Open University Cyprus **Hellenic *Open University***

Master's join degree/post graduate Programme
Enterprises Risk Management (ERM)

MASTER THESIS

Using Big Data Techniques in Risk Management

Antigona Marku

Supervisor

Pandelis Ipsilandis

May 2018

Open University Cyprus **Hellenic *Open University***

Master's join degree/post graduate Programme
Enterprises Risk Management (ERM)

MASTER THESIS

Using Big Data Techniques in Risk Management

Antigona Marku

Supervisor

Pandelis Ipsilandis

This thesis was submitted for partial fulfillment of the requirements
Master's join degree/post graduate programme
«Enterprises Risk Management (ERM)»
Faculty of Economics and Management

Open University of Cyprus **Hellenic Open University**

May 2018

Acknowledgements

I would like to thank my supervisor Mr. Ipsilandis for the help and the guidance that he provided me during my master thesis, as well as my teachers from the master program for the excellent collaboration we had and the knowledge they shared with me.

I would also like to say a special thank you to my family, my parents and my brother who supported me in every possible way in order to reach my goals.

Finally, I would like to thank my friends and my classmates for standing by my side in every difficulty that arisen to me.

Contents

List of Tables

List of Figures

SUMMARY

In a world of increasing complexity and demand the ability to capture access and utilize Big Data will determine risk management success. Risk managers haven't understood its power yet. In this thesis we will analyze a quantitative and qualitative research about the role of big data in supporting decisions regarding risk management.

The purpose of the thesis is to show how Big Data technologies can help Risk teams gain more accurate risk intelligence, drawn from a variety of data sources in almost real-time. In this thesis I will implement big data software in a case study with real data to proactively detect potential risks, like credit card fraud and show how financial service industry can benefit from analyzing big data in order to make a better mitigation strategies and better strategic decisions.

The most significant factor in an effective analytics is a good data. Good data will always be able to safe mediocre analytics performance, but poor data will also bring poor performance. This research will show the importance of big data in risk management and how big data could add 'high-veracity' and 'high-value' to analysis which will drive to reliability and offers potentially significant cost savings by combating the risks that can cost financial institutions billions.

The main part of the thesis begins with the theoretical approach of big data and makes a reference to big data techniques in combination with the risk management. Subsequently, I will develop a model for a case study using big data software.

Chapter 1

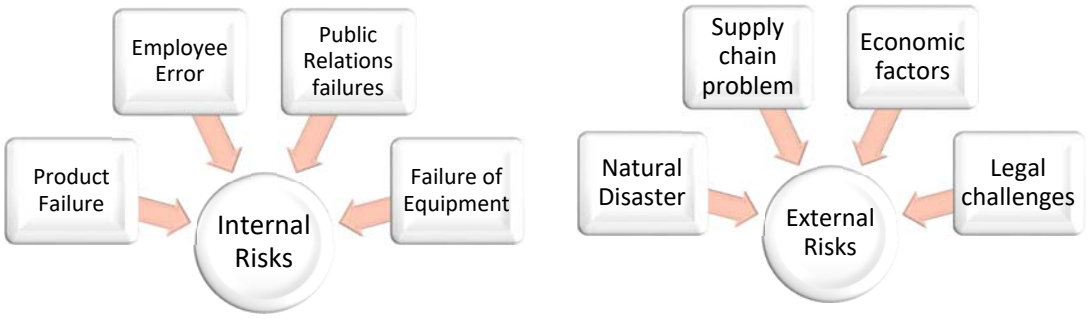
Introduction to Risk Management

The last decade the globalization of companies is a true fact and experienced an upward trend, as a consequence companies should think and act in an international level and peruse global developments and events and try to identify the risks that arise from them. Our era is an era of uncertainty interdependence so this leads us in a greater effort to identify risks and find solutions to avoid them. But first we should know what a risk is. In a theoretical way risk is the impact of uncertainty on people or organizations, which may have harmful or positive outcomes that can impact business objectives as well as its reputation. A classical definition of risk is [Fraser, J., Simkins, B.J., 2010]:

''Risk is proportional to measure for the probability P of the occurrence of an event and its consequences, C in the case the event does occur '':

$$R=P*C$$

A risk can create opportunities and can help deliver business objectives, it is a fact that organizations cannot develop without taking risks. Risk can occur within the business for example due to product failure, employee error, public relations failures, failure of equipment or can be developed externally like supply chain problems, natural disaster, economic factors, legal challenges etc (Figure 1).



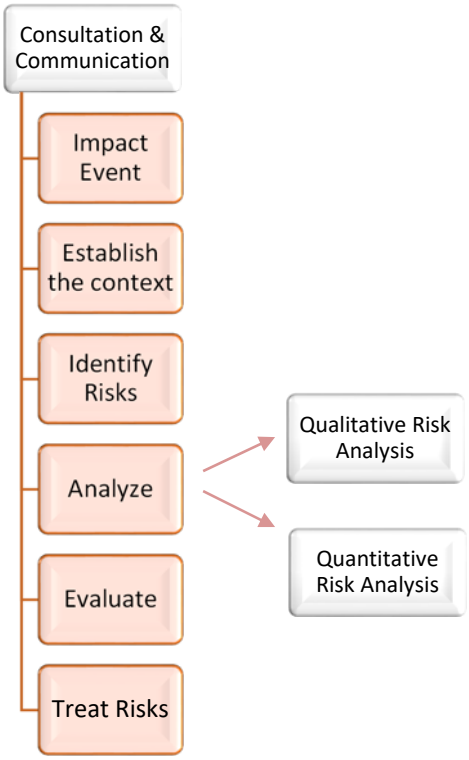
Source: J.Davidson Frame, (2003)

Figure 1. Internal & External Risks

Risk is inevitable, there will always be a level of probability that things will not work out as expected or that something unexpected will happen. Although the types of risks vary from business to business we can predict the likelihood of some risks occurring and we can measure the impact of these risks on the business. These types of risks that are high-priority due to their impact and likelihood are presented below:

- *Financial Risk* is one of the most high-priority risks that involves financial losses to businesses, arises due to instability and losses in the financial market.
- *Market Risk* is a change in value of assets due to changes in the underlying economic factors such as interest rates, foreign exchange rates, macroeconomic variables and stock prices.
- *Credit Risk* is a type of financial risk; it is the change in value of a debt due to changes in the perceived ability of counterparties to meet their contractual obligations.
- *Strategic Risk* is a loss that arises from an unsuccessful business plan, might arise even from making a poor business decision or from a failure to respond to changes in the business environment.
- *Operational Risk* is the risk of loss resulting from inadequate or failed internal processes.
- *Reputational Risk* is a threat of losing the good name of the business, it is a risk that can happened in many ways, as a result of the actions of the company, it can also arise from the actions of employees, by perceiving damage to the environment caused by business' extraction activities etc.
- *Compliance Risk* it is the danger of not managing to meet regulations within a workable timeframe and budget.

There are many risks and in different kind as a result it is important to be analyzed by people with skills and expertise in each domain and then brought together to form a complete view of the risks that an organization faces, this is where the Risk Management takes place. The purpose of Risk Management is about managing the impact of uncertainty specifically, can help organizations and people protect themselves and provide confidence that the ways in which they respond to risk are good enough to meet their needs [National Cyber Security Center, 2016]. Risk Management process consists of 5 steps which are quoted in the Figure 2, but the most important step is the risk analysis from which you drive ways to handle risks.



Source: Fraser, J., Simkins, B.J. (2010)

Figure 2. Risk Management Framework

At the end of this risk management process the business should have a list of risks attached to the business, the next and more important step is to analyze and attempted to remove the risks. Risk analysis is important for businesses because it helps anticipate and neutralize possible problems when planning projects and also helps decide whether or not to move forward with a project. In addition a business can benefit from risk analysis when planning for changes in its environment such as competitors coming into the market or changes to government policy. Finally, it is useful in managing a potential

risk in the workplace or in preparing for events such as equipment failure and natural disasters.

As we see above there are two categories of risk analysis, qualitative risk analysis and quantitative risk analysis. The two methods are different because the quantitative analysis uses a numerical scale on the other hand qualitative analysis uses a descriptive scale to measure probability. Specifically, a qualitative analysis may use one of three techniques:

- *Interviewing techniques* are used to quantify the probability and impact of risks on a project. The information needed depends upon the type of probability distributions that will be used, for example information would be gathered based on a risk matrix that has a scale of Low, Medium, High to indicate the likelihood and impact of risk event occurring (Figure 1).
- *Probability distributions* represent the uncertainty in values, for instance durations of schedule activities and costs of project components.
- *Expert judgment* from internal or external statistical experts validates data and techniques.

Also a qualitative risk analysis will help you determine if there is any type of risks that would require special attention or any risk event that need to be handled in the near term.

Likelihood	**Impact**				
	Insignificant	Minor	Moderate	Major	Severe
Almost Certain	Moderate	High	High	Extreme	Extreme
Likely	Moderate	Moderate	High	High	Extreme
Possible	Low	Moderate	Moderate	High	Extreme
Unlikely	Low	Moderate	Moderate	Moderate	High
Rare	Low	Low	Moderate	Moderate	High

Source: Fraser, J., Simkins, B.J. (2010)

Figure 3. Impact & Likelihood Matrix

One the other hand a quantitative analysis will determine the probability of each risk event occurring. Below we present three modeling techniques of quantitative risk analysis:

- *Sensitivity analysis* determine which risks have the most potential impact on the project and also which variation of a project element affects a project objective when all the other uncertain elements are held at their baseline values.
- *Expected monetary value* calculates the probability of each possible outcome and determines the average value of all outcomes.
- *Decision tree* is an analytical method that is used in case that we have to choose between two or more alternatives, it describes a decision under consideration and the implications of choosing one or another of the available alternatives.
- *Simulation* uses a model that translates the uncertainties specified at a detailed level into their potential impact on objectives at the level of the total project.

Some tools that are used in the process of quantitative analysis to measure the risk are standard deviation, beta, Value at Risk (VaR) and Conditional VaR.

- ✓ ***Standard deviation*** it is a measure that it is used to scatter the data from its expected value. It is also used from investors in making an investment decision to measure the amount of historical volatility or risk associated with an investment relative to its annual rate of return.
- ✓ ***Beta*** measures the amount of systemic risk a security has relative to the whole market. The market has a beta of 1, if security's beta is equal to 1 the security's price moves in time step with the market, one the other hand if security has a beta greater than 1 indicates that it is more volatile than the market, conversely if a security's beta is less than 1 it indicates that the security is less volatile than the market.
- ✓ ***VaR*** is a statistical measure that is used to assess the level of risk associated with a portfolio or company; it measures from a specified period the maximum potential loss with a degree of confidence.
- ✓ ***Conditional VaR*** is a measure that estimates the likelihood of a possible break in the VaR with a certain degree of confidence.

Although, traditional analyzing techniques and tools have been used from many companies all this years, nowadays are not enough anymore due to the fact that risk management faces new demands and challenges [EY. Insights on Governance, Risk and Compliance, 2014]. Modern organizations do not want to know what happened and why it happened, but also want to know what is happening right now and what is likely to happen in the future. For example banks and regulators expect more detailed data and

increasingly sophisticated reports. In addition, money-laundering scandals and fraud cases have prompted further industry calls for improved risk monitoring and modeling.

Traditional techniques and methodologies are the most spreading ones as they rely on classical tools which allow an ad hoc approach and which need to cross several sources of information.

As a consequence while many of the techniques to process and analyze data types have existed for some time, the massive demand for all this information and all these rapid technological developments combined with the decreased cost of computing models have encouraged broader adoption of new data analysing techniques that collaborate with new computing techniques like ''Big Data''. Big data present fresh opportunities to address these challenges and also has the potential to improve monitoring of risk.

Chapter 2

What is Big Data

2.1 Introduction to Big Data

We live in a world where data is exposing in unprecedented velocity, variety, volume and veracity (definition 2.1), it is now available almost instantaneously and it creates possibilities for near real-time analysis. There are a lot of sources now available for example business apps, social media, email documents etc, and the speed required for retrieval and analysis that with their complexity and variety bring fresh new approaches. Even though big data is already being embraces in many fields risk managers have not yet recognize its power. The ability to capture and access big data will determine risk management success. Big data technology has revolutionary potential; it can improve the predictive power of risk models, improve risk coverage and generate significant cost savings as it can use both structured and unstructured data. This system allows users to make analytical decisions all based on real time information.

2.1 Definition

Big data refers to the large, dynamic and disparate volumes of data being created by people, tools and machines. They are impossible to process with traditional methods as they require new and innovative technology instead. The turning of more businesses into e-commerce further enhances data production, larger e-commerce companies accept in a matter of minutes thousands of transactions from customers so it is important an analytically process the vast amount of data gathered in order to derive real time business insights that relate to consumers as well as to risk, profit, productivity etc [Oracle Enterprise Architecture White Paper, 2016].

2.2 Dimensions

Big data are defined by four features that are called the four ‘’V’s’’ (Figure 4) :

2.2.1 Volume

The magnitude of the data determines whether it will be consider as big data or not, the amount of data being created that companies use for analysis is not a sample but the whole data.

2.2.2 *Variety*

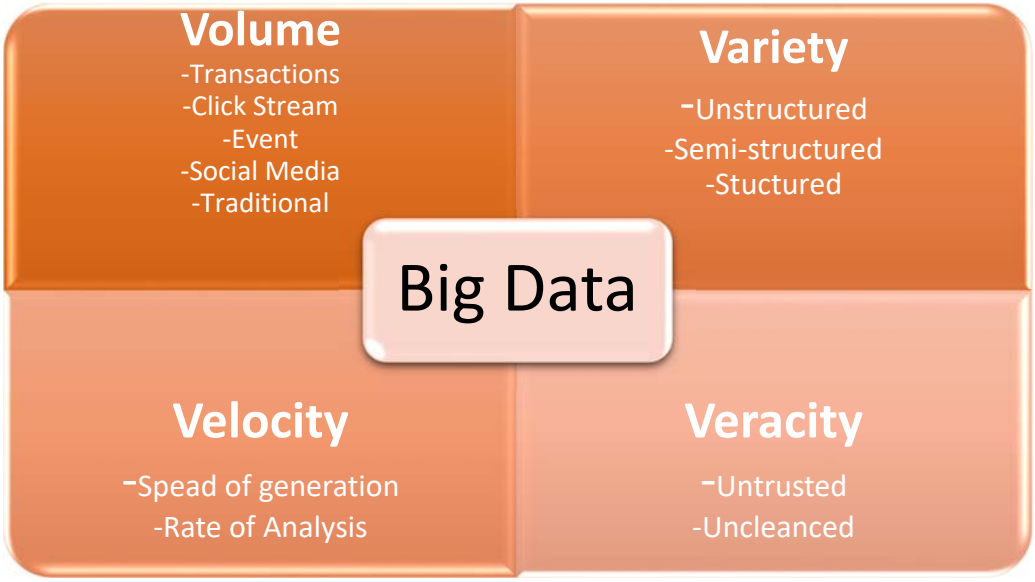
The different source from where data comes from make them varies, for example we want in social media to publish a thought and a photo to share a moment with friends. Generally, data is being created by machines, people, socials as well as transactions.

2.2.3 Velocity

Data is being generated extremely fast. Companies have the opportunity to analyze in real time, for example a customer who has seen a product on a company’s website but left the site without bought it will be target by company by seeing an advertisement of the product in the next website that he will visit.

2.2.4 Veracity

The data and their meaning are constantly changing due to the fact that data is sourced from many different places; as a result we need to test their quality. For example companies and also politicians are using social media to make an emotion analysis by trying to understand in millions of posts if the opinion expressed is interpreted as negative, positive or neutral in the media from automatic tracking. Despite the difficulty some companies are already trying to implement mechanisms to analyze the opinions of users that are set on daily basis in social media.



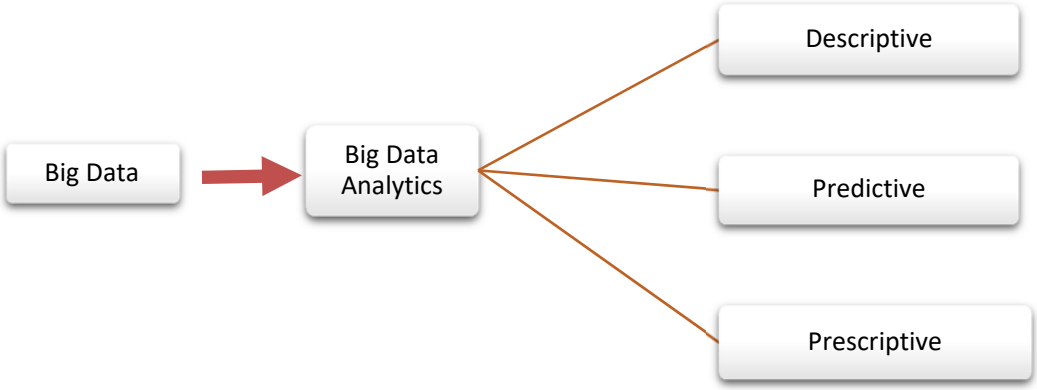
Source: EY. Insights on Governance, Risk and Compliance, (2014)

Figure 4: The four V's

2.3 Big Data Analytics Methods

Raw data is a set of numbers or characters which is sampled and analyzed for extracting decision information, these data sets are large and complex as much as it becomes difficult to process using on hand database management tool as a consequence specialized techniques and technologies are needed to analyze very large sets of data. Big data is an inevitable trend that all organizations will eventually need to be involved in, all enterprises will end up doing decision support by collected, stored and managing big data. Although the biggest challenge is not about managing them but also analyzing them and extract from them an upshot that will benefit managing decisions and especially the risk managing decisions.

Big data analytics is the process of exploring and analyzing big data in a way that you can extract hidden and valuable information and patterns and use it in decision-making process for your business. As a consequence it is vital to find the right software and most suitable techniques to analyze them. Big data works in the presence of unstructured data and techniques of data analysis that are structured to solve problem. Big data analytics can be classified as **descriptive**, **predictive** and **prescriptive** [Saumyadipta Pyne. B.LS. Prakasa Rao S.B. Rao, 2016] (Figure 5).



Source: *Saumyadipta Pyne. B.LS. Prakasa Rao S.B. Rao, (2016)*

Figure 5: Big Data Analytics Diversion

2.3.1 Descriptive Analytics

Descriptive analytics drills down into massive data, specifically historical data to extract patterns like variations in operating costs, sales of different products, customer buying preference. Analytics involve simple techniques such as regression (Table 1) to find correlation among various variables to identify trends in the data and visualize data in a meaningful way. For example enterprises can calculate the level of customer satisfaction that affect customer loyalty, also call centers can forecast the number of supports calls received which may be influenced by the one that they received the previous day.

Regression analysis it involves manipulating some independent variable to see how it influences a dependent variable [Jack Johnston and Dinardo, 1997]. It is devided into:

Table 1. Linear & Logistic Regression Analysis

Linear regression → $Yj = \beta o + \sum_{i=0}^{n} \beta i Xij$ which, is used to establish a relationship between dependent and independent variables and the expected value of the dependent variable given the values of the independent explanatory variable.
Logistic regression→$P(Y = 1) = \frac{1}{1+e^{-bo+\sum bixi}}$, which, is used to ascertain the probability of an event occurring, the event is captured in binary format, i.e. 0 or 1.

2.3.2 Predictive Analytics

Predictive Analytics can be divided in three actions *forecasting*, *prediction* and *scoring*. Specifically, it helps to combine massive data from different sources with the advantage of understanding what happened in the past and predicting future trends and events. Predictive analytics can predict the future by generating prediction models and forecasting trends. For example financial institutions invest a lot of resources in predicted credit risk for companies or individuals, also other organizations are able to forecast their sales trends or their performance and to predict customer behavior to target appropriate promotions.

2.3.3 Prescriptive Analytics

Prescriptive Analytics it is a new analytic method that involves techniques such as optimization, numerical modeling and simulation. It helps the professional in assessing the impact of different possible decisions. For example, it is used from explorers to the exploration process to optimize drilling location and at the same time explorers they avoid the cost and effort of unsuccessful drills. In addition oil and gas exploration industries are benefiting from applying big data perspective analytics as they use it in optimizing the exploration process. And also health care centers use prescriptive analytics in analyzing patient's medical history, allergies, medicines, environmental conditions etc.

2.3.4 Approaches of Data Analysis

Data Analysis can be divided in different approaches considering the project that they implemented to [Debbie Stephenson, 2013].

- Association Rules Analysis

Association Rules Analysis is one of the main techniques that are used by large retailers and enterprises to uncover associations between items; it allows retailers to identify relationships between the items that people buy, so for example if people buy milk and butter are more likely to buy diapers. It is also used to extract information about visitors to websites in purpose of increasing sales.

- Decision Tree Analysis

Decision Tree Analysis is a statistical method to identify categories that a new observation belongs to. Specifically, it is a tree-shaped graphical representation of several decisions followed by different chances of the occurrence; you can lay out options and investigate the possible outcomes of choosing those options.

- Genetic Algorithms

Genetic Algorithms are adaptive heuristic search algorithms which make uses techniques inspired from evolutionary ideas of natural selection and genetics by the way evolution works. It is an ideal mechanism that can be used in solving problems that are not well suited for standard optimization algorithms, including problems in which the objective function is discontinuous, no differentiable, stochastic or highly nonlinear.

- Machine Learning

Machine learning is a method that gives computers the ability to learn without being explicitly programmed. Machine learning has given us self-driving cars, has helps us distinguish between spam and non-spam email messages, has also help us making effective web search, and a vastly improved understanding of the human genome, as well as it determine the best content for engaging prospective customers. It is in every step of our daily lives, we probably use it lots of times a day without knowing it.

- Time Series Analysis

It is a method that arise when monitoring industrial processes or tracking corporate business metrics. The modeling and forecasting procedures discussed in identifying patterns in times series data involved knowledge about the mathematical model of the process. However, in real-life research we still need both to uncover the hidden patterns in the data and also to generate forecasts due to the fact that individual observations involve considerable error.

- Sentiment Analysis

Sentiment analysis is an analysis technique that analyzes opinionated text which contains people's opinions toward entities. It is used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention. Specifically we can say that is a process that determines the emotional tone behind a series of words. It is an analytical method that is very useful in social media monitoring due to the fact that

it allows us to gain an overview of the wider public opinion be certain topics, is a practice that is being widely adopted by organizations across world and sentiment analysis it's very helpful because it is broad and powerful in extracting insights from social media.

- Social Network Analysis

This method is used by Management consultants about their clients because it provides both visual and mathematical analysis of human relationship. In general we can characterize it as a mapping and measuring of relationships and flows between people groups, organizations, computers and other connected information entities.

All those analytical methods and techniques need the right tools and platforms to be implemented effectively. The use of open source tools is very important for big data analytics, as they contribute in data preparation, data exploration, data visualization, data modeling and data reporting. Below the top open source tools and platforms for big data analytics are presented [Saumyadipta Pyne. B.LS. Prakasa Rao S.B. Rao, 2016]:

2.3.5 Development Platforms

Big data platforms refer to a crowd of servers, storages, databases and other utilities that enable organizations in developing, operating and managing big data environment without complexity of multiple solutions, by just providing one unique cohesive solution [Hussein A.Abbass, 2015].

Apache Hadoop

A popular data storage and analysis platform, an open-source software framework for distributed storage of very large datasets on computer clusters which was produced by Yahoo. Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless. It gives also the ability to store any kind of data from any source and it can do very sophisticated analysis of that data easily and quickly. It is powerful enough to run sophisticated detection and prevention algorithms and to create complex models from historical data to monitor real-time activity. Hadoop can make networks more robust, if one cluster fails, it continues to run, can also provide scaling in the ideal case by enabling easier design and finally doesn't require applications to send amounts of data across the network.

MapReduce

MapReduce is a programming framework for processing large datasets in distributed environments, actually is a simple programming model for processing huge data sets in parallel. MapReduce takes large datasets, extracts and transforms useful data, distributes the data to the various servers where processing occurs, and assembles the results into a smaller, easier to analyze file. All in all we can sum up that the main purpose of MapReduce is to divide a task into subtasks, handle the subtasks in parallel, and aggregate the results of the subtasks to form the final output.

Apache Spark

Apache Spark is an open source cluster computing system that can be programmed quickly and runs fast; it is usually being used for research and production applications. Spark can be used to interactively query 1 to 2 terabytes of data in less than a second. It is a cluster computing framework that makes data analytics faster to run and also to write to distributed file systems. Spark supports RDDs which is a collection of objects spread across a cluster and stored in RAM or on disk. Spark provides an easier and alternative way in contrast with Hadoop and MapReduce and also offers performance up to 20 times faster than previous generation systems for certain applications. Combining these two frameworks together provides opportunities to solve credit card fraud like using big data analytics. Spark's advantages include an integrated advanced analytics; continuous micro-batch processing based on its own streaming API and is more efficient and faster than Hadoop and MapReduce [Donna-M.Fernandez, 2016].

Apache Flink

The Flink is an open source stream processing framework whose algorithms is appropriate for streaming and batch modes [Donna-M.Fernandez, 2016]. It is developed by the Apache Software Foundation. Due to its ability to process streaming data in real time is a ideal software for big data analytics, thus it is quickly with low data latency and high fault tolerance on distributed systems on large scale. One very important advantage of Flink program is that in the case of a failure with check pointing enabled will, upon recovery, resume processing from the last completed checkpoint.

2.3.6 Development Tools

In order to analyze and implement the deferent big data techniques that we mentioned before, open source tools are needed to handle the huge amount of datasets and to get

some significant value from them. In our opinion the most important and easy to use are those below:

Apache Mahout

Apache Mahout is a programming framework that is used to produce free implementations of distributed machine learning algorithms focused mainly in linear algebra and mathematically expressive DSL with the purpose of helping statisticians and data scientists implement their own algorithms. Mahout makes it easier and faster to turn big data into big information and find meaningful patterns in big data sets.

Python

Python is an open source programming language with a lot of utilities and libraries for data processing and analytics tasks. It is a flexible and powerful programming language which has the ability to integrate itself with web applications. Also Python it is an interpreted and portable language in which the variable types are defined automatically. For all this reasons above we can admit that python is a powerful language that actually can support the need of the business.

R

And last but not least R language program which is the open source that we will use and in our case study for credit card fraud detection. R programming is open source software available under software foundation; it is a statistical programming language which is used in data science and data analytics. Specifically, it is a one step process to calculate statistics (mean, variance, median), static graphics (plots, graphic maps), probability distributions (beta, binomial), to implement data mining and to visualize the data. There are no techniques that could not be implemented by R packages.

However, even though the traditional approaches are providing slow in identifying risks from huge datasets due to the increasing complex of environment and the increasingly creative methods of fraud that attackers use, which makes the detection harder and the analysis of risk cases more complex, unfortunately firms are still using traditional descriptive techniques for forensic data analytics according to the research we made, the results are presenting in table 11. We asked 47 companies from different sectors to inform us about the data analytic tools that they use in the analytic process. (Appendix 1A)

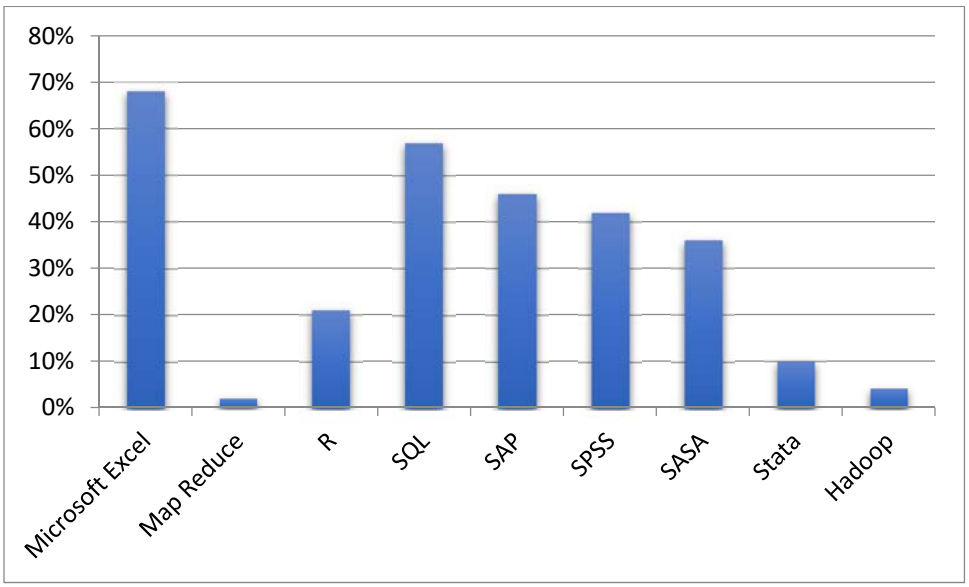


Figure 6. Proportion of each software tool use

These data evidence are disappointing since as we observe the Greek companies are not familiar with new technology software yet, even though the most countries nowadays deal with huge volumes of sensitive online data and are well aware of cyber fraud. Managers should persuade that big data and its real-time abilities are strong assets and enable companies to be more efficient in risk management. Big data analytics involves the ability to gather massive amount of digital information from structured and unstructured data sources to analyze, visualize and draw insight that can make it possible to predict and handle every type of risk. All these data are high volume, so the best way to analyze them is to apply big data analytics.

The techniques and analytics tools that we previously mention can create new opportunities for improving risk detection and prevention, can help the investigator in extracting meaningful and purposeful forensic evidence for detecting frauds from the large datasets. Specifically, algorithms can be developed to be more precise and reduce false positives and negatives that occur from traditional methods.

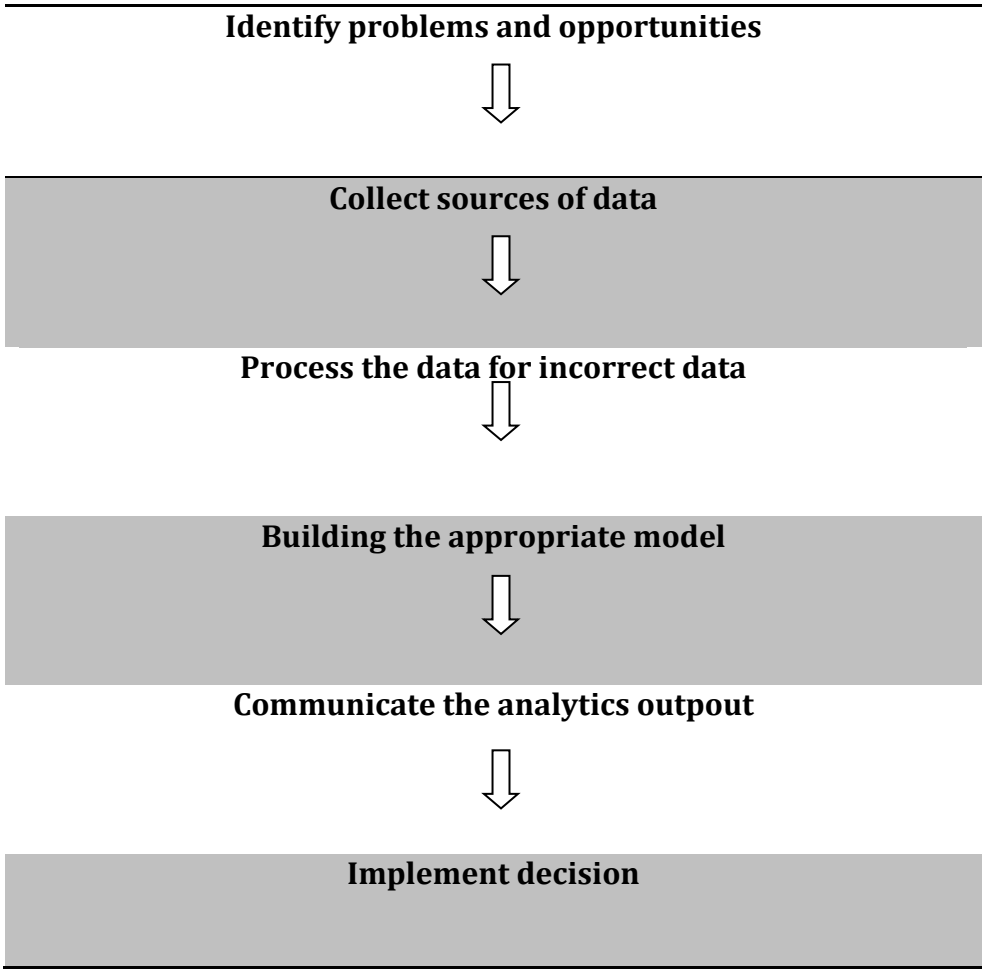
Chapter 3

Big Data and Decision Making

Nowadays enterprises depends on evaluating risks and then acting on those insights, more information should yield better risk assessments, which is why big data couldn't have arrived at a better time, big data present fresh opportunities to address the new challenges that risk management faces due to the crisis nowadays, even though enterprises have been skeptical about using big data. Enterprises are aware of the big data phenomenon and its potential benefits for risk management but they lack specific information to decide about implementing big data techniques due to the fact that big data solutions are on an early stage and enterprises are waiting for more proof of concept. Moreover, as most of the information technology investments, big data are costly because they are based on sophisticated technologies and enterprises have to pay the price to get such advanced technologies and replace their existing technology. However, some enterprises across various industry verticals have overcome these issues and have begun to leverage investments in big data analytics [Technavio, 2014]. Characteristic examples include:

- ***Microsoft***: The company has a special focus on providing big data solutions to SMEs.
- ***SAP AG***: It is a large software company that is its featured solutions include big data and rapid deployment solutions.
- ***Google Inc***: The company provides big data analytics through its Big Query platform.
- ***Amazon***: It is a website that has access to a huge amount of data like payments, customers' manes, search history etc, that have been used to improve customer relations and in advertising.
- ***American Express***: It is a company that uses big data to forecast consumer behavior in transactions and to predict customer loyalty.

Organizations need to use structured information to improve their decision-making process. To achieve this structured view they have to collect and store data and then analyze them and transform the results into useful information. To perform these analytical and transformational processes it is necessary to make use of an appropriate environment composed of a large repository. This repository must be filled with data originating from many different kinds of external and internal data sources and most recently big data. We can sum the process of decision making in the following figure.

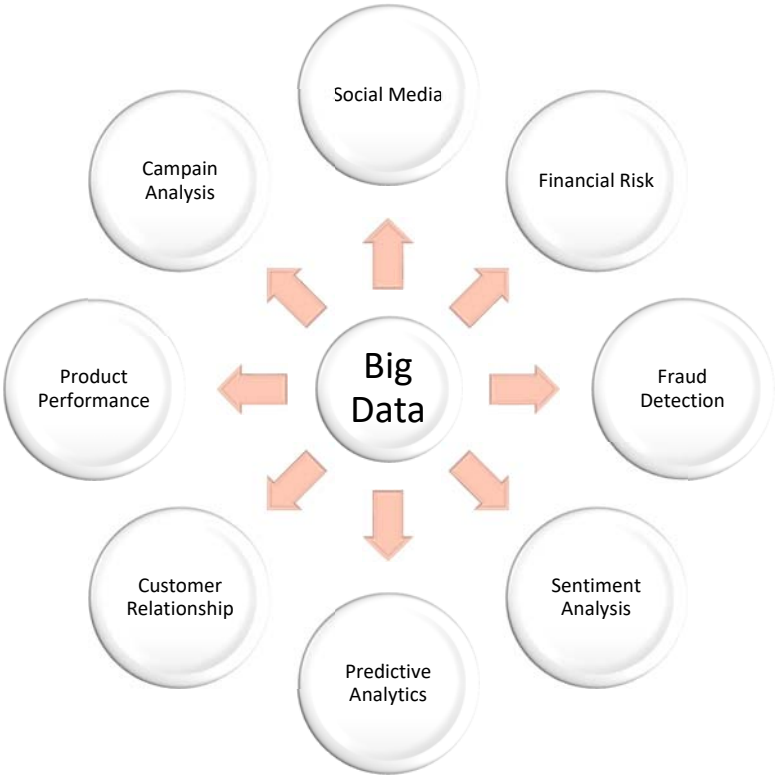


Source: Boris Delibasic, Jorge E.Hernandez, Jason Papathanasiou, Fatima Dargan, Pascale Zarate, Rita Ribeiro, Shaofeng Liu, Isabelle Linden, (2015)

Figure 7. Decision making process

The use of big data can lead to increase system intelligence Hussein [A.Abbass, 2015] which can be used to improve performance in sales, reinforce the internal risk management function and enhance fraud monitoring, product performance, it also

opens opportunities to ingest, stores and process data from new sources such as market data, social media data, communication etc. (Figure 8).



Source: *Hussein A.Abbass, (2015).*

Figure 8. The sectors where big data is used

With big data concept and its applications to risk management, enterprises can improve monitoring of risk coverage and the stability and predictive power of risk models, these models will support everyday Risk Officer Decision-making. Also it will help finance and accounting managers in budgeting, forecasting and planning according to risk measuring. Marketing managers who use big data tend to managed their goals as they check the market, source data from the systems, understand what consumers want and create a model and metrics to test the solutions apply the results in real time, and HR teams use big data to better predictive analytics [Thiago Poleto, Victor Diogho Heuer de Carvalho and Ana Paula Cabral Seixas Costa, 2015].

Specifically, below we present some typical decision scenarios [Mark van Rijneman, 2015] that departments of an enterprise can benefit from big data:

Table 2. Decision scenarios

Department	Tasks	Desicion Scenarios	Big data Implementation Benefit
Financial and Accounting	Credit controll & Collection	Is the customer we want to collaborate with credit worthy?	Identify bad credit risk
Sales	CRM	Where to find new and premium customers?	Identify customers with higher potential value
Marketing	Promotional planning and brand building	Which responses are to be expected on marketing campaigns in different countries?	Enable finer grained customer segmentation
H&R	Recruitment and training	How to find and hire the best talents?	Determine optimal job candidates
Logistics	SCM & purchase planning	How to identify new savings opportunities?	Faster price adjustments due to changing markets
Operation	KPI's for each employee in every project	How many employees are needed for each project?	Better forecasting due to better predictive models

In addition the use of big data technologies it overcomes traditional restraints in cost effective manner by enabling businesses to store data at the lowest level of detail,

keeping all data history under reasonable costs and with less efforts. For example, with a lower cost of storage per gigabyte it enables organizations to have a federated view of customers by shifting customer data from various separate business departments into a single infrastructure and then to run consolidated analytics and reporting on it.

To sum up we can briefly quote the key benefits and disadvantages of using big data analytics:

☑ **Advantages of Big Data analytics**

1. *Real time detection of errors and possible system violations*

In case of an error on security system, production or any administrative process of the enterprise with the use of big data analytics the error will be perceived automatically. By using such data techniques enterprises will manage to avoid similar phenomena in the future and this will lead to an ever-increasing acquisition of customers.

2. *Saving Resources and progressive services*

The quantitative methods of big data can offer quality services and better sales rates as a consequence and higher profits, due to the fact that with continues monitoring mistakes can be avoided.

3. *Comparative advantage towards the competitors*

The use of big data analytics makes it possible for enterprises to have a comprehensive picture if their competitors, as a consequence they can implement a better strategic plan.

4. *Additional knowledge of customers' spending patterns*

With big data analytics enterprises san analyze the spending patterns of consumers. The knowledge of consumers' trends can contribute in understanding the causes of lost sales and help in long-term planning about services and products that are more desirable from consumers.

☒ **Disadvantages of Big Data Analytics**

1. *New approach*

Many businesses in their up to date technologies were studying their data rarely, but

in case of big data is vital the monitoring in daily bases, because as we mentioned above big data are used daily in making decisions by enhancing businesses' strategies.

2. Possible failure in adapting big data analytics

For many businesses the study of big data is a big challenge, if a business try to apply the big data techniques without being ready to handle such a large volume data will have costly consequences.

3.The cost of adapting big data

Implementing this new data technologies can be time and cost spending for businesses, as they have to change their management, it requires recruiting new skilled people and also managing international data can cause privacy issues.

Despite of the disadvantages we can admit that organizations who embrace big data tools not only will see significant first mover advantages but will be considerably more agile and cutting edge in decision-making, as these businesses can decide faster, react more flexible, act not only for present but also predictive and take decisions outside the box.

Chapter 4

Research

4.1 Research Questions

The purpose of this methodology chapter is to describe the chosen methodology and it includes information about the steps and data gathering tools used for this thesis. Specifically, in this part we will discuss how the conceptual model will be tested. We will start by quoting the empirical steps that we followed with a choice for a methodological classification. Furthermore, the different steps will be elaborated, as a conclusion at the end of this part, it will be clear what and how will be measured to give an answer to the research question below:

- ***What is big data?***
- ***Why is big data analytics important for risk management?***
- ***How risk management can benefit from big data analytics?***
- ***Which is the contribution of open source tools in modeling development?***

4.2 Research Type

In order to determine the existence of big data and their definition and to understand what it really is, we made a review in the bibliography and scientific articles and we collect examples of them.

To measure the importance of using big data techniques in risk management and how big data technologies support decisions regarding managing risk I conduct a secondary research where we found industry examples of big data across the financial and related sectors, to do this we used academic books and articles, internet articles, we collect vendors commercial information and conference notes. Furthermore, we filtered the

industry example by grouping the cases and presenting the cases that are benefited from big data technologies. We also made some scenarios to show how departments of an enterprise can benefit from big data analytics (Table 2). Then we made a research by sending e-mails to several companies from different industrial sectors about the open sources tools that they use, the questioner (Appendix 1) was about the analytical software tools that are related to big data and with the result we understood that Greek companies are not yet familiar with new technologies.

Finally, we made an internal review, where we form comprehensive use case by implementing a model, in which will test the relation between different variables over time and could be controlled for other variables. Specifically, in our case we choose a risk that is vital for banks and generally for whole the market, as it affects negatively the merchants and also the consumers, it is credit fraud detection. We will try to show how risk analysis and modeling development can benefit from new technology software and especially with open source tools like R language. In order to detect credit fraud risk we will analyze certain factors and use logistic regression for modeling. We will implement step by step all the process of analyzing a model logistic regression with the traditional method compared to analyzing and implementing models in R, we will try to prove how easy and fast is to apply open source tools in analyzing and visualizing. The process of creating a classification credit card fraud detection model can be summed up in the following steps:

a) Selecting data
b) Installing the appropriate software in our case is R language
c) Creating data samplings for model estimation and testing
d) Model estimation
e) Model tests

4.3 Data Collection

During the data collection, the different variables are measured for the sample during the measurement period, actually is a dataset that is consisted of 284.807 transactions that occurred in two days in September 2013 by European cardholders. This is done with the use of several sources. Most of the information that is needed for the measurements could be drawn from annual reports that are provided by the website

https://www.kaggle.com/. Unfortunately, due to confidential issues we were not able to get access to nowadays transactions.

4.4 Methodological classification- Analysis of data

Analysis can be done using many ways but usually it is limited to the amount of data set size. It is unable to visualize large amount of data at a time. This project credit card fraud detection analysis overcomes the above mentioned issue in an efficient way by using R programming language. This proposed system can analyze and visualize large amount of data. In this case firstly the credit card data from the banks databases are collected. The credit card data which is collected is then converted into a pictorial format in R language. It is then reported as an analysis report and visualization report. Then, we will use descriptive statistics, which will be briefly discussed. These descriptive statistics will give an overview of the sample. When this is completed, the different measures could be related to each other. This will be done using Fisher's correlation, followed by logistic regression. From this logistic regression, the hypotheses could be tested and conclusions could be drawn. Thus, before final conclusions could be drawn about the research questions, the validity needs to be tested and discussed.

The steps of the model development are the follow:

4.4.1 Descriptive statistic

To give an overview of the sample, descriptive statistics will be given. First, the total number of bank observations in each period will be given, and the total number of observations in the sample. Then, the overall average, mean and standard deviations of the different measures will be given for the entire period. Further, these numbers will also be given for each measurement period.

4.4.2 Logistic regression

Logistic regression is a well-established statistical method for predicting binominal or multinational outcomes. Multinomial logistic regression algorithm can produce models when the target field is a set field with two or more possible values. Logistic regression methods built PASW Modeler are stepwise, enter, forwards and backwards. The stepwise method can be used in multinomial LR, the other methods can be used in both binominal LR and multinomial LR. Also two types of hypotheses are developed. First, a null hypothesis is developed, in which the effect of the independent variable on the

dependent variable is stated 0. Then, as alternative hypothesis, it is argued that the effect of the independent variable and the dependent variable is not 0. Specifically, in our case we will determine the size of the interest and we set independent variables that will be in the regression, we will test the effect of each independent variable on the model. Due to the fact that we will build a predicted model we need to evaluate its performance, to achieve that we need to create a test and training set in a part of the data and then we will check the efficiency of our model from the remaining sample.

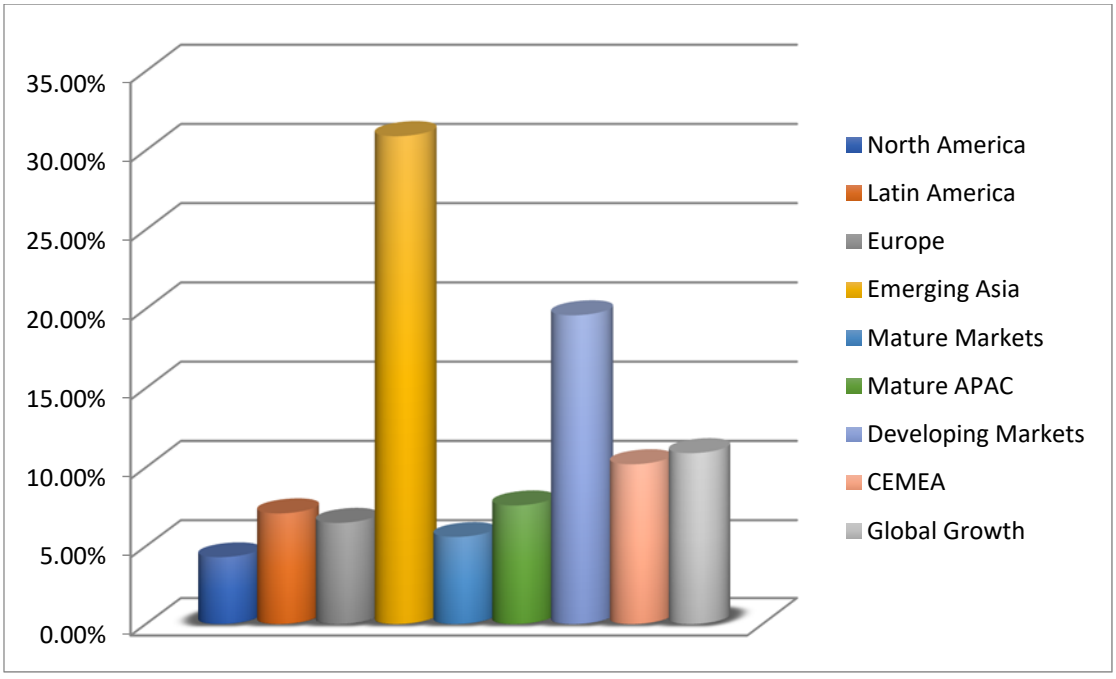
Finally, one of the things that we must consider when we create the predictive model is to determine the best model. A model is composed from all the variable options and manipulations applied to the data in addition to the selected algorithm and the relevant parameter. Simply, by trying to evaluate the model is the attempt to find the best model for implementation according to the current data.

Chapter 5

Big Data Analytics in Fraud Detection

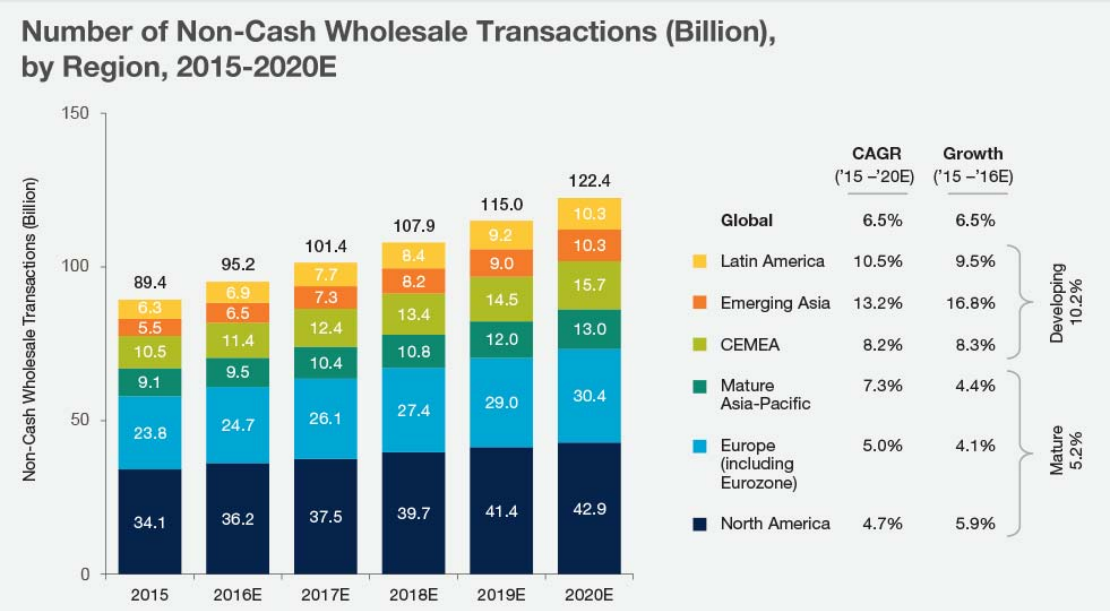
Due to the rise and rapid growth of E-Commerce, use of credit cards for online purchases has dramatically increased; actually global non-cash transactions broke a decade record for growth in 2014-2015 with volumes exceeding 11% to reach more than 433 billion in 2017, it is estimated that non-cash transaction volumes will record a CAGR of 10,9% during 2015-2020. As for Europe is expected that it will exceeding the growth of 6,5% over the next five years [World Payments Report, 2017].

Below are listed the growth rates of non-cash transactions across the world in 2017 table and also the number of non-cash transactions in billion by region between the years 2015 to 2020 figure :



Source: World Payments Report (2017)

Figure 9. Growth rates of Non-cash transactions 2015-2020



Source: *World Payments Report (2017)*

Figure 10. Number of non-cash wholesale transactions

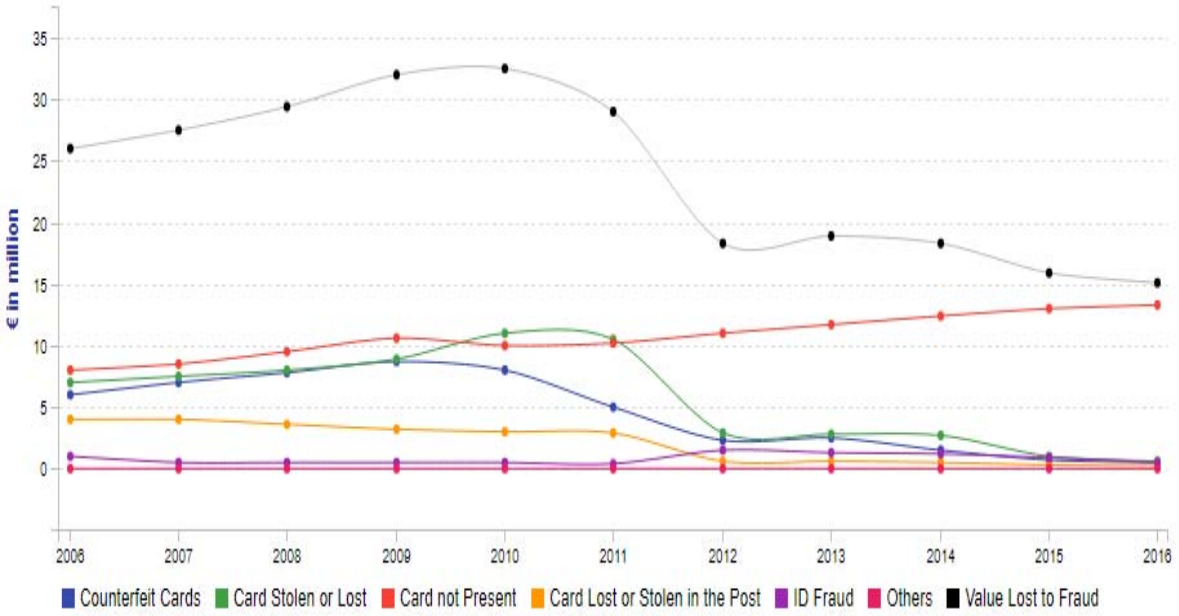
As we observe from the above statistical data non-cash transactions have become the most popular mode of payment for both online as well as regular purchase, as a consequence cases of fraud associated with it are also rising. Along with the great increase in non-cash transactions, specifically in credit card transactions, credit card fraud has become increasingly rampant in recent years. In Modern day the fraud is one of the major causes of great financial losses, not only for the individual clients but also for merchants in whom the affect of fraud has negative consequences in the reputation and image of merchant due to the fact that if a customer is victim of fraud with a certain company he no longer trusts the company and chooses another competitor and also may harm the business's reputation by urging other customers to avoid this business. In addition fraudulent cases affect mostly the banks in which it costs millions.

However there is no only one way that credit card fraud occurs, it appears in many ways [V.Dheepa, Dr.R.Dhanapal, (2009], some of them are:

- *Counterfeit Cards* is a method of criminally copying the magnetic stripe on a legitimate credit or debit card through a small handheld device called skimmer.
- *Card stolen or lost* it is the most common method of fraud it is easy to detect due to unexpected usage pattern of the credit card with respect to common practice.

- *Card not present* it is practice the fraudster steal the data of the card and not the card itself, it is a fraud that demands a prompt detection.
- *Card lost or stolen in the post* this kind of fraud occurs more rarely since the card should be activated first in condition to be used.
- ID fraud it is a practice where fraudsters use fake information to create a new individual in purpose of opening up a new credit card with the goal of creating credit records and boosting the credit profile.

Figure 11 shows us the losses in every type of credit fraudulent in Greece between the years 2006 to 2016.



Source: World Payments Report (2017)

Figure 11. Credit card Fraud Types

However frauds are continuously evolving, different fraud types and new criminals are taking part in the game and trying new strategies, so it is important to update the detection tools. The actions that should be taken against fraud can be divided into two stages, the ***fraud prevention*** and the ***fraud detection***. The first stage attempts to stop fraudulent transactions at source, on the other hand the second one attempt to identify if a new authorized transaction belongs to the class of fraudulent and also must be cost-effective in a way that the cost invested in transaction screening should not be higher than the loss due to frauds. It is obvious that it is preferable to prevent fraud but even though it cannot be prevented it is vital to detect it as soon as possible.

Although, fraud detection systems face difficulties and challenges enumerated bellow [V.Dheepa, Dr.R.Dhanapal, 2009].

- *Imbalanced data*- When the percentages of fraudulent credit card transaction are very small. (In the dataset of credit card transactions in the case study below), in this case there are several ways to tackle them
- *Different misclassification importance*- Misjudgment in the classification of transaction as fraudulent or on the opposite as normal.
- *Overlapping data*- The misclassification importance contribute in false positive rate and false negative rate, which is a key challenge of fraud detection system.
- *Lack of adaptability*- The inefficiency of detecting new patterns of normal and fraud behaviors, respectively.
- *Fraud detection cost* – It is vital for the detection system to identify the cost of detecting fraudulent in accordance with the cost of preventing fraudulent, because there is no gain in preventing a fraudulent transaction of a few amount of money.

Despite of the detection difficulties there are software tools that contribute in tackling them; one of these is the R language program. Below we will analyze the contribution of R language in the analysis process of large volume data and we will list the steps of statistical model development with R.

Handling big data and analyze them especially in case of developing a statistical model is one of the topics of high performance computing. Programming big data in R enables high-level distributed data and provides an easy visualization as it is an open source statistical software with a wide range of packages.

Chapter 6

Case Study Analysis

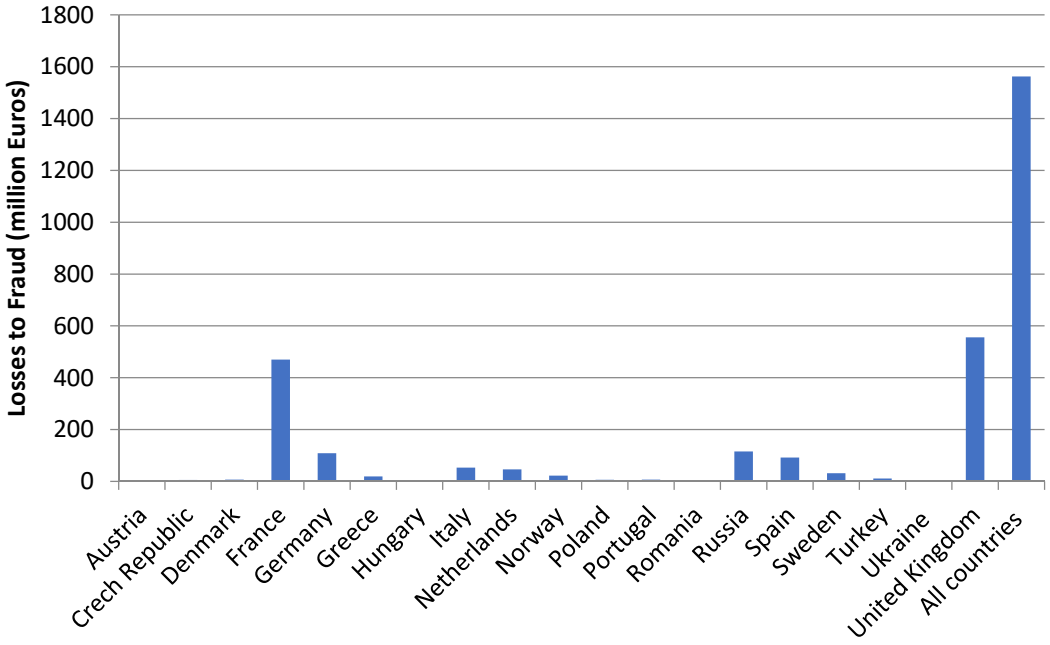
6.1 Fraud Detection Powered by Big Data

Predicting the fraudulent transactions based on the given data is done using programming software like R. By Analyzing and Visualizing the Credit card frauds, banks can implement a way to stop the fraud. Big data technology is a tool that can use banks to measure and manage risk at an individual customer level as well as at a product of portfolio level. Especially in combination with predictive analytics can help in predicting fraud because only when you have collected a large sample of outliers you can think about how to predict them.

In this chapter we will develop a methodology to analyze a dataset with transactions made by credit cards. This dataset is consisted of 284.807 transactions that occurred in two days in September 2013 by European cardholders. The data are public available and were taken from the website //www.kaggle.com/. The dataset was used by many analysts and data scientists before, one of this is ''Andrea Dal Pozzolo, 2015''.

But we will not count as much on the data as we will try to show the benefits that can be generated in managing risk and in risk decision-making by analyzing big data using open source tools such as R programming language. We will try to show how easy and quickly we can analyze and visualize data in cooperative with traditional statistical methods.

To begin with it is vital to first present the fraud losses from 19 European countries in the year of 2013, which is also the year from which we use the dataset of fraudulent transactions. The fraud losses are estimated totaling almost 1600M annually. As we understand the need to combat fraud is imperative.



Source: FICO Fair Isaac Corporation (2017)

Figure 12. Fraud losses in Europe

6.2 Analysis

To begin with credit card fraud belongs to a classification model whose include Logistics regression and hypothesis testing.

Dataset

Our dataset contains only numerical input variables, which are not the original features but are results of Principal Component Analysis (PCA) transformation due to confidentially issues. The features are separated in 28 variables V1, V2, V3… V28 that have been transformed with PCA and 3 variables Time, Amount and Class that are the original ones and have not been transformed with PCA. PCA transformation converts a set of observations of possibly correlated variable into a set of values of uncorrelated variables called although the solution will not be affected by this transformation.

Overall description of the method

Below we show a way to forecast the fraudulent card transactions applying Logistic Regression in R programming language, we will explore the ways which R can be used to calculate huge amount of data and to develop models that can be used from Risk Management Officers in order to predict risks related to fraud.

This will be achieved by running algorithms with the purpose of creating models and distinguish the most appropriate model for predicting the fraudulent. After creating the first model we will test it and validate it by using visualization, accuracy rate which describe the usefulness of the model, descriptive statistical methods as well as logistic regression. If our model it does not meet our criteria and we cannot accept it we will develop the next one. We will first implement the undersampling method to create a new predictive regression model. If it meets our criteria we will check its predictive power by estimating the ROC curve. However, even if the results are the preferable, we cannot rely on that, we have to check also the precision and recall due to the fact that poorly fitting model may have good discrimination.

Model 1- Logistic Regression

Logistic regression is a method with which we can analyze a dataset, the dataset contains one or more independent variable that determine an outcome , the regression analysis shows us the relationship between a dependent and independent variable [Jack Johnston and Dinardo, 1997]. It is usually used for forecasting. In our dataset the dependent variable is ''Class'' and all the other variables are independent.

$f(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$

where $z = \beta o + \beta 1X1 + \beta 1X2 + \cdots \beta \kappa X \kappa$

The depended variable of the model is:

Z= 1 if the transaction id fraudulent; Z= 0 if the transaction is honest

The independent variables of the model are:

X1= Time

X2= V1

X3= V2

.

.

.

X30=Amount

It is obvious that the variable which is depended is ''Class'' as we have to find a model that will adequately explain it, the other variables are the independent ''Time'', V1, V2…..V28 and Amount.

We outline some of our data in Table 3, we could not presented all due to their large volume. They are in total 284807 data. However all the dataset is public available to ''Kaggle'' website.

Table 3. Credit card transactions

Total	1	2	3	4	. . .	284806	284807
Time	0	0	1	1	. . .	172788	172792
V1	-135.980.713	119.185.711	-135.835.406	-0.96627171	. . .	-0.24044005	-0.53341252
V2	-0.07278117	0.26615071	-134.016.307	-0.18522601	. . .	0.530482513	-0.189733337
V3	253.634.674	0.16648011	177.320.934	179.299.334	. . .	0.702510230	0.703337367
V4	137.815.522	0.44815408	0.37977959	-0.86329128	. . .	0.68979917	-0.50627124
V5	-0.338320770	0.060017649	-0.503198133	-0.010308880	. . .	-0.377961134	-0.012545679
V6	0.46238778	-0.08236081	180.049.938	124.720.317	. . .	0.623707722	-0.649616686
V7	0.239598554	-0.078802983	0.791460956	0.237608940	. . .	-0.686179986	1.577.006.254
V8	0.098697901	0.085101655	0.247675787	0.377435875	. . .	0.679145460	-0.414650408
V9	0.36378697	-0.25542513	-151.465.432	-138.702.406	. . .	0.392086712	0.486179505
V10	0.090794172	-0.166974414	0.207642865	-0.054951922	. . .	-0.39912565	-0.91542665
V11	-0.551599533	1.612.726.661	0.624501459	-0.226487264	. . .	-1.933.848.815	-1.040.458.335
V12	-0.617800856	1.065.235.311	0.066083685	0.178228226	. . .	-0.96288614	-0.03151305
V13	-0.99138985	0.48909502	0.71729273	0.50775687	. . .	-104.208.166	-0.18809290
V14	-0.311169354	-0.143772296	-0.165945923	-0.287923745	. . .	0.4496244432	-0.0843164698
V15	1.468.176.972	0.635558093	2.345.864.949	-0.631418118	. . .	1.962.563.121	0.041333455
V16	-0.470400525	0.463917041	-2.890.083.194	-1.059.647.245	. . .	-0.608577127	-0.302620086
V17	0.207971242	-0.114804663	1.109.969.379	-0.684092786	. . .	0.509928460	-0.660376645
V18	0.025790580	-0.183361270	-0.121359313	1.965.775.003	. . .	1.113.980.590	0.167429934
V19	0.403992960	-0.145783041	-2.261.857.095	-1.232.621.970	. . .	2.897.848.773	-0.256116871
V20	0.251412098	-0.069083135	0.524979725	-0.208037781	. . .	0.1274335158	0.3829481049
V21	-0.018306778	-0.225775248	0.247998153	-0.108300452	. . .	0.265244916	0.261057331
V22	0.2778375756	-0.6386719528	0.7716794019	0.0052735968	. . .	0.80004874	0.64307844
V23	-0.110473910	0.101288021	0.909412262	-0.190320519	. . .	-0.163297944	0.376777014
V24	0.0669280749	-0.3398464755	-0.6892809565	-11.755.753.319	. . .	0.123205244	0.008797379
V25	0.1285393583	0.1671704044	-0.3276418337	0.6473760346	. . .	-0.56915886	-0.47364870
V26	-0.18911484	0.12589453	-0.13909657	-0.22192884	. . .	0.54666846	-0.81826712
V27	0.1335583767	-0.0089830991	-0.0553527940	0.0627228487	. . .	0.1088207347	-0.0024153088
V28	-	0.0147241692	-0.0597518406	0.0614576285	. . .	0.1045328215	0.0136489143

	0.0210530535						
Amount	149.62	2.69	378.66	123.50	. . .	10.00	217.00
Class	0	0	0	0	. . .	0	0

Source: Website Kaggle

Specifically, in the table above the following data are presented:

1st row: Total number of transactions

2nd row: Time~ The seconds elapsed between each transaction

29th row: Amount~ The transaction amount

31st row: Class~ The response variable which takes value 1 in case of fraud and 0 otherwise

3rd to 28th row: V1, V2, V3…V28 ~ We consider that they symbolize transaction id, account id, type of transaction, operation, balance, bank, account, merchant id etc.

We will first start with descriptive statistic in order to represent nearly every dataset and understand how well our variables behave, as well as to show the first potential flaws in our analysis. First, we will calculate mean, median, 1st Qu and 3rd Qu.

$$\text{Mean} = \sum x \frac{f(x)}{n} ,$$

$$1^{st}\ \text{Qu} = \frac{n+1}{4} ,$$

$$\text{Median} = \frac{n+1}{2} ,$$

$$3^{rd}\ \text{Qu} = 3\frac{n+1}{4} ,$$

$$3^{rd}\ \text{Qu} = 3\frac{n+1}{4}$$

The results are presented in Figure 13.

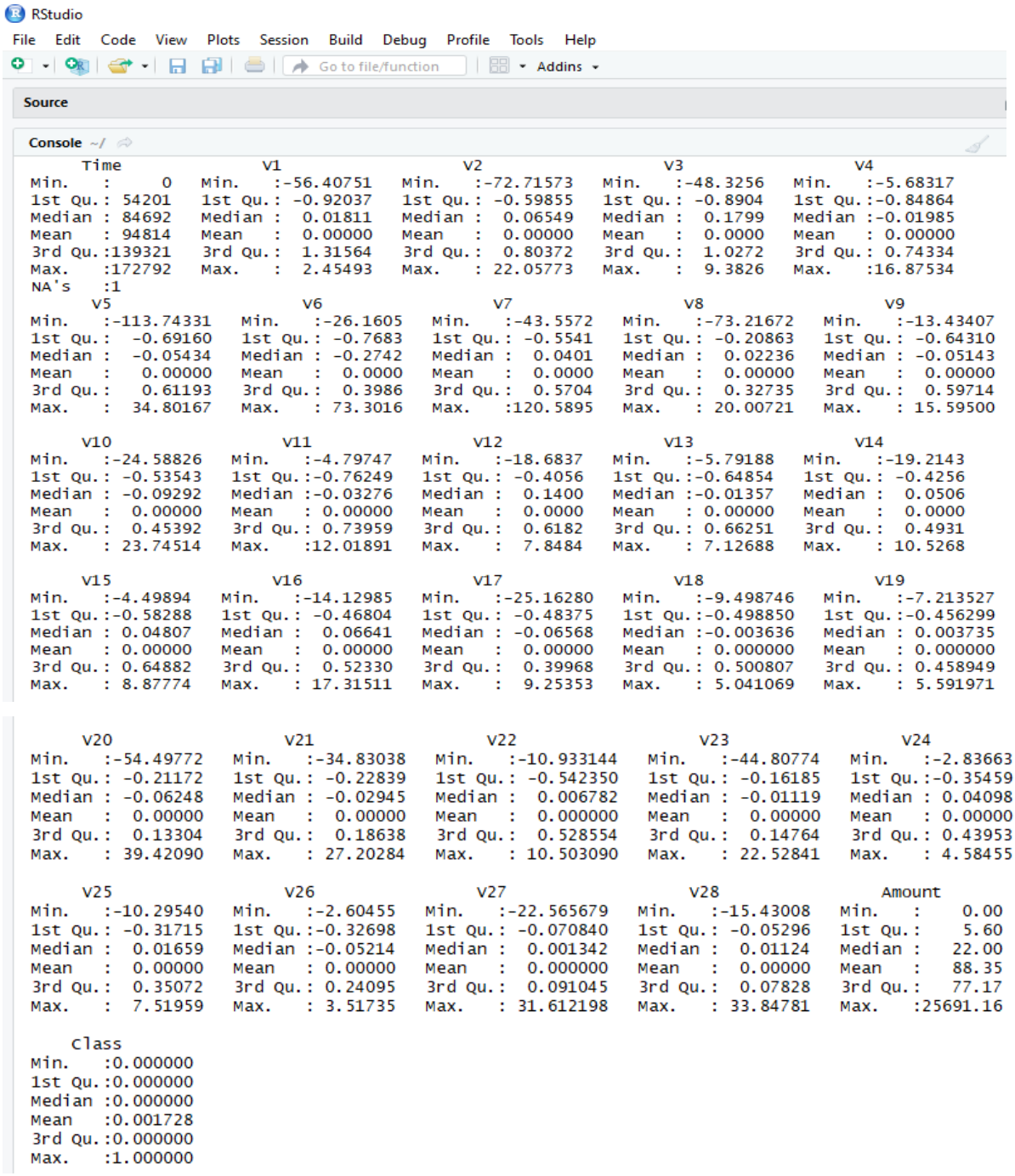


Figure 13. Descriptive statistic from R studio

As we observe the most data are not distributed normally, the only data that are normally distributed are V11, V13, V15, V18, V19, V22, V26. To prove it we will look at the corresponding histogram.

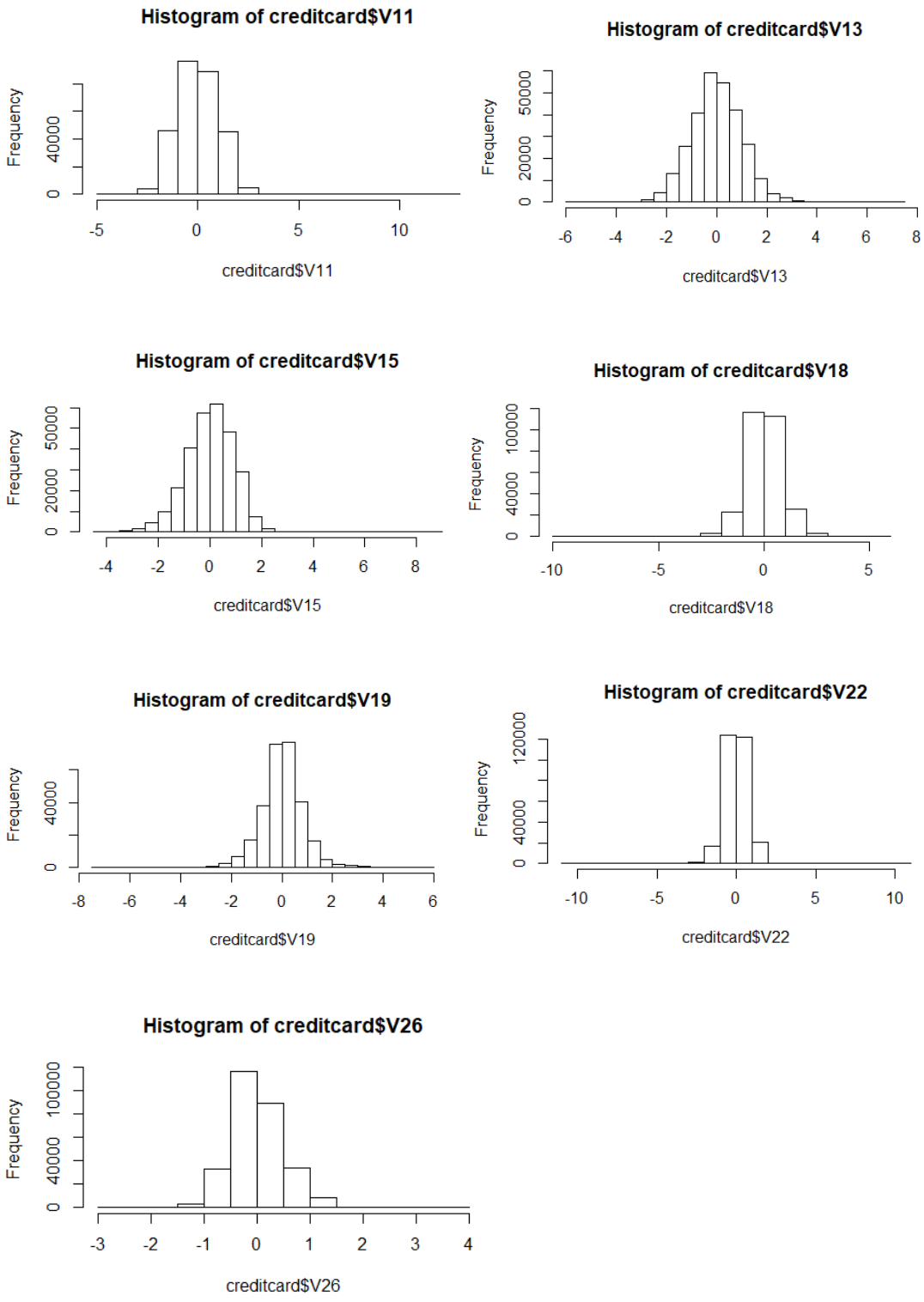


Figure 14. Histograms with normally distributed variables

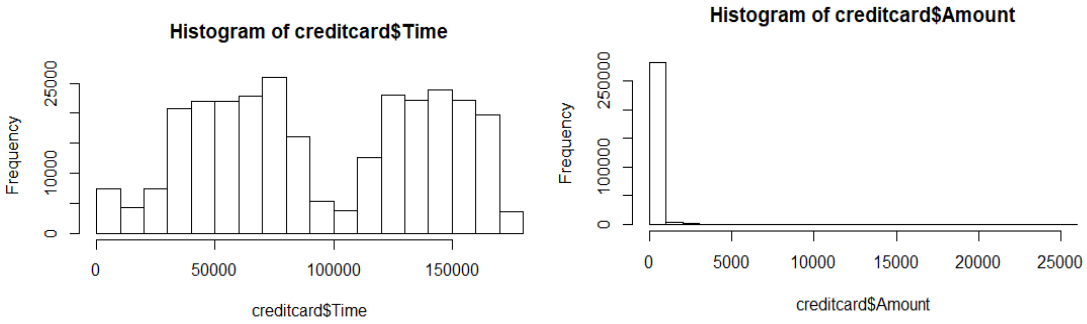


Figure 15. Histograms that are not normally distributed

A simple way to get a first idea of whether and how the know variables are correlated is to construct a Scatter Plot Diagram between the variables. The diagrams are presented below:

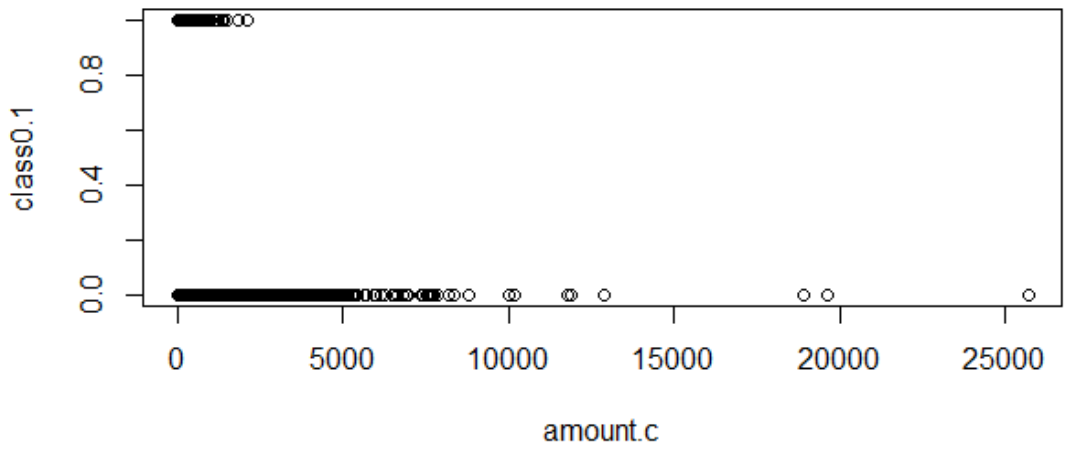


Figure 16. Scatter Plot between ''Class'' and ''Amount''

We observe that fraudulent transactions have been made in lower amount of money that the honest transactions, which means that those transactions may have become intact.

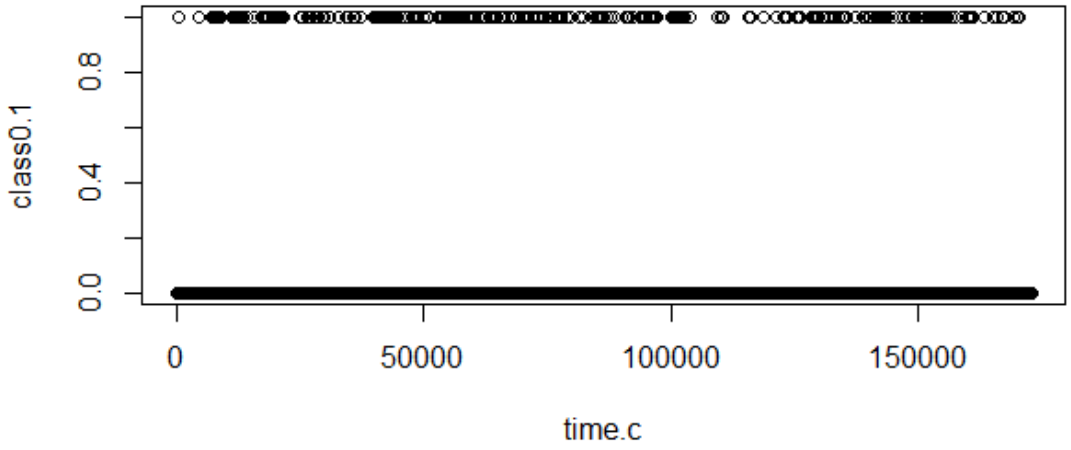


Figure 17. Scatter Plot between ''Class'' and ''Time'

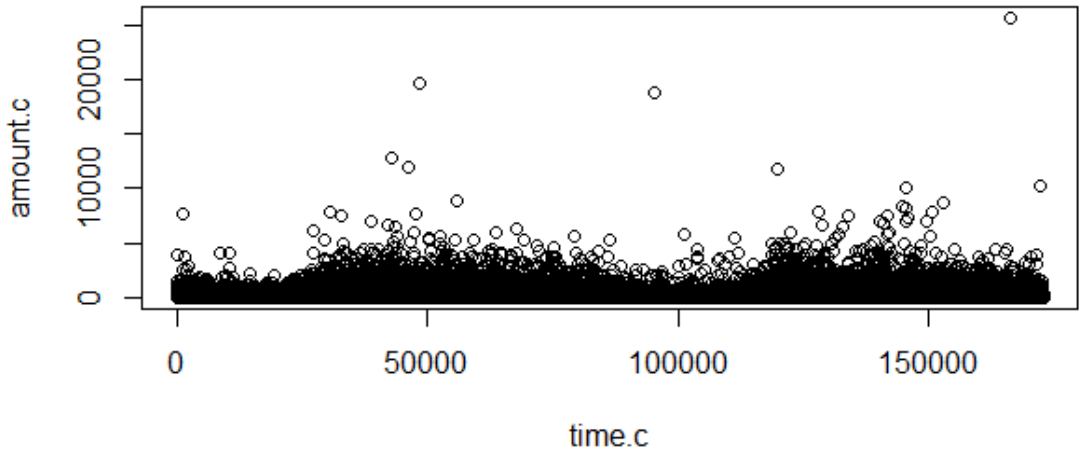


Figure 18. Scatter Plot between ''Amount'' and ''Time''

Observing figures 17 and 18 above we see that the variable "Time" does not really matter, due to the fact that it does not affect the outcome of the variable "Class" so we can remove it from our model.

Next we can see the counts of transactions divided into "fraudulent" and "honest" ones:

Table 4. Distinguish of transactions into fraud and honest

Honest	Fraudulent
0	1

284315	492

Base line accuracy = $\frac{284315}{284315+492}$ = 0.9982751

As we see base line accuracy rate is 99,9%, as a conclusion dataset is highly imbalanced, we cannot use this model because the model did not find any case of fraudulent transactions. So we need to perform a method to deal with imbalanced data. There are two options the first is oversampling which mean that we generate synthetic positive examples, and the second one is undersampling which means that we remove negative examples. But first we will present the regression model of all our dataset.

The following figure contains the result of the logistic regression using R.

```
glm(formula = Class ~ ., family = binomial, data = creditcard)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8738  -0.0293  -0.0193  -0.0123   4.6098

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.392e+00  2.494e-01 -33.652  < 2e-16 ***
Time        -3.742e-06  2.256e-06  -1.659 0.097095 .
V1           9.599e-02  4.239e-02   2.264 0.023556 *
V2           9.372e-03  5.806e-02   0.161 0.871769
V3          -7.926e-03  5.308e-02  -0.149 0.881303
V4           6.986e-01  7.390e-02   9.454  < 2e-16 ***
V5           1.295e-01  6.663e-02   1.944 0.051873 .
V6          -1.198e-01  7.365e-02  -1.626 0.103926
V7          -9.691e-02  6.669e-02  -1.453 0.146169
V8          -1.739e-01  3.045e-02  -5.711 1.13e-08 ***
V9          -2.843e-01  1.110e-01  -2.561 0.010441 *
V10         -8.176e-01  9.696e-02  -8.432  < 2e-16 ***
V11         -6.208e-02  8.147e-02  -0.762 0.446085
V12          9.089e-02  8.702e-02   1.044 0.296255
V13         -3.312e-01  8.161e-02  -4.058 4.95e-05 ***
V14         -5.571e-01  6.226e-02  -8.949  < 2e-16 ***
V15         -1.141e-01  8.578e-02  -1.330 0.183416
V16         -1.908e-01  1.250e-01  -1.526 0.126996
V17         -2.163e-02  7.004e-02  -0.309 0.757467
V18         -1.312e-02  1.290e-01  -0.102 0.918989
V19          9.625e-02  9.696e-02   0.993 0.320875
V20         -4.581e-01  8.170e-02  -5.607 2.05e-08 ***
V21          3.898e-01  6.002e-02   6.494 8.37e-11 ***
V22          6.297e-01  1.338e-01   4.707 2.52e-06 ***
V23         -9.506e-02  5.837e-02  -1.629 0.103404
V24          1.289e-01  1.474e-01   0.874 0.381889
V25         -7.610e-02  1.307e-01  -0.582 0.560259
V26          1.954e-02  1.898e-01   0.103 0.917995
V27         -8.188e-01  1.225e-01  -6.686 2.29e-11 ***
V28         -2.937e-01  8.816e-02  -3.332 0.000862 ***
Amount       9.159e-04  3.740e-04   2.449 0.014339 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7242.5  on 284805  degrees of freedom
Residual deviance: 2229.5  on 284775  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 2291.5

Number of Fisher Scoring iterations: 12
```

Figure 19. Coefficients Regression in R studio

The figure above shows us the coefficients regression and the p-values that correspond to each estimated coefficient, and the statistical significance of each p-value. Actually the one marked with ‘’***’’ is statistically significant only with space confidence level that we have define, one the other hand the other with one ‘’*’’ it is statistically significant only with space confidence level of 95% and above. So we can analyze the fitting and interpret what the model is telling us, we can see that V4, V8, V10, V13, V14, V20, V21, V22, V27, 28 are statistically significant with space confidence level that we have define, on the other hand V1, V9 and ‘’Amount’’ are statistically significant only with space level of confidence of 95% and above.

Also we observe that the level of AIC is equal to 2291.5 which is a high rate so we cannot accept the model as its predictability is unsatisfactory.

Akike’s information criterion (AIC) is an estimator of quality of a set of statistical models to each other, will choose the best model from the set.

We can calculate it from the formula:

$$AIC= -2(\log - likelihood) +2K$$

Where, K is the number of model parameters

The model with the lowest AIC value is considered as the best.

Due to the fact that our dataset is imbalanced and the level of AIC= 2291.5 is high we need to identify if the prediction equation gives us correct estimates forecasts, so we have to predict fraudulent with our model and then compare it using new data. With this method we will determine if our model is wrong or correct and accurately predicts fraudulent. In order to do this we will first implement the undersampling method to create a new predictive regression model.

Model 2 – Logistic Regression

First, we select all rows that have the value 0 in variable ''Class'' and we will present them in the table below; actually we will present the first part of the table due to huge amount of variables. The first part of the table from the variables that their class is equal to zero is in table 13.

Table 5. Head of honest data

```
# A tibble: 6 x 30
     V1      V2    V3     V4      V5      V6      V7      V8      V9     V10     V11     V12     V13     V14
  <dbl>   <dbl> <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 -1.36  -0.0728 2.54   1.38  -0.338   0.462   0.240   0.0987  0.364   0.0908 -0.552  -0.618  -0.991  -0.311
2  1.19   0.266  0.166  0.448  0.0600 -0.0824 -0.0788  0.0851 -0.255  -0.167   1.61    1.07    0.489  -0.144
3 -1.36  -1.34   1.77   0.380 -0.503   1.80    0.791   0.248  -1.51    0.208   0.625   0.0661  0.717  -0.166
4 -0.966 -0.185  1.79  -0.863 -0.0103  1.25    0.238   0.377  -1.39   -0.0550 -0.226   0.178   0.508  -0.288
5 -1.16   0.878  1.55   0.403 -0.407   0.0959  0.593  -0.271   0.818   0.753  -0.823   0.538   1.35   -1.12
6 -0.426  0.961  1.14  -0.168  0.421  -0.0297  0.476   0.260  -0.569  -0.371   1.34    0.360  -0.358  -0.137
# ... with 16 more variables: V15 <dbl>, V16 <dbl>, V17 <dbl>, V18 <dbl>, V19 <dbl>, V20 <dbl>, V21 <dbl>,
#   V22 <dbl>, V23 <dbl>, V24 <dbl>, V25 <dbl>, V26 <dbl>, V27 <dbl>, V28 <dbl>, Amount <dbl>, Class <fct>
```

On the other hand the first part of the table from the variables that their class is equal to 1 are presented in the table 14.

Table 6. Head of fraudulent data

```
# A tibble: 6 x 30
     V1    V2     V3    V4     V5     V6     V7      V8     V9     V10    V11     V12     V13     V14
  <dbl> <dbl>  <dbl> <dbl>  <dbl>  <dbl>  <dbl>   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
1 -2.31  1.95 -1.61   4.00 -0.522 -1.43  -2.54   1.39   -2.77  -2.77   3.20   -2.90  -0.595  -4.29
2 -3.04 -3.16  1.09   2.29  1.36  -1.06   0.326 -0.0678 -0.271 -0.839 -0.415  -0.503  0.677  -1.69
3 -2.30  1.76 -0.360  2.33 -0.822 -0.0758 0.562 -0.399  -0.238 -1.53   2.03   -6.56   0.0229 -1.47
4 -4.40  1.36 -2.59   2.68 -1.13  -1.71  -3.50  -0.249  -0.248 -4.80   4.90  -10.9    0.184  -6.77
5  1.23  3.02 -4.30   4.73  3.62  -1.36   1.71  -0.496  -1.28  -2.45   2.10   -4.61   1.46   -6.08
6  0.0084 4.14 -6.24   6.68  0.768 -3.35  -1.63   0.155  -2.80  -6.19   5.66   -9.85  -0.306 -10.7
# ... with 16 more variables: V15 <dbl>, V16 <dbl>, V17 <dbl>, V18 <dbl>, V19 <dbl>, V20 <dbl>, V21 <dbl>,
#   V22 <dbl>, V23 <dbl>, V24 <dbl>, V25 <dbl>, V26 <dbl>, V27 <dbl>, V28 <dbl>, Amount <dbl>, Class <fct>
```

Subsequently, we will take a random sample of 10000 data from our ''creditcard'' dataset which their class is equal to 0, and also we will take all the data that their class is equal to 1, then we will combine this two data frames by rows and we will create a new sample that consists of 10000 honest and 492 fraudulent. The rows of our data sample are equal to 10492.

We will split the data set for training and testing the model by dividing the data set into two pieces, the training set 70% and the test 30%. The results of each data sample are presented below in Table 7 and 8.

Table 7. Fraud and honest transactions from training model

Honest	Fraudulent
0	1
7000	344

Table 8. Fraud and honest transactions form testing model

Honest	Fraudulent
0	1
3000	148

From the tables above we observe that our data is well balanced compared to the previous one. In order to test our model we will use the training set to fit our model which will be testing over the testing test.

We fit the model and we obtain the results below.

Call

```
glm(formula = Class ~ ., family = "binomial", data = train, control = list(maxit = 50))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6029  -0.0653  -0.0267  -0.0115   4.3248

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.1052402  0.4656651 -10.963  < 2e-16 ***
V1           0.3939362  0.1530751   2.573 0.010068 *
V2          -0.5293970  0.4064274  -1.303 0.192724
V3          -0.6722233  0.1933149  -3.477 0.000506 ***
V4           0.9899609  0.1839637   5.381 7.40e-08 ***
V5          -0.0675556  0.2375127  -0.284 0.776081
V6          -0.7516068  0.2617502  -2.871 0.004086 **
V7           0.4338926  0.3084682   1.407 0.159545
V8          -0.4904498  0.1381988  -3.549 0.000387 ***
V9          -0.6703248  0.2546586  -2.632 0.008482 **
V10         -0.9098721  0.3062631  -2.971 0.002969 **
V11         -0.3535053  0.1515632  -2.332 0.019680 *
V12         -0.1836310  0.1665059  -1.103 0.270092
V13         -0.8147529  0.1511795  -5.389 7.07e-08 ***
V14         -0.7516864  0.1608895  -4.672 2.98e-06 ***
V15         -0.6915734  0.2028090  -3.410 0.000650 ***
V16         -0.2118794  0.2702639  -0.784 0.433056
V17         -0.4881918  0.1790762  -2.726 0.006407 **
V18         -0.0833591  0.2710508  -0.308 0.758432
V19         -0.0701674  0.2268406  -0.309 0.757075
V20         -0.4430484  0.3681361  -1.203 0.228786
V21          0.6283065  0.2374725   2.646 0.008150 **
V22          1.2462421  0.3723477   3.347 0.000817 ***
V23         -0.4327163  0.2864649  -1.511 0.130906
V24         -0.8148313  0.3804430  -2.142 0.032210 *
V25         -0.5522904  0.3699038  -1.493 0.135420
V26         -0.2513033  0.4252117  -0.591 0.554515
V27         -0.5514150  0.4802891  -1.148 0.250931
V28          0.0706456  0.5153505   0.137 0.890966
Amount      -0.0006308  0.0032918  -0.192 0.848040
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2777.60  on 7343  degrees of freedom
Residual deviance:  368.75  on 7314  degrees of freedom
AIC: 428.75

Number of Fisher Scoring iterations: 10
```

Figure 20. Coefficients Regression in R studio- model 2

As we see model 2 meets our criteria and the AIC= 428.75 rate is lower than the first model. To evaluate the predicted power of our model we plot the ROC curve and calculating the AUC.

The receiver operating characteristic (ROC) is a method of evaluating the efficiency of our model, the characteristics of the indicator's function ROC are calculated for all points and displayed graphically for interpretation, the axes of the chart are the sensitivity and the peculiarity calculated from the classification rates [Kelly H.Zou, Aiyi Liu, Andriy I. Bandos, Lucila Ohno-Machado, Howard E.Rockette, 2012].

The curves of ROC are calculated:

True positive rate is TPR =$\int_T^{\infty} f1(x)dx$

False Positive Rate is FPR = $\int_T^{\infty} fo(x)dx$

ROC curve plots parametrically TRT(T) versus FPR(T) with T as the varying parameter

The area under the curve:

A = $\int_{-\infty}^{\infty} TPR(T)FPR'(T)dT = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(T' > T)f1(T')fo(T)dT'dT = P(X1 > Xo$

Where X_1 is the score for a positive instance and Xo a score for negative instance and fo and f_1 are probability densities.

The ROC curve shows what a false positive rate should be expected according to true positive rate, the accuracy of the test is measured by the area under the ROC curve, specifically we could separate the area as:

50% - 70% =Acceptable

70% - 85% = Good

85% - 100% = Very Good

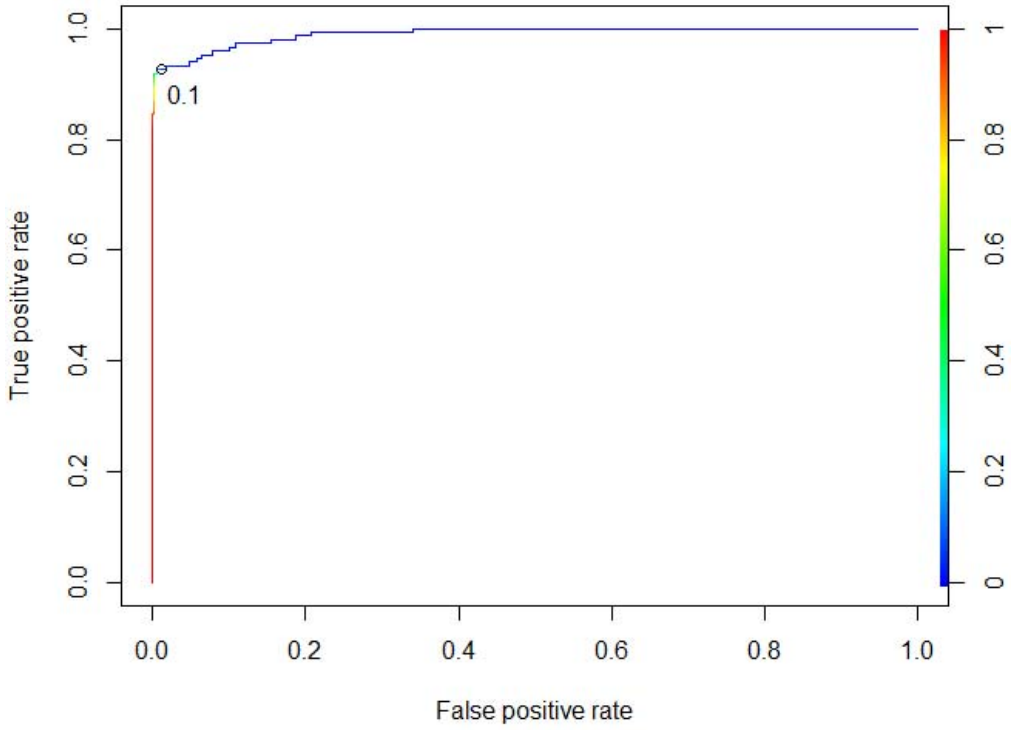


Figure 21. ROC graph in R studio

The area under the ROC curve it shows us the predictive power of the model, actually it shows which ranges provides a measure of the model's ability to discriminate between those observations that experience the outcome (Z=1) of interest and those that do not experience the outcome of interest (Z=0). In our case the result is satisfying as it equals with 0.98 which means that our model is a model with good predictive ability. It will help us understand the impact of a chosen classification threshold visually (Table 9).

We calculate the True Positive Rate and the False Positive Rate with the formulas:

TPR= TP/P =TP (TP+FN)

FPR= FP/N = FP/(FP+TN) = 1-SPC where SPC= TN/N = TN/(TN+FP)

Where,

TP (True positive) = the number of cases correctly identified as yes

FP (False positive) = the number of cases incorrectly identified as yes

TN (True negative) = the number of cases correctly identified as no

FN (False negative) = the number of cases incorrectly identified as no

Table 9. Confusion matrix for threshold > 0.5

	FALSE	TRUE
0	2994	6
1	12	136

However, the accuracy rate is still high so it is not impossible that poorly fitting model may have good discrimination, as a consequence we can use also another way to understand the meaning of the area under the ROC we will check also the precision and recall.

Accuracy is a measure of correctly predicted observation to the total observations

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ = 0.9942821

Precision is a measure of correctly predicted positive observations to the total predicted positive observations.

Precision = $\frac{TP}{TP+FP}$ = 0.9577465

Recall is a measure of correctly predicted positive observations to the all observations in actual class.

Recall = $\frac{TP}{TP+FN}$ = 0.9189189

F1 score is a score that takes into account false positives and false negatives rates as it is the average of precision and recall.

F1-Score = $2\frac{Precision* Recall}{Precision+Recall}$ = 0.937931

As we see we have high precision and recall, which means that high precision, relates to a low positive rate and also high recall relates to a low false negative rate, as a consequence model 2 is the model that we were looking for because the classifier is returning accurate results and a majority of all positive results.

This credit card fraud model is established on the basis of sufficient understanding the method of credit card fraud prediction. This modeling can be used by credit card risk control officers and his teams as a reference for evaluating the risk degree of certain transaction. When evaluating the risk degree of certain transaction can assess this transaction by taking the credit card fraud model as a reference to make reasonable

decisions. If a transaction is listed in the transaction being list provided by this credit card fraud model then the probability of this transaction being a fraud is high.

Although data analytics models do not remain the same with the passage of time they may be changes in the techniques, methods, and mode of fraud may also change, so continuous tracking is needed. Thus we also need to keep pace with technology because as we seen with the use of open tools like R language we can manage the unstrained. Specifically, R is a programming language that is easy to be understood, is extensible language and offers great functionality to users so they can easily and quickly create their own tools and analyze data.

All in all, we conclude that with powerful software, efficient hardware and big data analytics methods and techniques banks can take advantage of transactional data in real time to provide improved business insight and lower risk to their organization, to merchants and customers by more accurately and quickly indentifying potential fraudulent activity.

Chapter 7
Conclusion and Recommendations

7.1 Final Thoughts

It is a fact that we live in a world of continuously increasing data, the effective use of Big Data and the rise of intelligence that arise from new technologies is recognized by many organizations as key to gaining a competitive advantage and outperforming peers. Big data has motivated businesses to invest due to the fact that the same big data infrastructure used to mitigate risks can be used to pursue new sources of revenue. Also has embraced new technologies and techniques and has shown a fast and reliable path to risk management analysis, by applying such analytics to big data it is certain that valuable information can be extracted and exploited to enhance decision making. Although big data is not just about technology, it is about developing ways of working and collaborating as it promotes changes in culture and learning in organizations.

Organizations must understand the importance of big data associated with decision-making, they should emphasis on creating opportunities from these decisions as we live in a world that is always connected and everything changes so rapidly, big data analytics has the power to provide a basis for advancements on technological, scientific event in humanitarian levels.

All in all the main purpose of the thesis is to present the opportunities and prospects of the implementation of big data techniques to better and more effective decision-making and risk management. The case study that we apply about credit card fraud detection may contribute in understanding the importance of using big data technologies in decision-making on risk management as well as generally in business management.

References

Fraser, J., Simkins, B.J. (2010). ‘’*Enterprise Risk Managemen*t’’, The Robert W.Kob series in finance

Parag Kulkarni, Srang Joshi, Meta S.Brown, (2016). ‘’*Big Data Analytics*’’

Paola Cercheillo, Paolo Giudici, (2016).‘’*Big data analysis for financial risk management*’’

J.Davidson Frame, (2003). ‘’*Managing Risk in Organizations: A Guide for Managers*’’

Boris Delibasic, Jorge E.Hernandez, Jason Papathanasiou, Fatima Dargan, Pascale Zarate, Rita Ribeiro, Shaofeng Liu, Isabelle Linden, (2015). ‘’*Decision Support Systems V- Big Data Analytics for Decision Making*’’

Thiago Poleto, Victor Diogho Heuer de Carvalho and Ana Paula Cabral Seixas Costa, (2015). ‘’*The Roles of Big Data in the Decision-Support Process: An Empirical Investigation*’’

Saumyadipta Pyne. B.LS. Prakasa Rao S.B. Rao, (2016). ‘’*Big Data Analytics, Methods and Applications*’’

Jack Johnston and Dinardo, (1997). *‘’Econometric Methods*’’

Ramanathan, Muthu Mathirajan, A. Ravi Ravindran, (2015). ‘’*Big Data Analytics Using Multiple Criteria Decision-Making Models*’’

Hussein A.Abbass, (2015). ‘’*Computational Red Teaming, Risk analytics of Big Data to Decisions Intelligent Systems*’’

Sourab Mazumder, Robin Singh Bhadoria, Ganesh Chandra Deka, (2017). ‘’*Distributed Computing in Big Data Analytics*’’

Alexander, Rico Bergmann, Stephan, (2014). ‘’*The Stratosphere Platform for Big Data Analytics*’’

V.Dheepa, Dr.R.Dhanapal, (2009). '' *Analysis of Credit Card Fraud Detection Methods*''

Kelly H.Zou, Aiyi Liu, Andriy I. Bandos, Lucila Ohno-Machado, Howard E.Rockette, (2012), ''*Statistical Evaluation of Diagnostic Performance: Topis in ROC Analysis*''

EY. Insights on Governance, Risk and Compliance, (2014*). ''Big Data- Changes the way businesses compete and operate''.* [Online]. Available at: http://www.ey.com/Publication/vwLUAssets/EY_-_Big_data:_changing_the_way_businesses_operate/%24FILE/EY-Insights-on-GRC-Big-data.pdf [Accessed 10 Nov. 2017]

Oracle Enterprise Architecture White Paper, (2016). ''*An Enterprise Architect's Guide to BIG data*''. [Online]. Available at: http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf [Accessed 10 Nov. 2017]

DATAFLOQ, Mark van Rijneman, (2015) ''How *Big Data will Improve Decision Making in Your Organisation''* [Online]. Available at: https://datafloq.com/read/big-data-will-improve-decision-making-in-your-orga/307 [Accessed 20 April. 2018]

Paola Cerchillo and Paolo Giudici, (2016). ''*Big Data Analysis for Financial Risk Management*. [Online]. Available at: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0053-4 [Accessed 10 Nov. 2017]

The Economist Intelligence Unit, (2014). ''*Retail Banks and Big Data: Big Data as the key to better Risk Management''.* [Online]. Available at: https://www.eiuperspectives.economist.com/sites/default/files/RetailBanksandBigData.pdf [Accessed 18 Dec. 2017]

Debbie Stephenson, (2013). ''*7 Big Data Techniques That Create Business Value*''. [Online]. Available at: https://www.firmex.com/thedealroom/7-big-data-techniques-that-create-business-value/ [Accessed 5 Jan. 2018]

FICO, (2017). ''*Evolution of Credit Card Fraud in Europe 2016''* [Online]. Available at: http://www.fico.com/europeanfraud/index [Accessed 5 Jan. 2018]

World Payments Report (2017). [Online], Available at: https://www.worldpaymentsreport.com/#non-cash-payments-content [Accessed 5 Mars. 2018]

Comunidad de Madrid*, ''Risk Analysis and Quantification''*.[Online]. Available at: http://www.madrid.org/cs/StaticFiles/Emprendedores/Analisis_Riesgos/pages/pdf/metodologia/4AnalisisycuantificaciondelRiesgo%28AR%29_en.pdf [Accessed 5 Jan. 2018]

National Cyber Security Center, (2016). *''Risk Management And Risk Analysis in Practice''*. [Online]. Available at: https://www.ncsc.gov.uk/guidance/risk-management-and-risk-analysis-practice [Accessed 5 Jan. 2018]

Technavio, (2014). *''Small and Medium-Sized Enterprise (SME) Big Data Companies''*. [Online]. Available at: https://www.technavio.com/blog/top-12-small-and-medium-sized-enterprise-sme-big-data-companies [Accessed 3 Feb. 2018]

Hong Shu, (2016). *''Big Data Analytics: Six Techniques''.* [Online]. Available at: http://www.tandfonline.com/doi/full/10.1080/10095020.2016.1182307 [Accessed 3 Feb. 2018]

[Online]. Available at: https://www.kaggle.com/ [Accessed 10 Nov. 2017]

Donna-M.Fernandez, (2016). *''Comparing Hadoop, MapReduce, Spark, Flink and Storm''.* [Online]. Available at: http://www.metistream.com/comparing-hadoop-mapreduce-spark-flink-storm/ [Accessed 5 April. 2018]

Appendix A

Questioner about the use of open source tools

Appendix A1: The e-mail that was sent

In order to identify if the Greek companies are familiar with new technology software's a questioner was sending via e-mail in 47 companies.

''Dear Sir/Madam

I am a master's student in the department of Enterprise Risk Management at Open University of Cyprus. Due to the authorship of my master thesis I kindly ask to answer me in the question below:

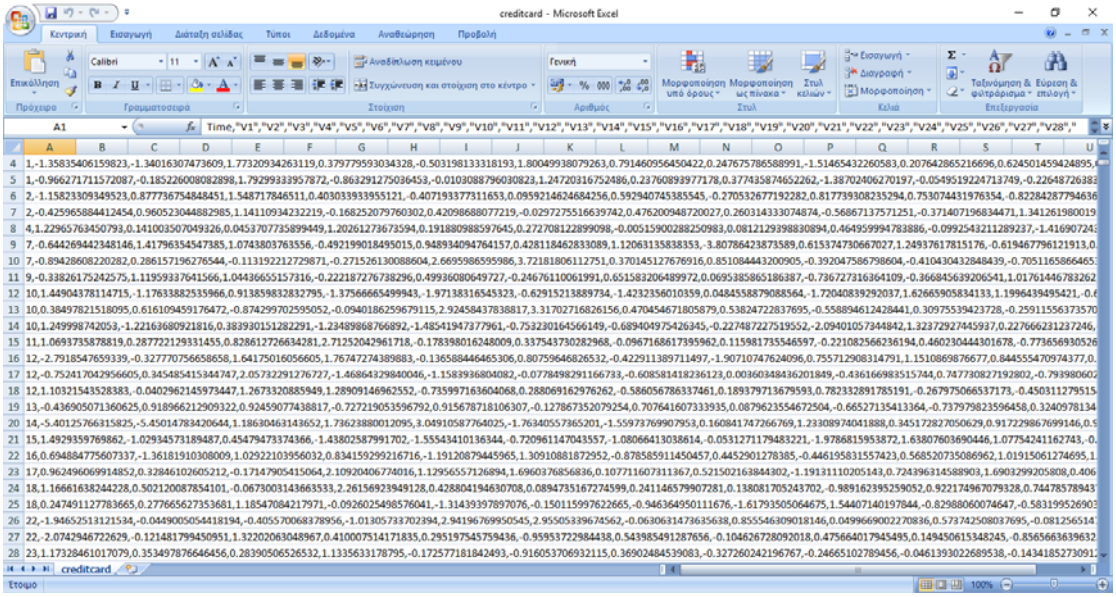
Which of the following software tools do you use in the analytic process?''

The results are presented in the table below:

Data Analytics Tools	Totals
Apache Hadoop	0
Apache Spark	0
Apache Flink	0
Apache Mahout	0
Microsoft Excel	32
Map Reduce	1
R	10
SQL	27
SAP	22
SPSS	20
SASA	17
Stata	5
Hadoop	2

Appendix B

Appendix B1. Dataset of transaction



Appendix B2. Logistic Regression Code in R language

First of all, to make the analysis in R some packages are required to be installed in our R studio, which is presented below:

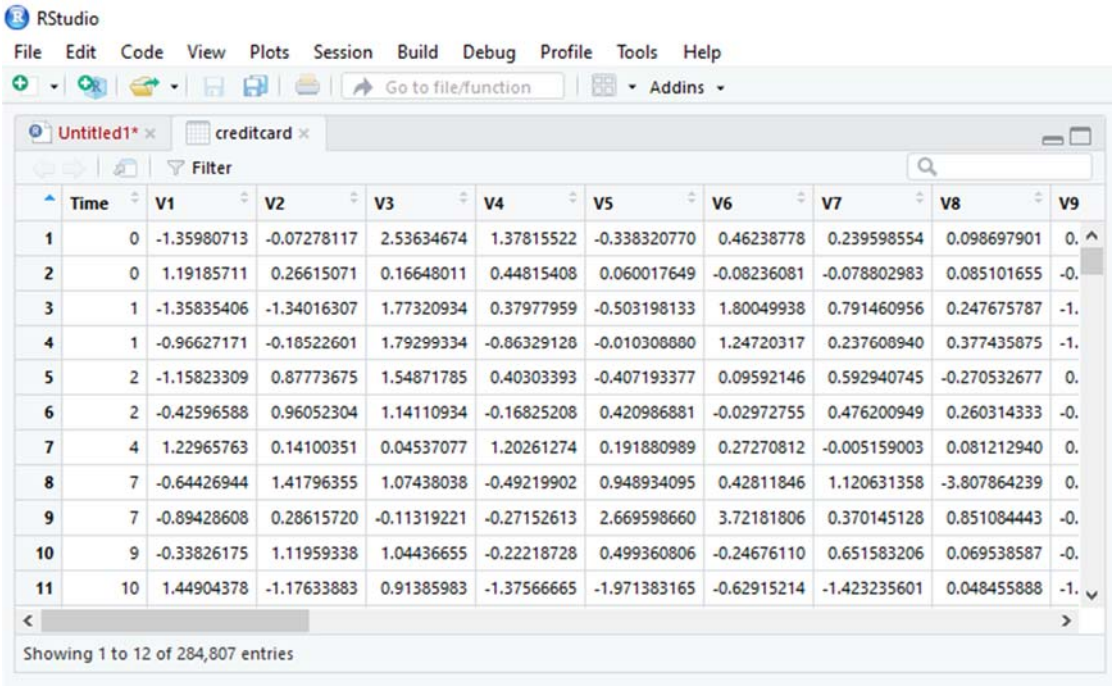
caTools, e1071, rpart, rpart.plot, reader, caret, ggplot2, glm.deploy, glm.predict, gplots, gtools, PredictiveRegression, ROCR, xtable

Then we will load into R the file that contains the dataset with the functions below:

```
> library(readr)
> creditcard <- read_csv("C:/Users/antigoni/Desktop/creditcard.csv")
```

The first line contains the names of the variables. View of the data set in R language

```
> View(creditcard)
```



	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9
1	0	-1.35980713	-0.07278117	2.53634674	1.37815522	-0.338320770	0.46238778	0.239598554	0.098697901	0.
2	0	1.19185711	0.26615071	0.16648011	0.44815408	0.060017649	-0.08236081	-0.078802983	0.085101655	-0.
3	1	-1.35835406	-1.34016307	1.77320934	0.37977959	-0.503198133	1.80049938	0.791460956	0.247675787	-1.
4	1	-0.96627171	-0.18522601	1.79299334	-0.86329128	-0.010308880	1.24720317	0.237608940	0.377435875	-1.
5	2	-1.15823309	0.87773675	1.54871785	0.40303393	-0.407193377	0.09592146	0.592940745	-0.270532677	0.
6	2	-0.42596588	0.96052304	1.14110934	-0.16825208	0.420986881	-0.02972755	0.476200949	0.260314333	-0.
7	4	1.22965763	0.14100351	0.04537077	1.20261274	0.191880989	0.27270812	-0.005159003	0.081212940	0.
8	7	-0.64426944	1.41796355	1.07438038	-0.49219902	0.948934095	0.42811846	1.120631358	-3.807864239	0.
9	7	-0.89428608	0.28615720	-0.11319221	-0.27152613	2.669598660	3.72181806	0.370145128	0.851084443	-0.
10	9	-0.33826175	1.11959338	1.04436655	-0.22218728	0.499360806	-0.24676110	0.651583206	0.069538587	-0.
11	10	1.44904378	-1.17633883	0.91385983	-1.37566665	-1.971383165	-0.62915214	-1.423235601	0.048455888	-1.

Showing 1 to 12 of 284,807 entries

```
>summary(creditcard)
> hist(creditcard$V11)
> hist(creditcard$V13)
> hist(creditcard$v15)
> hist(creditcard$V15)
> hist(creditcard$V18)
> hist(creditcard$V19)
> hist(creditcard$V22)
> hist(creditcard$V26)
> hist(creditcard$Amount)
> hist(creditcard$Class)
> table(creditcard$class)
> class0.1<-(creditcard$Class)
> time.c<-creditcard$Time
> amount.c<-creditcard$Amount
> plot(time.c, class0.1)
> amount.c<-creditcard$Amount
> plot(amount.c, class0.1)
> plot(amount.c, time.c)
> plot(time.c, amount.c)
```

```
>accuracy_rate<-284315/(284315+492)
> model_1 <- glm(Class ~ ., data = creditcard, family = binomial)
> View(model_1)
> summary(model_1)
> credicard$Class = as.factor(creditcard$Class)
> cclass.0 <- subset(creditcard, creditcard$Class == 0)
> head(cclass.0)
> cclass.1 <- subset(creditcard, creditcard$Class == 1)
> head(cclass.1)
> nrow(cclass.0)= 284315
> nrow(cclass.1)= 492
> cclass.0 <- cclass.0[1:10000, ]
> nrow(cclass.0)= 10000
> data <- rbind(cclass.0, cclass.1)
> nrow(data)= 10492
> set.seed(1)
> split <- sample.split(data$Class, SplitRatio = 0.7)
> train <- subset(data, split == TRUE)
> cv <- subset(data, split == FALSE)
> table(train$Class)
> table(cv$Class)
> model_2 <- glm(Class ~ ., data = train, family = "binomial", control = list(maxit = 50))
>summary(model_2)
> model_2_predict <- predict(model_2, cv, type = "response")
> summary(model_2_predict)
> ROCRpred= prediction(model_2_predict, cv$Class)
> ROCRpref = performance(ROCRpred, "tpr", "fpr")
> plot(ROCRpref, colorize =TRUE, print.cutoffs.at =seq(0.1,0.1), text.adj=c(-0.2, 1.7))
```