052

Leek, Jeff. 2013. The key word in "Data Science" is not Data, it is Science. Date of retrieval 05.02.2019. http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/

McCarthy, John. Feigenbaum, Edward Feigenbaum. 1990. Arthur Samuel: Pioneer in Machine Learning. Date of retrieval 01.02.2019. http://www.aaai.org/ojs/index.php/aimagazine/article/view/840/758

McCorduck 2004, Crevier 1993. "the conference is generally recognized as the official birthdate of the new science.", Russell & Norvig 2003, who call the conference "the birth of artificial intelligence."

Mura, Roberta. 1993. Images of Mathematics Held by University Teachers of Mathematical Sciences.

Narayanan, Arvind; Bonneau, Joseph; Felten, Edward; Miller, Andrew; Goldfeder, Steven. 2016. Bitcoin and cryptocurrency technologies: a comprehensive introduction. Princeton: Princeton University Press.

Ng, Andrew. 2017. Machine Learning by Stanford University. Date of retrieval 11.09.2018. https://www.coursera.org/learn/machine-learning/home/welcome

Ovenden, James. Blockchain Top Trends In 2017. 2017. The Innovation Enterprise. Date of retrieval 22.02.2019. https://channels.theinnovationenterprise.com/articles/blockchain-top-trends-in-2017

Parbhakar, Abhishek. 2018. Mathematics for AI: All the essential math topics you need. Date of retrieval 01.02.2019. https://towardsdatascience.com/mathematics-for-ai-all-the-essential-math-topics-you-need-ed1d9c910baf

R FAQ. 2018. Date of retrieval 02.06.2019. https://cran.r-project.org/doc/FAQ/R-FAQ.html#What-is-R_003f

R. Kohavi and F. Provost. 1998. Glossary of terms. Machine Learning.

Radke, Parag. 2017. Basic Probability Theory and Statistics. Date of retrieval: 30.06.2019. https://towardsdatascience.com/basic-probability-theory-and-statistics-3105ab637213

Rathi, Ram. 2018. Basic Logistic Regression. Date of retrieval 19.03.2019. https://github.com/ramrathi/IECSE-ML-Winter18/wiki/Basic-Logistic-Regression

Romeijn, Jan-Willem. 2014. Philosophy of Statistics. Date of retrieval: 30.06.2019. https://plato.stanford.edu/entries/statistics/

Rouse, Margaret. 2016. Supervised learning. Date of retrieval 21.02.2019 https://searchenterpriseai.techtarget.com/definition/supervised-learning

Schickard, Wilhelm. Wilhelm Schickard – Ein Computerpionier. Date of retrieval: 28.05.2019. http://www.fmi.uni-jena.de/fmimedia/Fakultaet/Institute+und+Abteilungen/Abteilung+für+Didaktik/GDI/Wilhelm+Schickard.pdf (PDF) (in German).

Schmidhuber, J. 2015. Deep Learning in Neural Networks: An Overview. Neural Networks. Date of retrieval: 04.05.2019.

Schüffel, Patrick (2016). Taming the Beast: A Scientific Definition of Fintech. Journal of Innovation Management. Date of retrieval: 30.11.2018.

Scott-Briggs, Angela. 10 most popular fintech sectors. Date of retrieval 18.02.2019. https://www.techbullion.com/10-popular-fintech-sectors/

Seif, George. 2018. The 5 Clustering Algorithms Data Scientists Need to Know. Date of retrieval 01.01.2019. https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Sullival A. 2018. 10 Big Financial Technology Trends for 2018. Date of retrieval 18.02.2019. https://thefinancialbrand.com/69779/financial-technology-trends-data-ai-blockchain/

University of Helsinki. Elements of AI. 2018. Date of retrieval: 30.03.2019. https://www.elementsofai.com/

van Otterlo, M., Wiering, M. 2012. Reinforcement learning and Markov decision processes. Reinforcement Learning. Adaptation, Learning, and Optimization. Date of retrieval: 30.04.2019.

Webster's Revised Unabridged Dictionary. Date of retrieval: 28.06.2019. https://web.archive.org/web/20150428142545/http://machaut.uchicago.edu/?resource=Webster%27s&word=probability&use1913=on

Wertheimer, Eric. 2006. Stanford University Press. Underwriting: The Poetics of Insurance in America, 1722-1872. Date of retrieval: 30.11.2018. https://books.google.com/books?id=ZCu4ctqn9OsC&pg=PA25

Yegulalp, Serdar. 2017. 11 open source tools to make the most of machine learning. Date of retrieval 20.02.2019. https://www.infoworld.com/article/2853707/11-open-source-tools-machine-learning.html

Zigurat. Evolution of Fintech. Date of retrieval 19.02.2019. https://www.e-zigurat.com/innovation-school/blog/evolution-of-fintech/