COMPREHENSIVE STUDY OF METHODS TECHNIQUES AND TOOLS USED FOR DATA ANALYTICS IN FIELDS OF FINANCE

by

Parth Sarangi

A special study report submitted in partial fulfillment of the requirements for the degree of Master of Engineering in Information Management

Examination Committee:	Dr. Vatcharaporn Esichaikul (Chairperson) Dr. Matthew N. Dailey Prof. Sumanta Guha
Nationality:	Indian
Previous Degree:	Bachelor of Technology in Electronics and Communication National Institute of Technology Srinagar, India
Scholarship Donor:	AIT Fellowship

Asian Institute of Technology
School of Engineering and Technology
Thailand
April 2017

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Data analysis is an very robust topic in the field of data science and encompasses the various mathematical functions. The functions are statistical in nature and are performed on the data obtained. The goal of data analytics is to support (or reject) the hypothesis which the data scientist postulates. " By processing a steady stream of real-time data, organizations can make time-sensitive decisions faster than ever before, monitor emerging trends, course-correct rapidly and jump on new business opportunities." [BIG Data Analytics: A Framework for Unstructured Data Analysis pdf]

This paper tries to enlist most of the up-to-date techniques used by researchers and mathematicians to make sense of the data. Also the paper presents them in three groups of analytics. But then there arises a question which is, why the need for data analytics ? Well, to answer that the paper proposes the literature from an article of Kdnuggets (Balm, 2015).

Data analytics is already the next big disrupter in the financial sector (Culp, 2017). New startups are build solutions to automate mundane manual tasks of reconciliations and consolidations. Artificial intelligence is helping companies to redesign the traditional process flows and restructure work-flows for optimization. With complex algorithms of machine learning, artificial intelligence, "Bot's" are designed which learn the product usage and popularity among customers. These "Bot's" can also interact autonomously to users and deduce patterns of interaction. Such an exisiting example is cited in Cnbc's article (Taylor, 2016), about Bank of America deploying "Erica" an digital assistant based on models of artificial intelligence.

According to Paul following are use cases of data analytics :

1. Analytics powers our decisions – we do not need to guess while making bold new decisions, we should use the information from data at hand.
2. Your data analysis weighs down your opponent's argument.
3. Cut down on loss making ventures with data analytics.
4. Can be applied to all domains be it health-care, banking, marketing, sales, operations etc.

On the scenario when describing Data analytics it is very important to put the focus on Hypothesis.

Shown in the table 1.1 is some business areas

Application area	**Applications**	**Specifics**
Insurance	Fraud detection	Identify claims meriting investigation
Telecom	Churn	Identify likely customer turnover
Telemarketing	On-line information	Aid telemarketers with easy data access
Human resource management	Churn	Identify potential employee turnover
Retail	Affinity positioning Cross-selling	Position product effectively Find more products for customers
Banking	Customer relationship management	Identify customer value Develop programs to maximize revenue
Credit card management	Lift Churn	Identify effective market segments Identify likely customer turnover

Table 1.1: Data mining use cases

Figure 1.1 is a view of the data analytics with respect to the main fields.

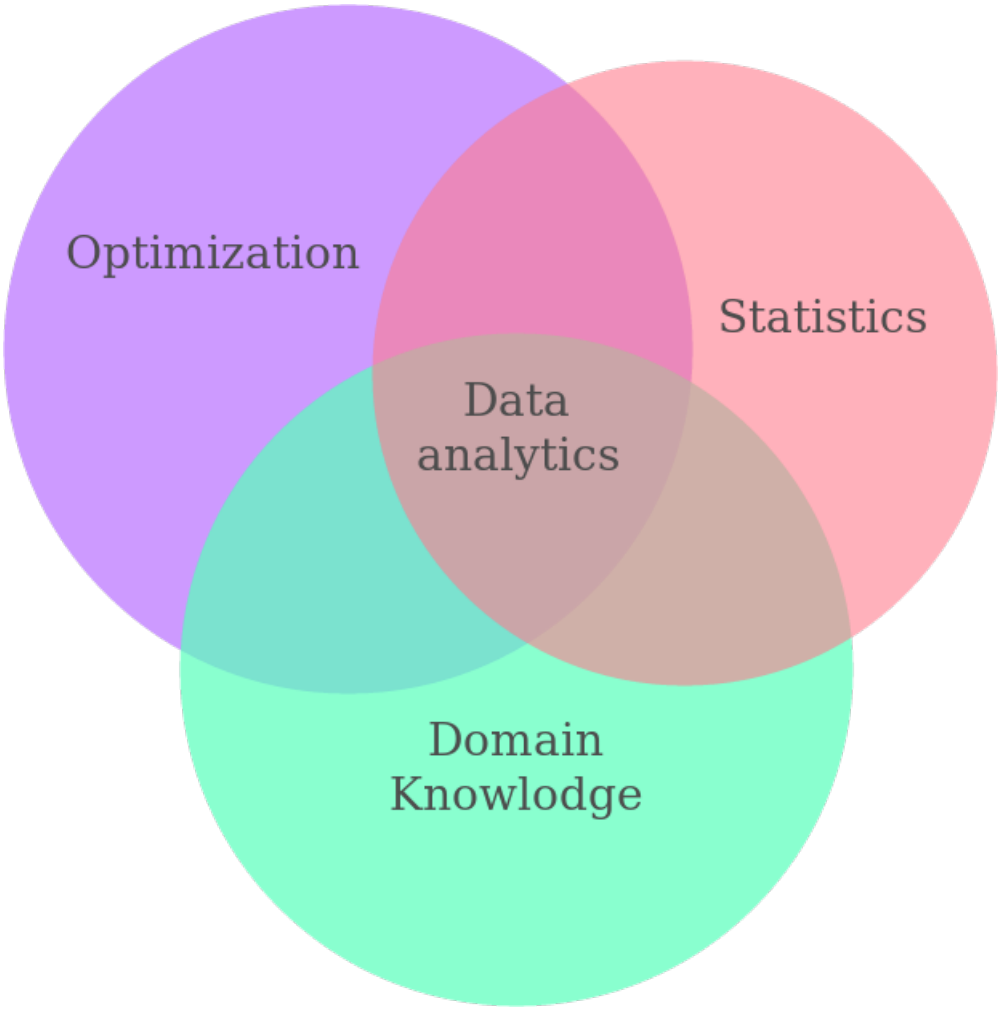

Figure 1.1: Data Analytics

Below we see a representations for the advanced analytics and business analytics 1.2.

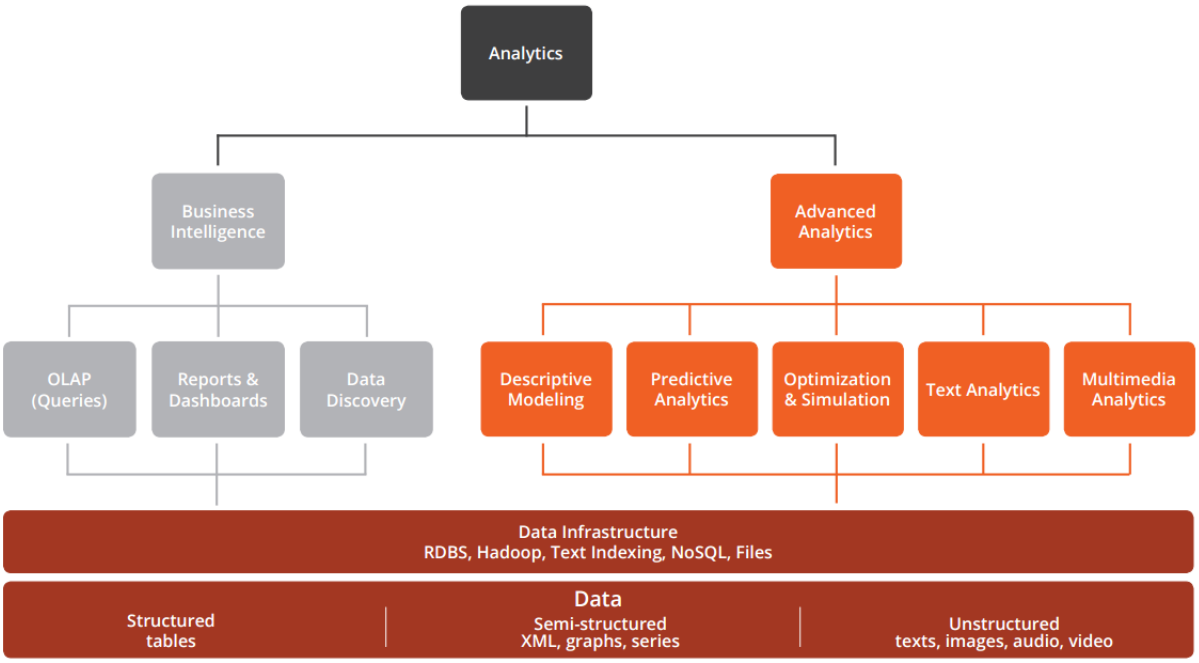

Figure 1.2: Advanced analytics

1.2 Objectives

The overall objective of the study report is to build an understanding of the tools and techniques of analytics used by financial analysts and institutions. This study presents a brief of the current state of affairs analytics and subsequently a study of the techniques. Also there is a demonstration of some common methods.

Chapter 2

Analytics in financial institutions

2.1 Why?

According to a joint research study, by Boston Consulting Group and Morgan Stanley with analytics professionals, it was reveled that the financial institutions lagged behind other verticals in the use of data analytics (Malhotra et al., 2017) and it is shown in figure 2.1. One of the findings of the research was that FI's are investing a lot of capital, an estimated total of about $1bn. In addition it was found that for near term value creation the FI's expected data analytics to optimize customer acquisition, customer retention, operational efficiency, and risk mitigation.

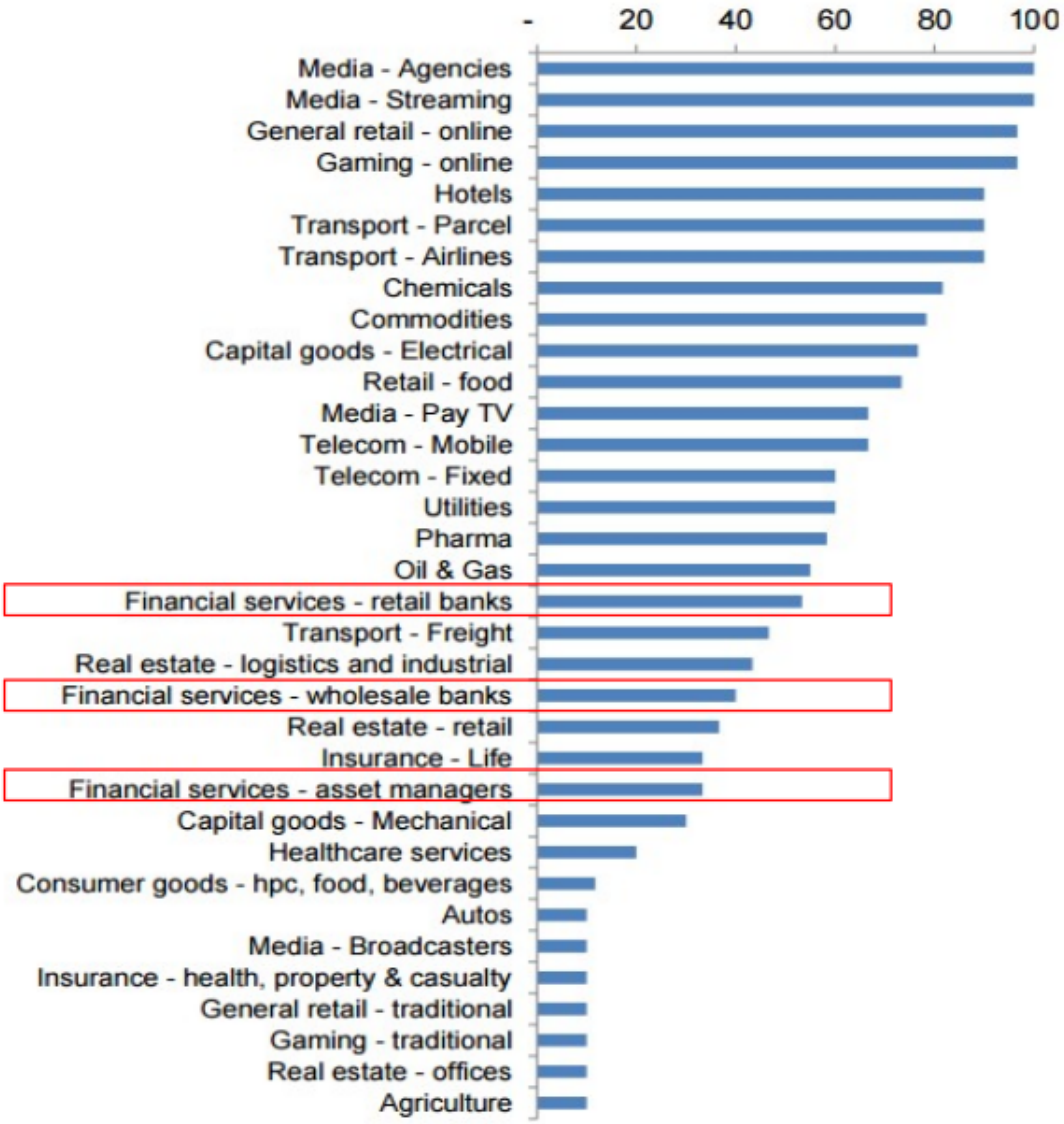

Figure 2.1: Reprinted from the Morgan Stanley Digitization Index ranks

2.2 State of analytics in financial companies

From the survey of FI's, a mix of interviewees representing payment companies, service providers, insurance, commercial banks, BCG made some interesting discoveries. It was found that most organizations have invested in analytics techniques to generate market perceptions. Some of the most used were those of social media, log, text and location analytics as shown in figure 2.2.

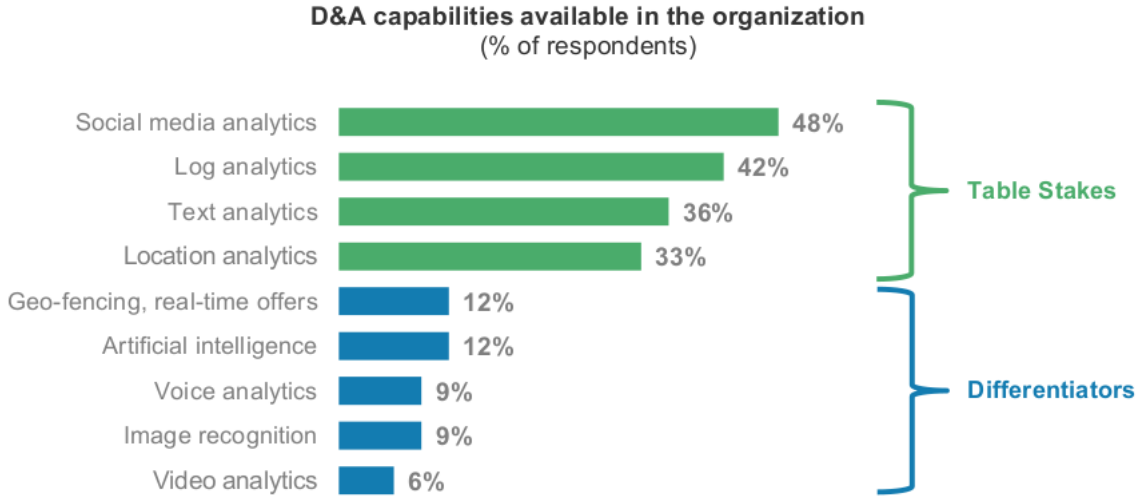

Figure 2.2: Reprinted from the work of Boston consulting group

Researchers have highlighted that most of the interviewees claimed that institutions are yet to make substantial gain from investments in analytics. They identified that companies can make gains from automation and digitization of manual processes. Additionally, it was noted that financial institutions are adopting digital processes to automate data collection for KYC (Know Your Customer) service and Anti-money laundering Customers are loyal consumers of products, which give value for their monetary investments. Big institutions tend to over look the need for value creation and focus their analytics and metrics to improve profit (Reichheld, 1996). In the HBR report Reichheld makes certain observations that customers of large companies witness degrading value standards and secondly increasing rate of customer churn is a good variable to predict cash flow from consumer to company. He also noted that companies can replace old customer with new ones but the profits are lower as cost of induction is high.

2.3 Challenges faced by analytics

For every new technology there are certain common challenges, such as those of acceptability and usability. Finding out the actual benefit or Return on investment for newer technologies and solutions such as those of digital assistants remains a mystery. There is no quantifiable or tangible benefit, but only a perception that it could revolutionize the method of banking and investing. Some hurdles, such as of vision, budget, and getting all members on board, are faced early in the life-cycle of adopting such bold complex projects. Institution leadership faces the tough task of decision making and if rests on them to perceive value at the end of project. Even after adopting analytics and tools, management faces challenges of educating shareholders, staff and customers to be literate enough to invest and interact with solutions for effective usage.

Chapter 3

Descriptive Techniques

The sole purpose of using the methods in this category would be to describe the Hypothesis. Descriptive analytics or data mining are at the bottom of the big data value chain, but they can be valuable for uncovering patterns that offer insight. A simple example of descriptive analytics would be assessing credit risk; using past financial performance to predict a customer's likely financial performance. Descriptive analytics can be useful in the sales cycle, for example, to categorize customers by their likely product preferences and sales cycle.

To summarize data into meaningful charts and reports, for example, about budgets, sales, revenues, or cost. They allow managers to obtain standard and customized reports, and drill down into the data and to make queries to understand the impact of an advertising campaign, for example, review business performance to find problems or areas of opportunity, and identify patterns and trends in data. Typical questions that descriptive analytics help answer are: How much did we sell in each region? What was our revenue and profit last quarter? How many and what types of complaints did we resolve? Which factory has the lowest productivity? Descriptive analytics also help companies to classify customers into different segments, which enable them to develop specific marketing campaigns and advertising strategies. [1]

There are two main approaches to apply in this topic Data warehousing and Visual analytics with reporting.

Descriptive analytics is very common and basic form of data mining technique to derive meaning and trend from data. Almost all of the companies in financial sector utilize the tools based on these techniques. The techniques in this are relatively basic compared with those of predictive models and prescriptive models. Institutions generally use tools which can generate reports regularly, some are printed monthly while some others are generated as a yearly routine. Such demands for reports describe the status quo of the balance sheets, organizational debts and cash flows. Daly transactions and updates of balances are maintained in the books or accounts and reports are generated daily for accountants and auditors. Thus it is imperative that vital functions of reporting, updating and auditing are functionalities of some business analytics.

There are several techniques employed by businesses to analyze and display data. They are as follows:

1. OLAP
2. Datawarehouse
3. Graphical views
4. Performance dashboard and KPI's
5. Social media analysis
6. Text analysis

3.1 Social media analytics

In a paper titled “Social media big data and capital markets—An overview“ (Bukovina, 2016), the researchers have studied the behavior of relationship between social media and the capital markets. The study relates to as behavioral finance.

3.2 Text analytics

In a paper titled “When machines read the news“ (Groß-Klußmann & Hautsch, 2011), the authors have used text mining to derive the relation between scrip news and intra-day stock trading. They try to relate the stock price fluctuations and liquidity, to the unscheduled news flash. The researchers used Reuter‘s NewScope is an news engine which classifies stock/company related news into positive, negative and neutral segments. The techniques used was VAR model known as vector autoregressive model. This model is used for multivariate time series.

The data was sourced from Reuter‘s NewsScope sentiment engine. It contains around 29,500 records of news headlines in the period of 1 year from the month of July 2007 to June 2008. Each news item has 3 attributes of sentiment, relevance and novelty. 40 stocks listed on the London stock exchange were selected for analysis. These are actively traded on the exchange and thus a dynamic market sentiment and frequent news coverage. These stocks cover about 70% of the FTSE100 market capitalization. According to the data analysis it was found that high frequency trading activity reacts positively to company specific news marked as relevant.

3.3 Location analytics

Valuation of goods and products varies from place to place. A certain item could be priced higher in a remote location compared to a normative value when sold in urban store. Location analytics is suitable for such marketers to price products according to demand and location. Improper valuation may decrease demand and eventually lead to unprofitable business. For example a bottle of water could cost 10 bucks in a city store, but could be priced about 5 times as high in an airport terminal. A research study published in the journal “Procedia Economics and Finance“, describes the real estate valuations in Romania (Droj & Droj, 2015). It compares the evaluation of studies the economic effects of using Location analytics and the standard practice of The study incorporates the use of Cadaster system for spatial analysis in the domain of GIS - Geographical information system. Cadaster system is a technique of mapping The researchers used a decision support system integrated with GIS functionality. It can improves manual valuations by taking in financial analysis of the property and fusing with the physical and social environmental influences. Normal pricing evaluations of properties are based on Hedonic models, ie. they price the individual elements of the properties and add them up. The pricing of bedrooms, kitchen, stories, size, are summed up. Factors taken into valuation of real estate are as follows :

- Physical location
- Social effects of proximity to schools, commercial spots, hospitals, nurseries, parks, criminality, recreational centers.
- Infrastructure of neighborhood such as roads, public transportation facilities, water and sewage systems, communications networks and Internet facilities.
- Environmental factors of air pollution, noise, industrial pollution, aircraft noise, traffic congestion etc.
- Economic price trends
- Legal constraints of building or modifying structures

In this study, researcher could automate the valuation of properties and determine the precise value. The cadaster system database is continuously updated by the administration with up to date taxation details. Hence it was possible to generate accurate valuations of living spaces.

Chapter 4

Predictive Techniques

The basis of Predictive modeling is the use of data mining techniques to forecast future results. Data mining is the process of sorting data to find patterns or infer relationships. Thus a formulation of a statistical model with relevant variables is essential for prediction.

Predictive analytics use big data to identify past patterns to predict the future. For example, some companies are using predictive analytics for sales lead scoring. Some companies have gone one step further use predictive analytics for the entire sales process, analyzing lead source, number of communications, types of communications, social media, documents, CRM data, etc. Properly tuned predictive analytics can be used to support sales, marketing, or for other types of complex forecasts.

ref from site (http://searchdatamanagement.techtarget.com/definition/predictive-modeling) and (http://searchsqlserve
mining)

Data mining techniques are used in credit card systems to detect fraud, in loan approval systems, identifying customer types and targeting specific promotional schemes. It is used to model customer behavior and detect churning to certain financial products. Insurance schemes and bank deposits are volatile instruments due to the uncertainty of the free economy. site (https://hbr.org/1996/03/learning-from-customer-defections).

There are many techniques to modeling predictive analytics

1. Clustering methods
2. Regession methods
3. Classification methods
4. Association rules

4.1 Geo fencing

Geo-fencing (or also known as Location Analytics) is the term for creating a location based analytics algorithm which can track customers in a certain area and deliver valuable marketing or services. The technology works based on geographic location services and tracks any potential clients with their mobile gps. Geo fencing tools are software applications which reside on the

Use-cases for geofencing :

- Manage a group of cargo vehicles

- Detect and optimize harbor docking for port management of ships
- Density flow of vehicles and management on roads or air space near airports.
- Exhibition centers can manage the flow of attendees to optimally view all exhibits, ensuring all items on display can be viewed.
- Shopping malls and shops can monitor the passage of people and optimize the display of commercials or products.

Geo fences are designed into the application which generally store the latitude and longitudes of the location to be monitored. When the user uses the app, it reads data from the gps module and keeps checking if the client has crossed any geo fence en-route. Now-a-days geo fencing.

Google maps geo fencing is available via API's

4.2 Artificial intelligence

AI was developed a long time ago, but it has for recent years gained much traction and entered our lives. AI was present from the 1950's, but it has only gained popularity. According to an article by PwC (PWC, 1996), some companies have invested in AI, Machine and Cognitive learning tools and have implemented solutions for Chatbots, Personal assistants etc. Big data coupled with faster computing and ubiquitous implementation with cloud computing has certainly boosted research and development. As per a Forbes research projection, in the next 10 years, implementations of AI will increase economic growth by 100% in about 20 countries. Also there could be an increase in productivity of around 40% of banks financial labor (Culp, 2017).

AI is used as a technique to automate various categories of tasks such as following (Rich & Knight, 1991) :

- Mundane tasks
  - Perception via Vision and Speech
  - Natural language processing
  - Robot control and common sense
- Formal tasks
  - Playing games such as chess, go, checkers etc
  - Solving mathematical problems of geometry, logic, calculus
- Expert tasks
  - Engineering design, fault finding, Manufacturing planning
  - Scientific analysis
  - Medical diagnosis
  - Financial analysis

Some techniques which fall in the space of AI are Support vector machines, Heuristics, Neural networks, Markhov decision process, and NLP natural language processing.

In an example case presented by a researcher from Romania (Costea, 2014), the techniques of artificial intelligence combined with fuzzy logic for classification. The case is applied to a scenario where the National Bank of Romania classifies NFI's into "good" or "poor" depending on the periodic financial reporting. The Non - Banking financial institutions of Romania are regulated by governmental policy to send their financial reports to the NBR for scrutiny. NBR appoints their staff to manually screen the reports and classify for goodness or badness and then need to proceed for on site inspection. The artificial neural network is built with genetic algorithm and trained to perform classification. The process involved introducing mutations to chromosomes of both the parents and offspring.

The dataset is made of 990 records having 11 attributes grouped into 3 dimensions viz., Capital adequacy, assets' quality and profitability. The data ranged between 2007 to 2012. There were 68

non banking financial institutions. 4 clusters were chosen using the Fuzzy C Means algorithm. The ANN structure contained 8 neurons on the first hidden layer and 5 on the next. The accuracy rate achieved was around 92.32% and validation accuracy around 91% and testing accuracy about 89%. Thus in this study the researcher was able to get high accuracy artificial neural network classification to classify the financial statements of non- banking institutions in Romania.

4.3 Deep learning

In the paper ref (Heaton et al., 2016) the authors have presented a list of techniques which are being used to classify and predict financial domain problems ranging from :

- securities pricing
- portfolio design and entity selection
- risk management

.

Deep learning models are being used in financial product design by working with big data. These models produce better results than the traditional techniques of economics.

4.4 Voice analytics

4.5 Image recognition

4.6 Video analytics

4.7 Combating financial fraud

In a research (Chen et al., 2015) , the Alibaba company developed an analytics solution to tackle fraud in the company. Big data from the company was leveraged to study patterns of financial transactions, user behavior and network analysis to predict in real time with machine learning algorithm, predicting bad user's and transactions. AntBuckler is a predictive analytics solution developed in-house to detect and prevent fraudulent transactions. Big data center of Alibaba use many fraud risk models to deal with activities. They apply the fraud and risk models on all processes related to account opening, identity verification, order placement. AntBuckler generates RAIN score - risk score for merchants and shared with banks, merchants for their judgment on risk levels.
Shown below is the framework Alibaba uses in Alipay 4.1.

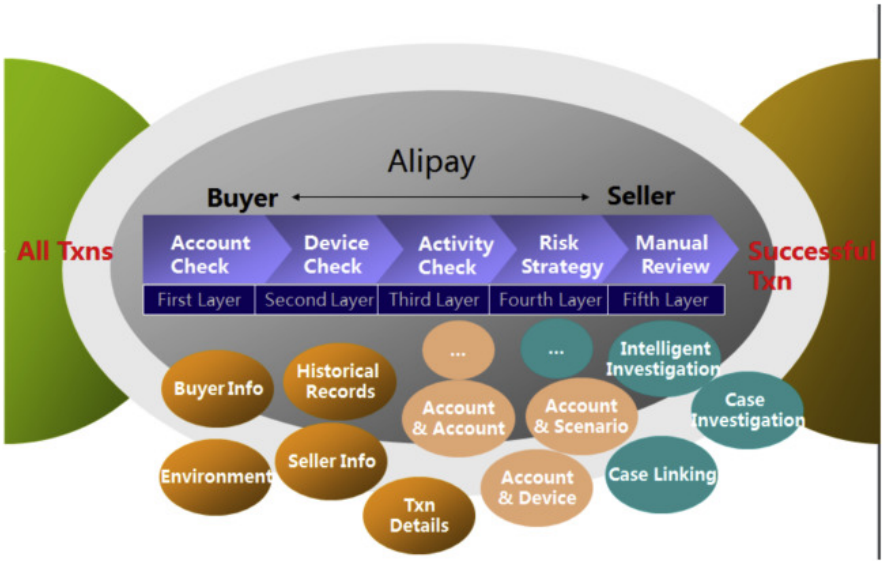

Figure 4.1: Fraud analytics in Alipay

Stages of fraud restrictions :

- Account check
- Device check
- Activity check
- Risk strategy
- Manual review

RAIN - model used by Alibaba stands for Risk of activity, identity and network 4.2. It identifies the risk of any object interacting with the Alibaba system such as a human, credit card or account. 3 dimension are defined in the system viz., activity, identity and network; and all variables are classified

into them, values of variables are computed for all the objects. Depending upon the scenario particular variables are chosen and RAIN score is computed. Such as in the case of credit card fraud, identity variables are adjusted with higher weights; in the case of credit scoring network group of variables are weighted more. Logistic regression part of a Machine learning is used to determine the weights for a given scenario.

Figure 4.2: Dimensions of RAIN score

Network based graph representation is also used to link accounts and ip addresses to visually represent the links between name, address, phone, credit card, etc.

Chapter 5

Prescriptive Techniques

A group of data analysis techniques which can suggest possible outcomes and prescribe recommendations and suggestions. Prescriptive analytics is really valuable, but largely not used. According to Gartner, 13% of organizations are using predictive but only 3% are using prescriptive analytics. Big data analytics in general sheds light on a particular subject, but on the other hand prescriptive analytics answers specific questions with laser-like focus. Generally the techniques are used to optimize output, schedule and manage inventory for supply systems

5.1 Game theory

5.2 Optimization Techniques

5.3 Simulation Techniques

In the paper title *"Monte Carlo method in risk analysis for investment projects"* (Platon & Constantinescu, 2014), the authors have studied the effect of random input values to the possible outcomes for risk assessment of environmental projects. The Monte carlo model is evaluated with a random input from a normal probability distribution and the output. Monte carlo simulation was formulate around 1944. The method is used to produce artificial values for a probabilistic variable by the use of:

- a random uniformly-distributed number generator in the $[0, 1]$ interval, and
- a cumulative distribution function associated with the stochastic variable

There were three steps in the process of simulation.

1. Select an investment project
2. Estimate risk of crossing value or cost of project
3. Estimate risk of extension in implementation

There exist alternate methods of plan and execution for risk mitigation :

- risk acceptance or tolerance

- monitoring of process risk
- risk avoidance strategy
- mitigation and outsourcing

Chapter 6

Research literature summary

In this section a summary of all the referred articles are presented in table 6.1.

Title & Author	Objective	Data	Results	Further Study
title	obj	20	30	something
title	obj	20	30	something
title	obj	20	30	something

Table 6.1: Literature review summary table

Chapter 7

Conclusion and Recommendations

This section presents the conclusion of study report.

7.1 Conclusion

In this paper we have tried to present a picture of the methods and tools of data analytics. They are summed up under three categories. Our motivation was to study the various methods and practices used by researchers and businesses. We hope this paper encourage new companies to apply softwares and schemes to analyze data. The analytic results could give them insights about To achieve what was not perceived, to deduce what was not understood is all made possible with data analytics. Thus it is our effort that the topics discussed in this paper would be of use to those who step into the world of Analytics without knowledge of background. References :

`Analytics`

7.2 Recommendations

Text..

References

Balm, P. (2015). *10 reasons why i love data and analytics — kdnuggets.* http://www.kdnuggets.com/2015/06/10-reasons-love-data-analytics.html. (Accessed: 2017-05-10)

Bukovina, J. (2016). Social media big data and capital markets—an overview. *Journal of Behavioral and Experimental Finance*, *11*, 18–26.

Chen, J., Tao, Y., Wang, H., & Chen, T. (2015, dec). Big data based fraud risk management at Alibaba. *The Journal of Finance and Data Science*, *1*(1), 1–10.

Costea, A. (2014). Applying Fuzzy Logic and Machine Learning Techniques in Financial Performance Predictions. *Procedia Economics and Finance*, *10*, 4–9.

Culp, S. (2017, Feb). *Artificial intelligence is becoming a major disruptive force in banks' finance departments — forbes.* https://www.forbes.com/sites/steveculp/2017/02/15/artificial-intelligence-is-becoming-a-major-disruptive-force-in-banks-finance-departments/#39f25de14f62. (Accessed: 2017-06-27)

Droj, L., & Droj, G. (2015). Usage of location analysis software in the evaluation of commercial real estate properties. *Procedia Economics and Finance*, *32*, 826–832.

Groß-Klußmann, A., & Hautsch, N. (2011). When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance*, *18*(2), 321–340.

Heaton, J., Polson, N., & Witte, J. (2016). Deep learning in finance. *arXiv preprint arXiv:1602.06561*.

Malhotra, S., Hizir, B., Badi, M., & Grealish, A. (2017, May). *The journey from insight to value — data analytics for financial institutions.* https://www.bcg.com/publications/2017/technology-digital-transformation-data-analytics-financial-institutions.aspx. (Accessed: 2017-06-17)

Platon, V., & Constantinescu, A. (2014). Monte carlo method in risk analysis for investment projects. *Procedia Economics and Finance*, *15*, 393–400.

PWC. (1996). *Artificial intelligence in financial services — from pwc's financial services institute.* https://www.pwc.com/us/en/financial-services/research-institute/artificial-intelligence.html. (Accessed: 2017-06-16)

Reichheld, F. F. (1996). *Learning from customer defections — hbr.* https://hbr.org/1996/03/learning-from-customer-defections. (Accessed: 2017-06-16)

Rich, E., & Knight, K. (1991). Artificial intelligence. *McGraw-Hill, New*.

Taylor, H. (2016, October). *Bank of america launches chat bot erica — cnbc.* http://www.cnbc.com/2016/10/24/bank-of-america-launches-ai-chatbot-erica--heres-what-it-does.html. (Accessed: 2017-06-27)