ISCTE IUL
Lisbon University Institute

Department of Information Science and Technology

Knowledge Extraction of Financial Derivatives Options in the Maturity with Data Science Techniques

Lídia da Conceição Silva Eira

A Dissertation presented in partial fulfilment of the requirements for the Degree of

Master of Information Systems Management

Supervisors:

Doctor Paulo Cortez, PhD, Associate Professor with Habilitation,

University of Minho

Prof. Raul M. S. Laureano, Assistant Professor, ISCTE Business School,

Department of Quantitative Methods for Management and Economics

September 2016

List of abbreviations

ASUM-DM	Analytics Solutions Unified Method for Data Mining/Predictive Analytics
BI	Business Intelligence
C&RT	Classification & Regression tree
CHAID	CHi-squared Automatic Interaction Detection
CRISP-DM	CRoss Industry Standard Process for Data Mining
DBMS	Database Management System
DM	Data Mining
DT	Decision Tree
DTB	Deutsche Terminbörse
EUREX	European Exchange
EMIR	European Market Infrastructure Regulator
FN	False Negatives
FP	False Positives
GUI	Graphical User Interface
IBM	International Business Machines Corporation
GCM	Global Clearer Member
KDD	Knowledge Discovery in Databases
KNN	K-Nearest Neighbour
LIFFE	London Financial Futures Exchange
MONEP	Marché des Options Négociables de Paris
MATIF	Marché à Terme International de France
NCM	Non Clearer Member
SEMMA	Sample, Explore, Modify, Model, Assessment SQL - Structured Query Language
SOFFEX	Swiss Options and Financial Futures
SPSS	Statistical Package for the Social Science
SunGARD	Sun Guaranteed Access to Recovered Data
SVM	Support Vector Machine
TN	True Negatives
TP	True Positives

1 Introduction

To improve the level of support in information systems and quality of services by questioning the daily routine of a team using a set of financial evidence has been an interesting and challenging problem for many researcher and decision maker professionals. As part of a well-known investment bank that deals financial instruments like European-style options derivatives, operational teams are well aware that the focus of their work are around the evolution on pricing until the expiry moment.

The choice of knowing more about financial derivatives options, especially in the maturity period, was made after a long process of study on economics and financial concepts in a certain institution. A special attention was given in subjects where information technology teams have less knowledge, which are the mathematical operation of derivative financial options and their implications in financial terms. As well, the identification of areas of business could be studied with greater interest for a specific organisation.

One difficulty encountered throughout the study, is to deal with the enormous amount of information in the organisation. According to it, it is confidential not only in disclosure of it but also the transfer of information within the same as full protection policy, as well as the prestige of its clients and the institution itself. Much of the narration here was an acquired knowledge on the point of view of the participant observer (the researcher), with the conditioned possibility check with the institution's documentation for an academic environment. Some notes mentioned herein are the result of reflections acquired from everyday experience, as a business analyst in the Listed Derivatives Information Technology (IT) Project team. After a short interview with a senior member of the technical support team of the whole architecture, it was found that one of the areas that have less information across the vast world of derivatives is the **Exercise** process (confirmation of an option) at maturity date. The study may be relevant to give more knowledge to the team of business operators of the bank as well as a support level of information systems. It should also give the perspective of new projects that current information technologies in the field are unable to provide, and not visible by just looking to data randomly. Given that it has not been found considerable articles that specifically combines the financial and technical options with Data Mining at the time where the buyer/seller performs the Exercise, this is therefore a timely topic, in addition to enhance knowledge of the respective courses that occurred in the first year of the current master.

It is as well at this point that the central objective is unravelled – the acquisition of knowledge of financial derivative options in the maturity date, trends and forecast events through Data Mining (DM) techniques – which is added to the context of financial derivative products found in different markets and governed by different rules under the purview of various operational teams working daily in joint projects with IT teams, in constant search of obtaining benefits for their stakeholders. Quoting the research manual in social science of Raymond Quivy: Gaston Bachelard summarised the scientific process in a few words: “The scientific fact is won, constructed and verified” (Quivy, 1992, p. 3), which means won over prejudices, constructed by reason and verified in facts. Commenting this statement, the research must respond to this challenge, the initial phase is focussing more on the thought to win and break preconceived ideas, in order to further build and verify during the course of the study. Finally, it matters referring that this introductory phase is divided into four parts: background with information relating to derivatives and DM techniques; goal setting, theme and assumptions; the explanation of the methodology which is believed to apply to answer the problems described below, and finally the structure of this document.

In this perspective, and after confirming the lack contingency planning on events during the maturity date in the Organisation, as well as the Option Watch application and its database is not revised since 2008, this knowledge extraction could be pertinent in order to assist the various operational and technical teams, to prepare the unpredictability during the day of maturity of options contracts in the future. With the exponential volume of financial options, the acknowledge of a sudden increase of data in tick moments would be a key factor to properly support operational teams geographically disposed, where a transitional period of mass automation is taking in place and delicate situations may occur in the future.

1.1 Problematic and motivation

This study started around the following question: how to get useful knowledge about derivatives options? And how information systems can contribute to this higher knowledge? The first step is to understand the concepts of each term. In a very simple way, a financial derivative option is an option contract celebrated by three parties, with an underlying asset and a maturity date in the future. The underlying asset may be, for example, a listed share on the stock exchange, and a maturity date could be of a specific day several future months. The parties involved are: the buyer, the seller and the entity that assumes the risk, the Clearing House (Investopedia, 2014).

As the Figure 1 shows, a contract of options has an extensive life cycle if we consider the settlement (the last stage where premium payments are completed). It is purchased in specific markets of derivatives and it is very sensitive to fluctuations of markets, depending of the underlying asset and the progress of external information relatively to its nature. It can be sold or purchased until the day of maturity. Once the maturity date arrived, it is exercised or abandoned. The process of exercise is a technical term in finance to the action of confirmation of an option, which means that a given option had its intention to achieve a confirmation state from its owner (who bought or sold), in order to collect the dividend agreed after the maturity period (Investopedia, 2014). For a financial derivative option specifically in the current Organisation, the exercise is performed manually by a professional operator and option by option. Some markets have connection to Option Watch application, thus providing automatic operations to that market (Operational Team MONEP of Organisation, 2013). Option Watch is an application used in the institution that provides an automatic exercise, instead of confirming manually option by option. This information system is being currently in a study phase to determine whether is widely used or not, and if so, perform the extension of this automation to other markets.

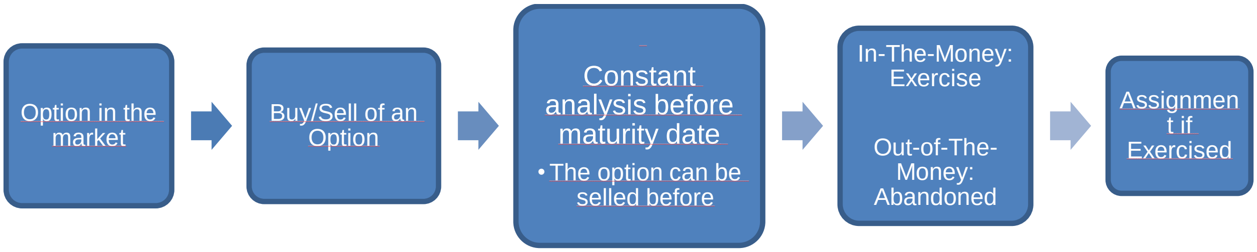

Option in the market
Buy/Sell of an Option
Constant analysis before maturity date
• The option can be selled before
In-The-Money: Exercise
Out-of-The-Money: Abandoned
Assignment if Exercised

Figure 1: Life cycle of a European derivative option

Adapted from (Organisation, 2013)

There are several types of options depending of its properties and relatively to its nature. To subset the universe and as the institution deals with more information on European markets, we will be concentrated on European-style options, which are options that can **only** be exercised (or abandon) at the maturity date whilst American-style options can be exercised **until** the maturity date.

Several studies have been made around options in a mathematic and financial way, but the system information's domain plays an essential role and must be considerate as well. It is intended that the research will lead to an improvement in the decision management for IT support process and reflected in the optimisation of resources (IT and human), in line with the strategic view of the institution (directly influenced by external factors such as international and national financial crisis that create more pressure on the rationalisation of resources available to the institution).

1.2 Objectives and contributions

It is intended to carry out a case study based on the data from application solution option derivatives of an International investment bank, the **Organisation**, in an attempt to identify behavioural patterns of options that allow optimal operation during the maturity date. To this end, Data Mining techniques and financial options methodologies areas are used for extracting useful knowledge. The data to be analysed corresponds to records saved between 2012 and 2015 and comes in several tables carried out by various databases (Oracle and SQL Server databases).

The application solution falls within the area of operational teams of two of the most important markets, since hedgers and traders are selected, gird up into the application to manage. Thus, being the focus of European-style financial options in the maturity date, the research work will focus on Data Mining and particularly to system information, i.e. the discovery of patterns that translates into useful and applicable knowledge in the case study. With the aim to acquire more knowledge on financial derivative options in order to the technical teams be able to support the operational team at the maturity, data mining techniques could help to focus on potential causes, by transforming data into knowledge.

Thus, the objective proposed for this research is to predict the classification of exercise of a particular European financial derivative option at the maturity date.

The fulfilment of this objective and, consequently, the knowledge generated could suggest measures in terms of information systems efficiency, in order to have a greater knowledge on the predictability of more accurate data over a given period of time with limited resources. Then, study may have influence on strategic Organisation of resources and rationalisation of available operators, but also a contribution in the scientific knowledge, as there weren't found substantial studies of this subject in the last years regarding the information systems and decision making field rather than the financial field.

1.3 Methodology overview

A case study on the data collected from application solutions of an international investments bank will be conducted, focusing primarily on the period of maturity of MONEP and EUREX financial derivatives options. To do this, techniques and data mining methodologies will be used directly to the gathered data. The data to be analysed corresponds to records obtained between August 2012 and July 2015 and come from the acquisition of contracts from the aforementioned markets. This acquisition is performed in two ways: manual and electronic. Manual i.e. performed by an operator entering the data manually in the Option Watch applications (OPW). Electronic, when the insertion occurs transparently through specialised applications without human interference.

This research is a participant observer case study, inserted in the IT project team of financial derivatives options of an investment bank market leader. The data was collected in this institution during the period of July 2015. There were selected 328 598 valid contracts of two financial European-style option markets – EUREX and MONEP – between the period of August 2012 and to July 2015, having selected 27 attributes after a selection from 5 different tables and from 2 different databases. The data is a bit raw, not normalised. Each column of the datasheet (attributes) must be analysed, treated and prepared at first. It is a high number of records and the majority is complex to extract, and the lack of technical knowledge financially speaking had increased the complexity. Therefore, the assistance of the course of financial derivatives was the step 0 before starting the study. DM techniques are adequate to extract some knowledge. The use of specific methodologies is essential for a better understanding of relating data. As a process model, the CRISP-DM methodology will be used because it is the most widely accepted approach in the field, plus some tools have been practice along with the important tools such as R and SPSS Modeler. CRISP-DM is chosen because it is focused in to only in data but also in understanding the business, by understand, prepare, transform, model and finally evaluate and deploy a chosen model. Thus, the uncovered challenge is to be able to find an algorithm that fits our objective. Finally a discussion can take place by point contributions, limitations and research opportunities.

1.4 Structure of this document

This document is organised into five chapters. The chapter 1 presents the introduction having the exposition of the problematic, objectives and framework. In the chapter 2, the state of the art is reviewed for the areas of finance. Chapter 3, more practical, presents the methodology adopted and all the data mining development and knowledge extraction. The results are presented and discussed in Chapter 4. Finally, Chapter 5 presents the respective conclusions, contributions and future work directions.

Additionally, it should be noted that the financial technical terms here presented are set out in the original language (English or French), having an explanation whenever possible on the basis of the current page with a brief description and translation of terms. This option is due to the fact the original terms are globally known in their respective specialized areas. To facilitate the reading of this document, there is also a list of abbreviations available at the beginning of this work.

As suggested by the CRISP-DM methodology, a glossary of terminology relevant to the project was compiled (Annex A). It includes two components: relevant business terminology, which can be a useful “knowledge elicitation” and education exercise; and a glossary of data mining terminology, illustrated with examples relevant to the business problem in question (Chapman et. al, 2000).

2 Evolution of derivatives and financial options

2.1 Concepts and two powerful markets

Stefan Bernstein wrote a practical definition of a derivative. In his opinion, "A derivative is no more than a financial instrument that derives from an underlying asset or action" (Bernstein, 2009, p. 18). In other words, a financial derivative product is a contract made between three parties for the purchase of a product, signed today but with the acquisition of the underlying product contracted at a future date.

Achieving that date, the buying decision is analysed which can be accomplished or refused. Instead of purchasing, for example, 10 shares "Abc" to a value of 0.5€ per share, it can be transacted an options contract worth a total of 2€ or 0.2€ per share, to buy them at a future date. If the price of the underlying asset (i.e. 10 shares) rise during this period, then the 2€ or 0.2€ per share were well invested and will be the advantage, than having just bought these shares on the spot market (Bernstein, 2009).

The derivatives are typically traded in large amounts and often targeted to large institutions, often acquired for hedging and therefore may allow gains (and losses) in a fast and dangerous way, stimulating the markets and investors. Who realizes these purchases and sales between the parties are the teams of operators, existing for at least one market, and are in constant communication with professionals who represent the parties to the contract (Hull, 2014). Thus, operators of financial derivatives of a known institution of international banking investments caution throughout the life cycle of a product derived from a given market. A large IT staff and a complex support architecture to information systems make possible these operations runtimes closer to real time and where everything is automated: from clearing ("pairing" between demand and supply) the tax calculation, the investor data constant checking and passing information between them, markets and regulators (example: EMIR - European Market Infrastructure Regulation) (European Commission, 2015), as well as the clearance of gains and losses on the contract maturity date (also called expiration date or and simply "the expiry" used in day-to-day by the operators) (Bernstein, 2009). Although, in derivatives, this process is essential, operators know very little about it. They just know that the application distributed by the supplier that performs this process without major problems, having a database created for this purpose but there is so far no study in the Organisation about it, despite producing gigabytes of records year after year.

Standard maturities are maturity dates previously defined, typically the third Friday of each month. Speaking of standard maturity dates, this process usually occurs every month (for standard maturities, it occurs on the third Friday of each month), and is the confirmation of the contract on a certain date, giving permission to pay the premiums, close the positions and report to the client all contracts made during the day of closing, week, month and year (Hull, 2014).

Actors that directly contributed to the data collected on the derivative options are varied (Operational Team MONEP of Organisation, 2013) and can be grouped into two fields: technical and business. It is essential to understand both sides in order to understand the problem in both perspectives. In the business side, markets, investors, brokers and operators deals with the data by using the information systems as a service, whilst in the technical side, software suppliers and IT teams deals with the permanent objective of sustainability of systems. For the sake of the institution that held information systems and in both sides (business and technically speaking), it is crucial to have the knowledge of peak moments where more resources (information systems and human) are needed (Avellaneda, Kasyan, & Lipkin, 2011). The Figure 2 shows the overview of operation needed in a derivative product.

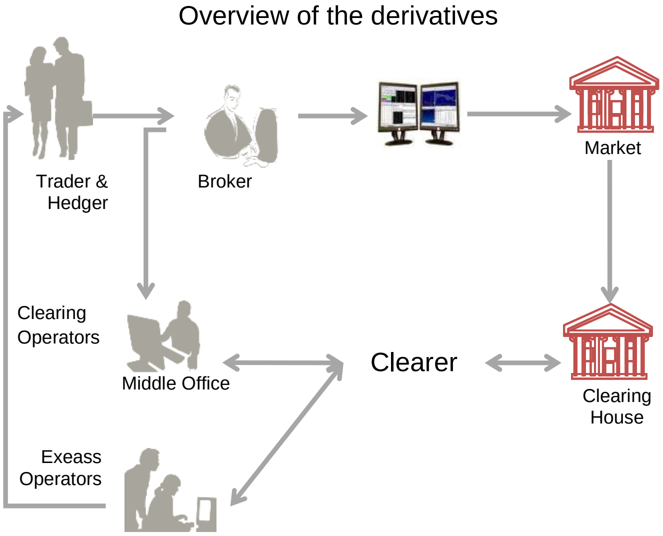


Figure 2: Actors of a derivative product in 2015

Adapted from (Operational Team MONEP of Organisation, 2013)

Hedgers and traders are the investors of derivatives products which have contributed to they evolution in opposite ways (Bernstein, 2009). They communicate their will to do business to the broker who introduces the information in the system of the institution (Operational Team MONEP of Organisation, 2013). Operators will conduct the strategy proposed by the broker affecting the clearing with a certain clearer. This clearer will communicate the clearinghouse and the market the ownership of the risk, and when the maturity arrives, it is exercised/assigned or abandoned my “Exeass” operators. “Exeass” is an informal designation made by each market operators for the exercise/assignment in the maturity period (Organisation, 2013).

As referred in the objectives, the study carried out concerns an analysis of the derivatives options at the maturity date, and so it is important to have a previous study of this financial instrument, as well as understand what techniques are available to collect more precisely the wanted data. As derivatives, the most important European markets are EUREX and MONEP exchanges but as we can see in tables 1 and 2, they endured profound changes. Since its inception in 1848 for MONEP exchange and in 1999 for EUREX exchange, markets were targeted with mergers through the years, the most important being summarised in the following tables (Organisation, 2013):

Table 1: Evolution of MONEP exchange

Date	*Evolution*
1848	Creation Market Chicago Board of Trade (CBOT- merchandise)
1970	Introduction of the first financial derivative market
1986, 1987	Creation of the MATIF (Marché to Terme International de France) and MONEP (Marché des Options négociables Paris).
1999	MATIF and MONEP merged creating the Paris Bourse SBF-1999.
2000	Joined the Paris Bourse SBF-the AEX (Amsterdam Exchange) and Belfox (Belgium Futures and Options Exchange), creating EURONEXT.
2001	Integration of Portugal and then the UK on EURONEXT-LIFFE
2007	2007 EURONEXT-LIFFE merged with the NYSE to create NYSE-EURONEXT

Adapted from (Organisation, 2013)

MONEP name market has been constantly changed, as it has been merged with other markets. To facilitate the nomenclature, it was decided to call the MONEP, even if this means the now giant NYSE-Euronext, containing (MONEP, MATIF, AEX, Belfox most recently LIFFE and NYSE). The EUREX market was an early proponent of electronic trading. The Called "Battle of the Bund" in the late 1990s (Figure 4), during which traders changed their volume of open outcry trading pit business on LIFFE to fully electronic, instead of manually trading by operators (Estelle Cantillony, 2008). The EUREX market was globally known as critical to the success of electronic trading (Figure 5).

Table 2: Evolution of EUREX exchange

Data	*Evolution*
1990	Creation of the Deutsche Terminbörse (DTB, derivatives Frankfurt market)
1998	Merger of DTB and SOFFEX Market Swiss options and futures, both founded in the late 1980s and created the EUREX (European Exchange).
1999	EUREX dethroned the LIFFE market started, to be considered the most sophisticated market, using almost exclusively electronic trading (because it is less subject to human error).
2014	EUREX Exchange is the largest European derivatives exchange and the third ranked globally as measured by traded volume

Adapted from (Organisation, 2013)

The first option arose in 1848 in Chicago, but only in the early 60's there was a strong growth and specialisation. Gone are the days when a derivative was traded manually in a noisy room open outcry call as a way to communicate between professionals and the markets, as bellow illustrated. In the 80s, with the arrival of electronic contracts the Trading Pit – area more traditionally driven to commodities: Buying of goods like cereal or Brent barrel –, business was decided in the room through the interaction of professional with many different coloured uniforms, as shown in Figure 4 (Bernstein, 2009).

Figure 3: Market trading room "New York stock Exchange" September 1963

Font: (Library of Congress, 2015)

Figure 4: Transaction area of commodities in the early 90s

Font: (Bernstein, 2009)

With the fusion of markets, the increased regulation from the subprime 2008 (European Commission, 2015) and the development of unprecedented information systems, the open outcry type rooms were fully replaced digital flat screens (Figure 5), high-performance servers and inserted a great IT architecture provided by the institutions.

Figure 5: The same room at January 22nd of 2015

Font: (Lothian, 2015)

Figure 6 is a graph that shows the systematic growth of the volume option contracts of EUREX business. Since 1998, the market has managed to maintain a high volume and has overcome having great offer-response capacity a mainly thanks to the revolution of information systems. The computation in financial services is crucial and low time response and how to anticipate a big volume of transaction in a certain period has been more and more important, not only to markets but as well to survival of financial institution like the Organisation.

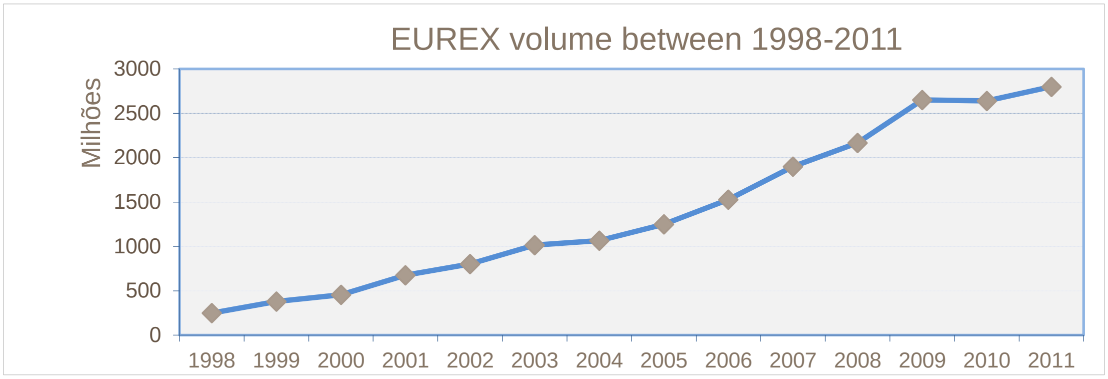


Figure 6: Annual turnover of EUREX

Adapted from (Organisation, 2013)

2.2 Derivative financial markets

Derivative financial products have been used for centuries to facilitate transactions and drain orders and market studies (Hull, 2014). A derivative is an instrument to which its value depends on your contract or underlying asset. Some examples of derivatives are contracts, options and swaps, among others (Placeholder1).

The term financial contract is a commitment to carry out transactions of a financial asset in a future period, and both parties (hedger and trader) agree to buy or sell in the future at a predetermined and agreed price at baseline (zero time). These contracts can be futures, forwards or swaps. It should be noted that both parties have different reasons, and each of them developed the market over the years in opposite directions. Thus, a hedger is the part of the contract that purports to cover, protect, avoid being exposed to vulnerabilities such as if a person owns a loan, taking into account the fluctuations in interest rates. But the trader is "opportunistic" analysing market conditions and can trigger events to take advantage (Bernstein, 2009). The figure 7 shows the relationship between the contract and its underlying asset, highlighting that they belong from two different types of market under different conditions. It means that their narrowed relation tail the trends (if the price of a stock rises, the price of the instruments that has this underlying product might suffer influences (Avellaneda, Kasyan, & Lipkin, 2011).

Alternatively, financial derivatives futures contracts are contingent commitments to make transactions with assets in the future, but dependent on events. This is the main difference between futures and options. As for the forwards are only tradable Over-The-Counter instruments, i.e. without any regulatory body protection of both parties (example: a clearing house), but all the same as futures contracts (Hull, 2014).

The main difference between the future financial instruments and options is that the future has the obligation to buy the underlying asset and an option has the option to buy the underlying asset at maturity date. This means that the options are, at first glance, more interesting to invest, but also more complex, because it has many more variables to consider and compare, including: dividend yield, volatility, interest rate without risk, the quantity, the strike price[1] and maturity date (Hull, 2014).

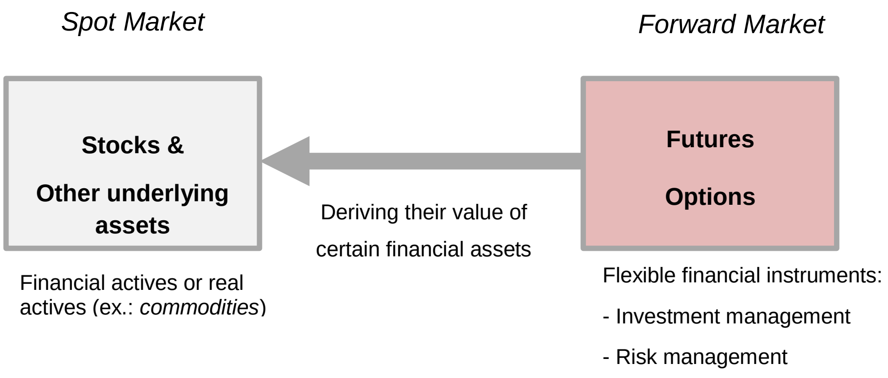

Spot Market
Forward Market
Stocks &
Other underlying assets
Futures
Options
Deriving their value of certain financial assets
Financial actives or real actives (ex.: *commodities*)
Flexible financial instruments:
- Investment management
- Risk management

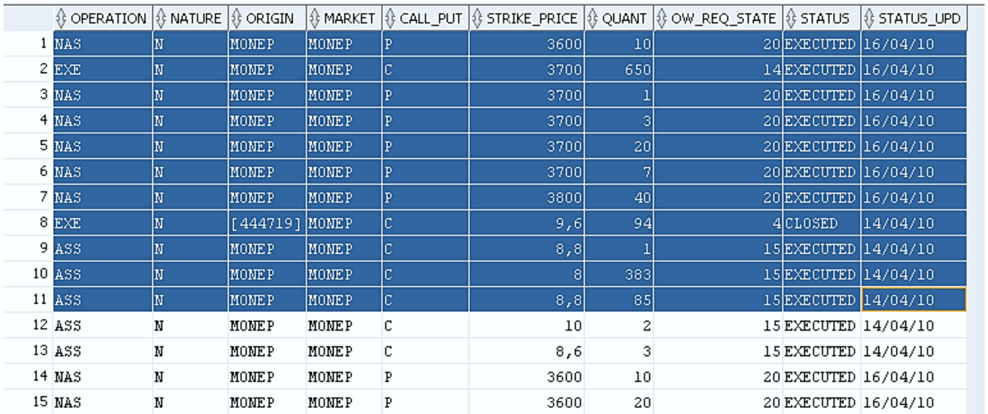
Figure 7: Underlying asset relation with complex contracts

The option is an asymmetric agreement. The option buyer (long position) has the right to buy an asset, whilst the seller is obligated to sell the asset if the buyer decides to exercise his right. For this reason, the holder of the long position is always an advantage and therefore the long position must initially pay the contract price (> 0), contrary to what happens to the future. Future contracts are symmetrical: both positions they can take (long or short) must be respected equally (Hull, 2014).

As Bernstein explained in page 30, in general the options offer great utility to: Ripening current portfolios against market fluctuations (hedging), cash generators from existing portfolios (for instance, monthly rents), and allows make a contribution to the stock market at a lower value. However, the gains are dangerously commensurate with its risk, it is important to have a thorough knowledge of the product. The exercise price is the amount of profit the option on the expiration date previously agreed (Investopedia, 2014), in 2014. The difference between the fixed exercise price and the market price at the time the option is exercised is what gives the value. Generally and according to Hull's bible at page 185, the larger the difference between the exercise and the market price at the time an option contract is written, the greater the required initial premium to acquire the option (Hull, 2014). The Assignment occurs when an option holder wants to exercise its option by notifying your broker (professional playing contracts awarded by the trader and main referral commissions and customers. It is also the main communication link between the trader and the Organisation (Bernstein, 2009), which then notifies the OCC (Option Clearing Corporation): notifies the seller that a given

[1] Strike price is the price of the contract agreed by the parties to which may be confirmed (or not) at maturity. It serves as a reference during its life cycle.

option was notified by the new holder of the option, and is made to transfer the obligation rights in Settlement. If an option is not "claimed" not exercised, then it is abandoned, by assigning to Abandon. The Novation process is the technical term for the process where contracts between clearers members are taken by the clearinghouse. For a better understanding, the Figure 8 illustrates the type of operation that occurs on the date of maturity of a given option. Here are all the MONEP market and operations are: EXE = Exercise, ASS = Assignment, NAS = Non-Assignment.



	OPERATION	NATURE	ORIGIN	MARKET	CALL_PUT	STRIKE_PRICE	QUANT	OW_REQ_STATE	STATUS	STATUS_UPD
1	NAS	N	MONEP	MONEP	P	3600	10	20	EXECUTED	16/04/10
2	EXE	N	MONEP	MONEP	C	3700	650	14	EXECUTED	16/04/10
3	NAS	N	MONEP	MONEP	P	3700	1	20	EXECUTED	16/04/10
4	NAS	N	MONEP	MONEP	P	3700	3	20	EXECUTED	16/04/10
5	NAS	N	MONEP	MONEP	P	3700	20	20	EXECUTED	16/04/10
6	NAS	N	MONEP	MONEP	P	3700	7	20	EXECUTED	16/04/10
7	NAS	N	MONEP	MONEP	P	3800	40	20	EXECUTED	16/04/10
8	EXE	N	[444719]	MONEP	C	9,6	94	4	CLOSED	14/04/10
9	ASS	N	MONEP	MONEP	C	8,8	1	15	EXECUTED	14/04/10
10	ASS	N	MONEP	MONEP	C	8	383	15	EXECUTED	14/04/10
11	ASS	N	MONEP	MONEP	C	8,8	85	15	EXECUTED	14/04/10
12	ASS	N	MONEP	MONEP	C	10	2	15	EXECUTED	14/04/10
13	ASS	N	MONEP	MONEP	C	8,6	3	15	EXECUTED	14/04/10
14	NAS	N	MONEP	MONEP	P	3600	10	20	EXECUTED	16/04/10
15	NAS	N	MONEP	MONEP	P	3600	20	20	EXECUTED	16/04/10

Figure 8: EXE, ASS and NAS of expiry options in Option Watch database (test environment, 3/2015)

Retrieved from (Organisation, 2015)

This picture shows important information of the table expiry options, with the expiry date of 16/04/2010. The column status displays one of the possible statuses that an option can have: “executed” or “closed”. If the status is “executed”, it means that the option was claimed, refused or assigned. This assumption seems confusing but if we observe the combination of this column with the column operation, which is the value that we want to forecast, the consequence is that a “closed” option means that the three parties of the agreement accepted the operation. The other columns represent other characteristics of the option and the columns market, call/put, strike price and quantity could influence the result of the column operation.

Figure 9 shows the Option Watch possible views for a regular user participating in the process of Exercise/Abandon. In this case, the user selects the option "Positions view" as seen in Figure 9. The next figure illustrates the layout of Option positions for the market MONEP, with the main two options available: Exercise or Abandon.

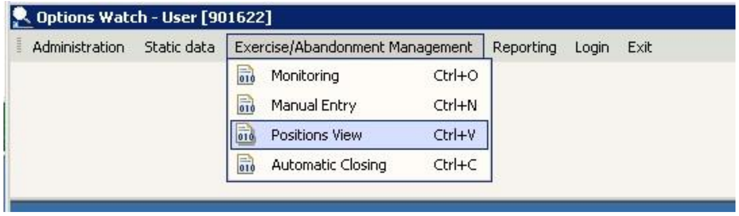


Figure 9: Option Watch user 901622 option views in the Exercise/Abandon Management

Retrieved from (Organisation, 2015)

All these flows in parallel are "marked" by CORIG, as shown in the table 3 some examples of FICNEG for EUREX market.

Table 3: Origin code (CORIG) application and its description

CORIG	*Description*
LISA_EUREX or LISA_MONEP	Integration of trades by Ubix module called *Tradefeed*
SRX_EUREX	Integration of EUREX trades for the member BNAPA (from the market to Ubix by SGW + SRX)
OPTIONWATCH-O	EXEAS of MONEP positions, through OPWs application

OPW interface writes in the daily tables FICNEG (‘fichiers de négociations’) and FICDEP (‘fichiers de dépouillement’ or portfolio), but it’s in HISNEG and IHSDEP history tables that the study is focused. The HISNEG and IHSDEP tables are similar tables. The difference is in the portfolio of each customer, and the ISHDEP contains all clients’ portfolios, therefore that is the target of the study, however, if we pay attention to prices, the table of negotiations is more relational variables.

2.2.1 Evolution price of a specific option

Evaluation of options has been the subject of many articles, but most work focused on options are in finding the best moment to exercise options, rather than performance or resource systems (Morales, Horrein, Baghdadi, Hochapfel, & Vaton, 2014, p. 1). In a finance point of view, the most exciting and debated topic of a European option is its price evolution: at the very last moment at the maturity date and just before the market is closed the option is *at-the-money* (ITM) or *out-of-the-money* (OTM)? However, as pointed the leading options trader Jeff Augen "unusual market forces create option price distortions" (Augen, 2009, p. 52). In other words, some market makers and traders tend to "pin" the price of the share at a strike price. Near option expiries (i.e., expiration date of options), market makers are able to control the price to a certain extent in order to ensure that the maximum "pain" is inflicted on the majority of long option positions. This means that the difference of prices between the underlying asset and the option contract might tend to 0, which is at-the-money (ATM) and the option typically is abandoned in the very last moment because it's not worthy anymore to exercise it, as it has premium costs to pay to whom the option is been bought (Avellaneda, Kasyan, & Lipkin, 2011, p. 3).

Regarding the evaluation of options, the challenge is to know the composition of the portfolio that replicates the option because it depends on the spot price of the underlying asset, at a time *t*. This has implications in the payoff, because in the next time t, this value is different. There are two algorithms that *in theory* (in perfect conditions and evaluated without any leverage or arbitrage argument) an option can be calculated: by Binomial Tree (Morales, Horrein, Baghdadi, Hochapfel, & Vaton, 2014) and by Black Scholes models

(Black & Scholes, 1973). Derivatives options are used since the nineteenth century and there have been studies since its beginning, due to impressive changes in the way business are being taking place according advances in informatics and efficiency in dealing with a lot of information, but the valuation of derivatives options are still under the old Blacks-Scholes model perspective (Black & Scholes, 1973). This model of giving a theoretical estimate of the price of European-style options seems to be closed to real value of an option once the maturity is achieved in most of cases (Zvi Bodie, 2008), but it could be adjusted by the volatility "option smile" model (Hull, 2014, p. 335). Instead of the expected flat surface, for a given expiration date, options whose strike price differs substantially from the underlying asset's price command higher prices than what is suggested by standard option pricing models. This is why we assist to the "smile" appearance in the graph of the Figure 10, which shows the volatility implied patter.

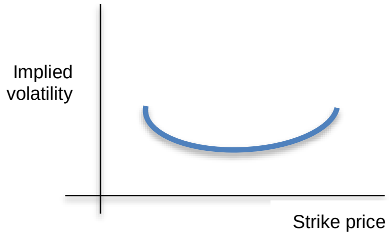


Figure 10: For a given expiration, options whose strike price differs substantially from the underlying asset's price command higher prices, implying volatilities

Adapted from (Zvi Bodie, 2008)

These options have thus implied volatility and are said to be deep ITM or OTM (Zvi Bodie, 2008). An option is a powerful instrument full of variables that need to be carefully considerate. A <u>European</u> option is a contract where the owner has the right (but not the obligation) of buy (<u>call</u>) or sell (<u>put)</u> a determine quantity (<u>contract size</u>) of an active product (<u>underlying asset</u>), at a determine price (<u>strike price</u>) at a future date (<u>maturity</u>). Increasing the complexity, a trader/hedger can have the call option to buy (long call) or the obligation to sell (short call), but also a trader/hedger can have the put option to buy (long put) or the obligation to sell (short put).

The payoff (the benefit received in options) can be interpreted as the Figure 11 shows, where *k* is the strike price and s the evolution of the price over time of the option. We can see the long call (the option to buy) is the one who's taking less risks and the short put position is the least envisaged, because it has the obligation (short position) to sell his/her option, unless there is an attractive compensation.

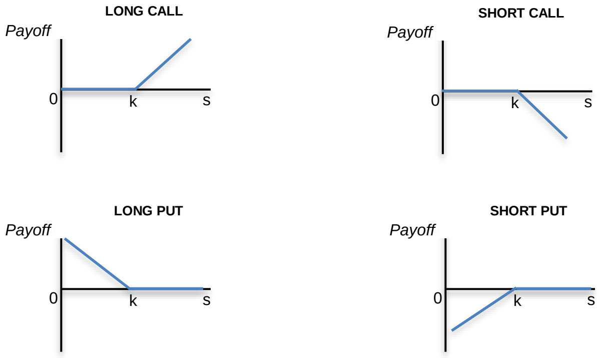


Figure 11: *Payoff* charts of the four possibilities of options

Adapted from (Hull, 2014)

2.2.2 Latest studies

With the advances on computation and, as described in the previous point, unprecedented experiences in finance, It has been studied lately how to confirm the accuracy of the Black Scholes equation by using computational techniques (Jeong, Yoo, & Kim, 2016). In this study, values of variables of the equation are essential input data to perform tests near the expiry date, as these variables are the cause of fluctuations. Mathematically speaking, the Black-Scholes equation is a partial differential equation (PDE) with the objective to find the price after a certain delta T, for a European option (Jeong, Yoo, & Kim, 2016). Furthermore, there is an on-going joint venture merger of the derivatives with OTC financial products (Over-The-Counter). The OTCs are financial products equivalent to derivatives but not regulated, a type of contracts transacted between two shares in markets where there is no clearing house to protect it from possible losses of both parties. They are a “telephone – and computer – linked network of dealers" (Hull, 2014, p. 2).

3 Methodology

This chapter is the work core of this study. It starts with an introduction to intelligent data analysis in section 3.1 followed the relevant by methods and techniques in context, sections 3.2 and 3.3. The second moment of this chapter is the description of each step of the process model chosen, the CRISP-DM itself in the last point, section 3.4.

3.1 Introduction to intelligent data analysis

The data that is being collected continuously from EUREX and MONEP option markets is a matter of high importance, since it's been systematically archived in order to be rescued in case of any problem, security, fiscal or organisational auditing. However, do we know what is being saved? Business data from customers, contracts, market and technical indexes are daily saved into databases where there have been accessed less often, not only to make it faster and have more efficient systems, but also to be able to safeguard business information. It is necessary to understand what kind of information is saved that the institution could use to obtain more knowledge (McCormick, Khabaza, Abbott, Mutchler, & Brown, 2013).

There are many ways to get information. However, in a globalized world as ours and with large quantity of data to be stored permanently, it is easily understood that it often gets difficult to know where to start. Data science is the subject that over the past few years has had significant advances in knowledge extraction content. Expressions as "Data Science" and "Big Data" proliferate in the scientific community, in online forums and specialised courses, but what they mean? The term "science" means knowledge acquired through a systematic study (Dhar, 2013). Vasant Dhar, professor at the School of Business of New York, also says that a data scientist must have knowledge of several areas: mathematics, machine learning, artificial intelligence, statistics, databases and enhancements as well as good knowledge to formulate solutions. The methodologies of data science hunches up over many domains, and strongly influences the areas of economy, business and finance (Bloomberg L.P., 2014).

There are several techniques derived by DM, according to the type of data to be analysed and treat: DM is focused to evaluate structured information, such as relational databases, and similar structure tables; Text Mining proposes to analyse integration of competitive intelligence (Gary Miner, 2014), a new and emerging area of study that include a large number of activities, including the DM and Data Analysis on all kinds of text and multimedia data, whilst the Web Mining analyses information from the internet.

3.2 Methods and techniques

The Knowledge Discovery in Database (KDD) term first emerged by Fayyad et al. in 1996 with particular emphasis on knowledge discovery, but easily verified that the DM and KDD are convergent concepts as they share the goal of discovering new knowledge with the crossing of information. Fayaad et al. considered that DM mainly consists of the means to which the patterns are extracted and enumerated from data. Of course, "loose" data by relational tables or relating two variables that have no relationship isn’t a proper way to create knowledge, and may even mislead draw lessons from this kind of assumptions at random (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Fayyad et al. also encourages these stages are carried out in an automatic way, since analysing data manually proves to be expensive and not without opinions. Figure 12 shows an overview of the KDD transformation areas in the authors' perspective: selection, pre-processing, processing, data mining and interpretation / evaluation. The selection phase aims to create a data set that focus on a particular set of data from the database (Data), calling it the Target Date. This endures a treatment called pre-processing, which aims to clean up the target data to more consistency and above all not to emphasize the poor quality of data, and then it will treat the phase transformation. Here, transform the data reducing their size through specific functions before reaching the standards of research phase (data mining phase). Only after the data is standardized that they are really able to make their interpretation and evaluation, last phase of the process. This process is interactive and iterative, varying depending on each state the decision to go back or forward through the stages.

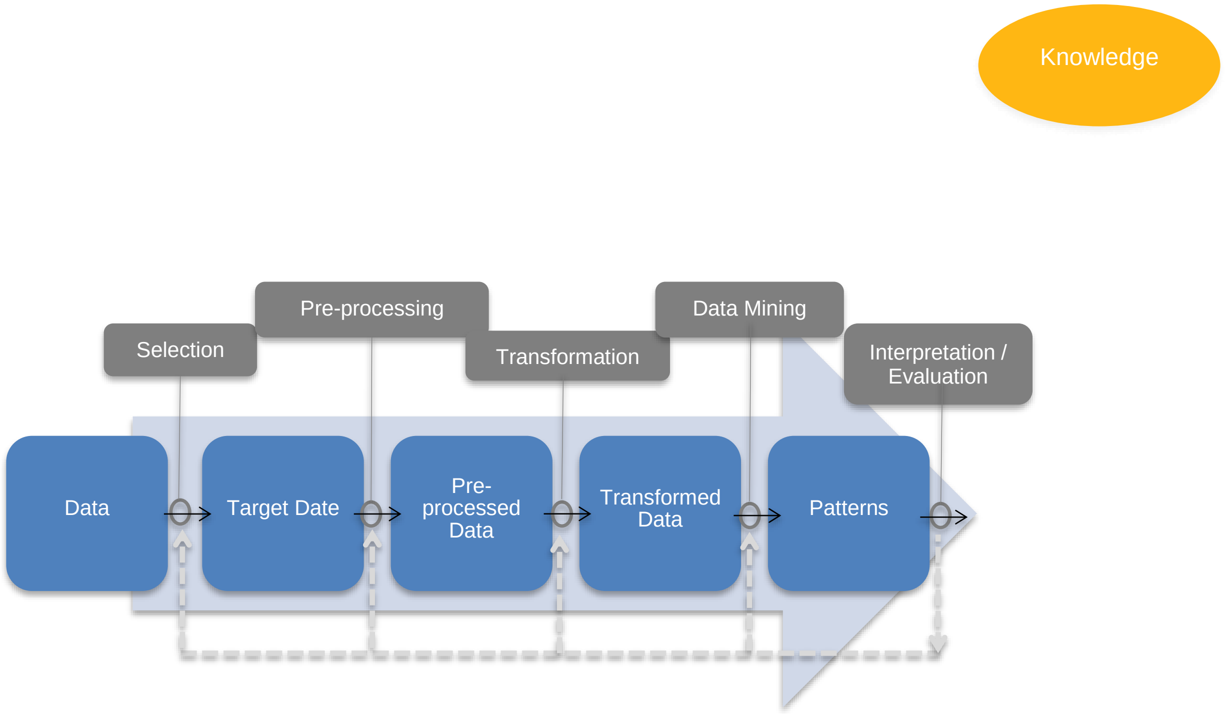


Figure 12: Knowledge discovery process steps overview

Adapted from (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)

Figure 12 illustrates the KDD phases showing that DM is a particular step of the KDD process. However, the scientific and business communities have adopted KDD and DM as synonyms over the years (Dhar, 2013) (Benoit, 2002).

The type of attributes in Data Mining has a simple taxonomy, as Figure 13 shows. Data can be categorical (e.g. the attribute “market” which has the value EUREX or MONEP), or numerical (e.g. the attribute “quantity”). “Market” is a categorical nominal attribute type because it hasn’t any particular order otherwise it would be a categorical ordinal type. “Quantity” is a numerical ratio attribute type, where zero means the absence of the attribute. The objective of this study is categorized as a classification problem because de dependent variable takes a finite set of values as outcomes, which is the decision possibility of two results to classify. For this type of problem, decision trees (DT) can help. DTs are widely used methods in DM. Subdivided in classification and regression analysis trees it is the most common and easy way to relate numerical and categorical variables having highly visual outcome and easy to interpret tree result (Chen, 2016).

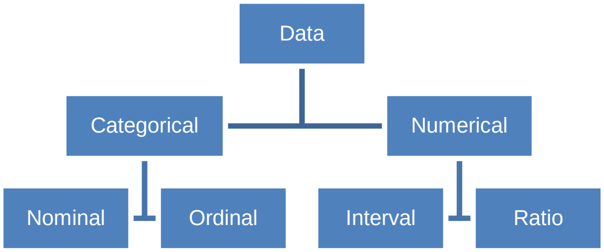


Figure 13: Simple taxonomy all types of attributes in Data Mining

Adapted from (Moro , 2011)

A DT is a predictive tree model, a parent-child typology tree, where each node represents a question and each child step from that node represents a possible answer to that question. Figure 14 illustrates the decision DT concept according to Suduan Chen (Chen, 2016). Indeed, the path from the root node through the internal branches having different weights to the selected leaf node can have different possible outcomes, being the simplest inductive learning method.

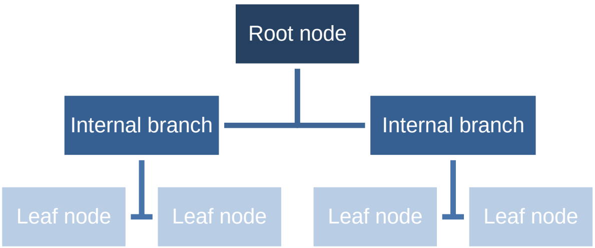


Figure 14: Illustration of the decision tree

Adapted from (Chen, 2016)

If the predicate that qualifies the root is true, then the left internal branch is the next predicate to evaluate; if not, the next predicate to evaluate is the right side of the tree, and this logic continuous until the leaf is reached. There are several algorithms that use the DT logic in order to detect the best accuracy for a problem and each of them like its specificity, with no specific order and as will be just briefly introduced below, the most common are C&RT, CHAID, C5.0 and QUEST (Dursun, Cemil, & Ali, 2013). It is therefore pertinent to use differentiated algorithms in the data extracted from the Organisation test them, understand their results and chose the best for the current problem. So lets characterize each of them.

The C&RT (Classification & Regression Tree) was introduced by Breiman et. al (1984) and it is a binary DT with the capacity to deal with continuous or categorical targets. Despite its complexity it can be extremely powerful and it runs fast, with the main idea of using binary trees to minimize the error in each leaf. It will be the first to be tested due to its simplicity: the tree grows by repeatedly repeating the algorithm. It works recursively by data splitting in to two subsets in order to enhance homogeneity, and it only stops when no more heterogeneous groups exist or when the stopping criteria is achieved. It has *ensemble* methods or techniques, for instance the bagging (or bootstrap aggregating) (Breiman, 1994), which consists in building multiple decision trees and repeats resample with training data replacing it, and voting the trees for an undeniable prediction. Predictors may be used several times throughout the DT, and in the end C&RT determines the best attributes for the best threshold maximizing the homogeneity.

With a strong predecessor as C4.5 and ID3 (Iterative Dichotomiser 3), the C5.0 algorithm created at 1993 by Quilan has a more efficient memory that creates smaller and faster DTs, having a more objectivity and precision results. C5.0 performs automatic boosting while building the tree: this adds more accuracy to the classification. The training is partitioned in more groups based on the output of single attribute test value. This test is then chosen, by giving a close to optimal partitioning. Additionally with the use of pre-pruning and post-pruning to establish DTs from top level, doing this procedure repeatedly until there is only one homogenous group so cannot partition more or the maximum depth is reached.

Created by Gordon Kass (1980), the CHi-squared Automatic Interaction Detector as known as CHAID, is one of the most popular supervised tree and an algorithm that uses three steps: merging, splitting and stopping. Based on adjusted significance testing, with main use for segmentation and tree growing, treats missing values automatically, it is a versatile algorithm: for prediction, classification, regression and even interaction between with the previous mentioned. CHAID algorithm is highly visual because it uses by default multi-way split, and is unique from other DTs because it can produce more two categories at any level of the tree, therefore it creates a big and not binary tree.

Finally the QUEST algorithm (Wei-Yin Loh, 1997), similar to the C&RT but relatively more recent, has however some differences. QUEST uses imputation (for missing values) instead of surrogate split and can handle categorical attributes with many categories. It deals with two different types of split selection separately and produces extensive trees but then applies automatic cost-complexity pruning to minimize its size (IBM, 2016). When Chen studied the detection of fraudulent financial statements he realized the CHAID-C&RT model was the most effective with 87.97% of accuracy, and in its literature state found that the most commonly used DT algorithms are C&RT, CHAID and C5.0.

3.3 Choosing the right methodology

Data science continues to be an emerging and interdisciplinary scientific area that allows and provides knowledge from given pre-existing information, allowing users to give knowledge through a well-defined methodology (IRMA, 2013). It involves six stages, the first four of description of the data, and the remaining two for the forecast with the computational process to discover patterns in data sets. According to

KDNuggets[2], the most popular DM methodologies that have been adopted by the data science community are the SEMMA and the CRISP-DM, not only given by its use but also because many of the data scientists have their own methodology or the company where their located, hence their choice in this analysis (KDNuggets, 2014). Both methodologies have grown in an industrial environment and are composed by a sequential set of steps guiding the deployment of new knowledge through the DM applications.

The SEMMA (Sample, Explore, Modify, Model, and Assessment) was developed by SAS Institute and refers mainly to the project DM driving process. It has the following phases: Sampling - take a sample that is large enough to represent the data by its relevance as its proportion, and possible to manipulate without problems; Explore - explore the data by checking what kind of anomalies, trends, ideas, and realize the quality of data; Modify - at this stage, the data is modified by creating and selecting which variables, and that transformation should be performed; Model - is the process of finding combinations of data to predict outcomes through a specialized software (closely related to *SAS Enterprise Miner software*); Assess - the final stage of the iteration where an assessment is made and see how it went, confirming the results.

The CRISP-DM (Cross-Industry Standard Process - Data Mining) is also an interactive and iterative process, having one more step. In addition, its 6 phases have a more comprehensive nature with business vision. It is also why it became a more consensual approach between the scientific communities and in business applications. Table 4 summarises the phases of the CRISP-DM:

Table 4: CRISP-DM stages and their description, as defined by (Chapman et. al, 2000)

CRISP-DM Steps	*Description*
1. Business Understanding	This stage is particularly important to define the objectives of the problem, the requirements in the business point of view and convert this knowledge to a data-mining problem.
2. Data Understanding	Combines the data collection procedures and activities. The aim is to become familiar with the possible complexity of the data and understand them in order to identify data quality problems, discover the focus of attention or detect subsets to create hypotheses to justify the missing information (hidden among the data).
3. Data Preparation	Encompasses all construction activities of the final dataset[3].
4. Modeling	Apply several data modeling techniques, calibrating the values and giving attention to the parameterization.
5. Evaluation	Evaluation of models, review the steps of the revised models to achieve the objectives.

[2] KDnuggets (KD stands for Knowledge Discovery) is a leading site on Data Mining and Data Science in general. Its president is Gregory Shapiro, a data scientist that among others is co-founder of KDD conferences (Piatetsky-Shapiro, 1997).

[3] Dataset - data set and data repository organised in a file that can have various formats: text, excel, csv, sql, among others.

CRISP-DM Steps	***Description***
6. Deployment	The creation of the model is not the end of the project. Even if the purpose of the project is to increase knowledge of the data, the knowledge gained has to be organized and presented in a manner that it can be used (Chapman et. al, 2000).

Objectively, the modelling and evaluation phases are in fact the main objective of this study: to acquire more knowledge about the derivative options on their date of maturity. The phases are not rigid and at the end of each phase it is always possible to make adjustments before moving to the next phase, even if it is necessary to go back. Making a comparison of the three methodologies KDD, SEMMA and CRISP-DM at first, the approaches have similarities, especially between KDD and SEMMA, even if their only similarity is the fact that they have a common creation / near the *SAS application Enterprise Miner*. However, compared to the KDD model with the CRISP-DM methodology, the similarities are not so evident. Cross Industry Standard Process for Data Mining, commonly known by its acronym CRISP-DM, was a data mining process model that describes commonly used approaches that data mining experts use to tackle problems. Figure 15 shows the 6 phases in CRISP-DM and its association with the package *rminer* from R tool.

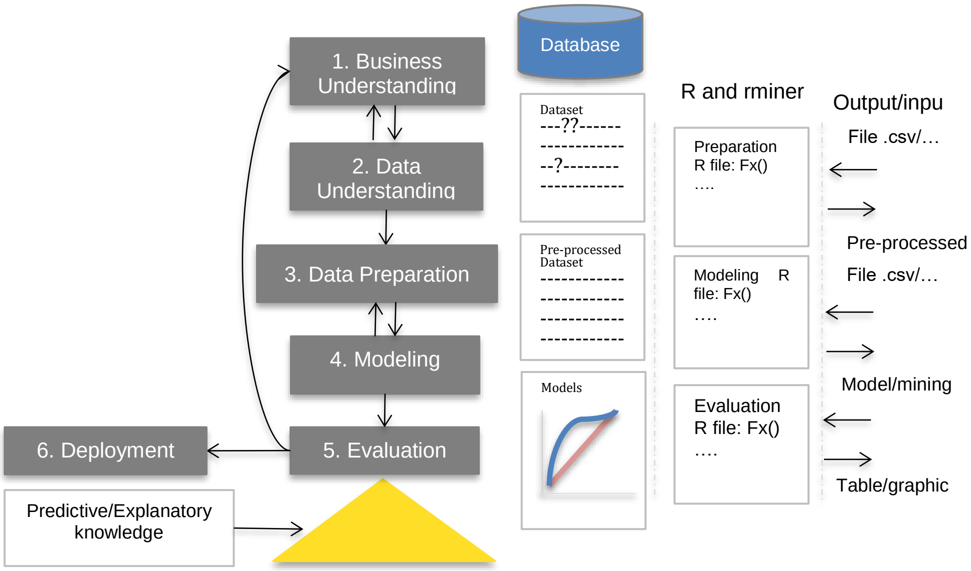


Figure 15: CRISP_DM methodology and suggest *rminer* package use, adapted from (Cortez, 2015)

Conceived in 1996, CRISP-DM is a robust and well-proven methodology being widely accepted by not only IBM but also the scientific community in general and used for projects in SPSS Modeler.

The KDNuggets website showed an interesting picture putting the evidence of each approach side by side and drew several conclusions (Table 5).

Table 5 - Comparison between KDD, SEMMA and CRISP-DM methodologies

KDD	***SEMMA***	***CRISP-DM***
Pre-KDD	-	Business Understanding
Selection	Sample	Data Understanding
Pre-processing	Explore	Data Understanding
Transformation	Modify	Data Preparation
Data Mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post-KDD	-	Deployment

Adapted by (KDNuggets, 2014)

After a closer look, we find that the SEMMA and CRISP-DM approaches can be followed in the implementation of DM problem and are both implementations of the proposed model by Fayyad et al. However, the CRISP-DM is more complete than the SEMMA and is more directed to the business, in addition to being widely accepted by the scientific community. For these reasons it was decided to follow the CRISP-DM methodology.

Recently, IBM Analytics Services have released a new implementation method for Data Mining projects called ASUM-DM. Its acronym stands for Analytics Solutions Unified Method for Data Mining/Predictive Analytics and is a refined and extended CRISP-DM launched in October 2015 (IBM, 2015). This model came as an alternative of the failed attempt to update CRISP-DM (between 2006 and 2008) but no success. As a results, the CRISP-DM.org website is no longer available. ASUM-DM ha the same steps as the CRISP-DM but have more focus on the evaluation and implementation. It is not considered the best candidate for the methodology of this research because the objective is to have a bigger focus on the first 3 stages of the CRISP-DM, rather then the evaluation and implementation.

3.4 CRISP-DM – iteration 1

Assumptions

The CRISP-DM is an iterative and an evolutionary process methodology. In order to be less confusing for both work and description of this document, the iterations will be as few as possible – ideally just one – and if necessary doing more advances and going backs from one step to another. The methodology will be followed with the guidelines of the CRISP-DM (Chapman et. al, 2000).

3.4.1 Business understanding

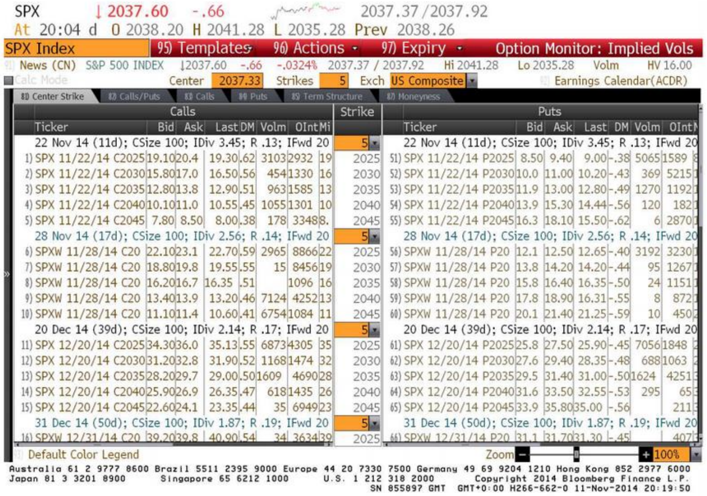
The Business Understanding is the first step of the CRISP-DM methodology. Here, the goal is to capture in business perspective about what the Organisation really wants to know. For this purpose, it is important to uncover crucial facts right from the beginning than wait for the further steps. "A possible consequence of neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions" (Chapman et. al, 2000, p. 16).

The universe of this project is the option derivatives product itself, bought or sold in the Organisation and with a future expiry date. A sample of this universe is a subset that must have a direct relation with the study, with the perspective of the operational and IT support teams. It is a non-probabilistic convenience sampling type, which is when the researcher only has access to part of the population that should use (Quivy, 1992). As mentioned in the introduction of this document, the objective proposed for this research is the discovery of patterns that translates into useful and applicable knowledge by predicting the probability of exercise in a particular financial option at the maturity. In terms of business success, the result of this study could be an additional point in favour to "tender" the decision makers in enhance IT data storage for a future better data quality. The search will start with three different studies: **exploratory**, **descriptive** and **explanatory**.

Exploratory studies: Data collecting

In this point, the question is what relevant information should be gathered to create a dataset? Internal and external factors, markets, timeline data, among others must be analysed and consider. After some research to collect quality data, it was possible to learn more about relevant information in this area by the use of a Bloomberg terminal provided by the ISCTE academic library.

Figure 16 shows a print screen of the Strike Option Monitor, for implied volumes of S&P 500 index, by the help desk online of the Bloomberg terminal (it is a software system by Bloomberg L.P. that provides messaging, security data, analytics and news to more than 300,000 professional market participants around the world).



SPX ↓2037.60 -.66 2037.37 / 2037.92
At 20:04 d O 2038.20 H 2041.28 L 2035.28 Prev 2038.26

News (CN) S&P 500 INDEX ↓2037.60 -.66 -.0324% 2037.37 / 2037.92 Hi 2041.28 Lo 2035.28 Volm HV 16.00

Calc Mode Center 2037.33 Strikes 5 Exch US Composite Earnings Calendar(ACDR)

Calls								Strike	Puts						
Ticker	Bid	Ask	Last	DM	Volm	OInt	Mi		Ticker	Bid	Ask	Last	DM	Volm	OInt
22 Nov 14 (11d); CSize 100; IDiv 3.45; R .13; IFwd 20								5	22 Nov 14 (11d); CSize 100; IDiv 3.45; R .13; IFwd 20						
1) SPX 11/22/14 C2025	19.10	20.4	19.30	.62	3103	2932	19	2025	51) SPX 11/22/14 P2025	8.50	9.40	9.00	-.38	5065	1589
2) SPX 11/22/14 C2030	15.80	17.0	16.50	.56	454	1330	16	2030	52) SPX 11/22/14 P2030	10.0	11.00	10.20	-.43	369	5215
3) SPX 11/22/14 C2035	12.80	13.8	12.90	.51	963	1585	13	2035	53) SPX 11/22/14 P2035	11.9	13.00	12.80	-.49	1270	1192
4) SPX 11/22/14 C2040	10.10	11.0	10.55	.45	1055	1301	10	2040	54) SPX 11/22/14 P2040	13.9	15.30	14.44	-.56	120	182
5) SPX 11/22/14 C2045	7.80	8.50	8.00	.38	178	3348	8.	2045	55) SPX 11/22/14 P2045	16.3	18.10	15.50	-.62	6	2870
28 Nov 14 (17d); CSize 100; IDiv 2.56; R .14; IFwd 20								5	28 Nov 14 (17d); CSize 100; IDiv 2.56; R .14; IFwd 20						
6) SPXW 11/28/14 C20	22.10	23.1	22.70	.59	2965	8866	22	2025	56) SPXW 11/28/14 P20	12.1	12.50	12.65	-.40	3192	3230
7) SPXW 11/28/14 C20	18.80	19.8	19.55	.55	15	8456	19	2030	57) SPXW 11/28/14 P20	13.8	14.20	14.20	-.44	95	1267
8) SPXW 11/28/14 C20	16.20	16.7	16.35	.51		1096	16	2035	58) SPXW 11/28/14 P20	15.8	16.40	16.35	-.50	24	1151
9) SPXW 11/28/14 C20	13.40	13.9	13.20	.46	7124	4252	13	2040	59) SPXW 11/28/14 P20	17.8	18.90	16.31	-.55	8	872
10) SPXW 11/28/14 C20	11.10	11.4	10.60	.41	6754	1084	11	2045	60) SPXW 11/28/14 P20	20.1	21.40	21.25	-.59	10	450
20 Dec 14 (39d); CSize 100; IDiv 2.14; R .17; IFwd 20								5	20 Dec 14 (39d); CSize 100; IDiv 2.14; R .17; IFwd 20						
11) SPX 12/20/14 C2025	34.30	36.0	35.13	.55	6873	4305	35	2025	61) SPX 12/20/14 P2025	25.8	27.50	25.90	-.45	7056	1848
12) SPX 12/20/14 C2030	31.20	32.8	31.90	.52	1168	1474	32	2030	62) SPX 12/20/14 P2030	27.6	29.40	28.35	-.48	688	1063
13) SPX 12/20/14 C2035	28.20	29.7	29.00	.50	1609	4690	28	2035	63) SPX 12/20/14 P2035	29.5	31.40	31.00	-.50	1624	4251
14) SPX 12/20/14 C2040	25.90	26.9	26.35	.47	618	1435	26	2040	64) SPX 12/20/14 P2040	31.6	33.50	32.55	-.53	295	65
15) SPX 12/20/14 C2045	22.60	24.1	23.35	.44	35	6949	23	2045	65) SPX 12/20/14 P2045	33.9	35.80	35.00	-.56		211
31 Dec 14 (50d); CSize 100; IDiv 1.87; R .19; IFwd 20								5	31 Dec 14 (50d); CSize 100; IDiv 1.87; R .19; IFwd 20						
16) SPXW 12/31/14 C20	39.20	39.8	40.90	.54	34	3634	39	2025	66) SPXW 12/31/14 P20	31.1	31.70	31.30	-.45		407

Australia 61 2 9777 8600 Brazil 5511 2395 9000 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2014 Bloomberg Finance L.P.
SN 855897 GMT GMT+0:00 H266-662-0 11-Nov-2014 20:19:50

Figure 16: Print screen of the Bloomberg terminal provided by the ISCTE academic library

Adapted from Bloomberg helpdesk online (Bloomberg Finance L.P., 2016)

After accessing to this helpful tool, it was possible to build a datasheet with the latest ratings of companies in the S&P 500[4], i.e., ascertain whether a possible rise or drop rate of concerned companies holding an option is relevant to the problem (tables 6 and 7 below). Tables are related each other by the ticker of the equity with its underlying product.

Table 6 – Example of Equities rates updated by an Excel add-in of Bloomberg[5], at 12th Nov. 2014

Ticker	*Exchange code that it trades on*	**Rate of Standard & Poor's 500**
A UN	A UN Equity	BBB+
AA UN	AA UN Equity	BBB-
AAPL UW	AAPL UW Equity	AA+
ABBV UN	ABBV UN Equity	A
ABC UN	ABC UN Equity	A-
ABT UN	ABT UN Equity	A+
ACE UN	ACE UN Equity	#N/A

Adapted from Excel add-in with the help of helpdesk online of Bloomberg

According to an exploratory conversation with an experienced operational of the Organisation, this rating is commonly considerate by traders and it figures one of the most important factors for client decision making in general. However, the datasheet can only be updated when a new physical access to the terminal is done, limiting the refresh of this rating. Despite that Bloomberg offers a service online to update outside the ISCTE library (Bloomberg Finance L.P., 2016), the historic data in Bloomberg is available for few months, and this rating could only be stored from the current date onwards.

Table 7 - Correspondent associated Equity, company and type of commodity at 12th Nov. 2014

Underlying Ticker	*Company Name*	*Currency*	*Exchange*	*Type*
IBM US Equity	International Business Machines Corp	USD	US	Common Stock
AAPL US Equity	Apple Inc. (U.S.)	USD	US	Common Stock
BABA US Equity	Alibaba Group Holding Ltd	USD	US	ADR

[4] The S&P 500 (Standard and Poor's 500) is one of the best single gauge of large-cap American equities, which its index includes 500 leading companies and captures approximately 80% coverage of available market capitalization (S&P Down Johns indexes, 2014).

[5] Excel add-in of Bloomberg: this add-in was can include special Bloomberg functions like BDP (Security, field) used in the populated tables. It was combined the security (ticker/exchange combination) with fields (mnemonics with information to download) (Vienna University of Economics and Business, 2009).

Underlying Ticker	***Company Name***	***Currency***	***Exchange***	***Type***
FB US Equity	Facebook Inc.	USD	US	Common Stock
AMZN US Equity	Amazon.com Inc.	USD	US	Common Stock
SPY US Equity	SPDR S&P 500 ETF Trust	USD	US	ETP
ARCP US Equity	American Realty Capital Properties Inc.	USD	US	REIT
VOD LN Equity	Vodafone Group PLC	GBP	LN	Common Stock
TWTR US Equity	Twitter Inc. USD	USD	US	Common Stock

Adapted from Excel add-in with the help of helpdesk online of Bloomberg

Considering this, the approach is not viable and was dropped, because the data stored in database are from the past and the goal is to have a dataset with thousands of records, if possible from three years in the past (from 2013 to 2015). The data collection was done before the interviews, as these interviews required a special authorisation and therefore all the process was delayed. The importation of data in test environment has been traduced in a long process due to security and policy constraints in retrieving it to outside the institution. The relevant database tables retrieved are records from:

- Current trades, orders and clearing;
- Historic trades, orders and clearing;
- Information detailed of trades and options (contributed changes for the cleared state (code: 101);
- Audit trail and its historic tables with technical information;
- Markets where as well present in tables clearing tool: the European markets as EUREX, MONEP, ICE, OMX, MEFF, IDEM, SAFEX; the American markets as KOP: CME, OCC, ME, CBOT and the Asian markets as SGX, SFE, and NZFOE.

Descriptive studies: characterisation of option data

As all this information was extensive (historic data could start from 2009 at the majority of markets, and each day has thousands of records of trades saved in the database), a *brainstorming* meeting with the team leader of IT support took place in front of Oracle and SQL databases. After having analysed together the quality of that data existing in the databases, it was concluded that narrowing the focus to the expiry date for **EUREX** and **MONEP** markets could be the best choice to have a consistent result. EUREX because it is the more differentiated market, having 665 different types of options contracts, 1538 types of future contracts (Eurex, 2014) in terms of product and have more specifics. Furthermore, this data has the particularity of possessing expiry date per week (non-standard options). MONEP (NYSE Euronext Paris, Amsterdam, Lisbon and Brussels) - for being the oldest market of the Organisation, the 2nd largest in Europe after London stock exchange (NYSE, 2014), as well as the best known by the bank and which Portugal is also part of. American-style markets weren't chosen because, in the case of options and as described in the literature review, the client cans exercise **up to** the expiration date and not **only** on the expiration date as with the rest of the world, which multiplies the perspectives and the speculation of each characteristic of the option.

Explanatory studies: the business focus

Finally, the explanatory studies explain the phenomena. In this context interviews should be introduced to enforce and explain a given directive agreed by the management of the teams. Interviews are oral conversations preferably face-to-face with people selected carefully. This way of getting information has to be very well thought out and has to meet the criteria of conduct and drafting in order to not influence the interviewed (e.g.: consciousness, culture, among others) (Quivy, 1992). It is therefore crucial to interview the team leader of the IT Support team in a technical point of view but as well to interview an experienced business player (trader/investor), to be able to understand the financial perspective. Initially, a third interview was upon the table dedicated to the human resources office but the truth is that the information that could be retrieved wouldn't be an added value for the research (as it would be closed to the trader's feedback). The interviews will be semi-structured with a script. The formulation of the questions and rules of conduct shall comply with the investigation manual of Quivy (Quivy, 1992). The questions will focus mainly on the experience of IT offices, the motivation of investors to exercise an option in the maturity period, the efficiency of these offices, and the current way of working and so regarding the operational team, the questions are more technical.

Tools

There is a wide range of intelligent data science tools currently available. Some of them are growing slowly but consistently (R language) and others like Python has been growing very fast (KDNuggets, 2016). R project continues to be the language preferred of the scientific community and not only among IT scientists because it is intuitive and is an open source (cran.rstudio.com). Inclusive, R is being adopted by major OS market leaders thanks to its instructions in R embedded in their tools for business intelligence and database technologies (Oracle, 2016). The R is an open source and multi-platform tool that can be downloaded from the official web site: http://www.r-project.org/. The ***rminer*** package can be easily installed by opening the R program and using its package installation menus or by typing the command after the prompt: `library(rminer)`. Only one package installation is needed per a particular R version. The *rminer* package has been adopted by users with distinct expertise domains and in a wide range of applications. From 2th May of 2013 to 20th May of 2015, the package has been downloaded 10,439 times from the RStudio CRAN servers (cran.rstudio.com). The *rminer* package is been used by both IT and non-IT users, e.g., managers, biologists or civil engineers (Cortez, 2015).

IBM has been linked to data mining since the 90s and in result the CRISP-DM approach was created to help projects achieving their goals. The SPSS Modeler, a tool created by IBM in 2009 predecessor of Clementine (Gersten, Wirth, & Arndt, 2013) and has been gained some ground the last years, is an application for data mining and text analytics. However not an open source software, it is used to predictive models

(McCormick, Khabaza, Abbott, Mutchler, & Brown, 2013). Besides its large history towards CRISP-DM, this tool was chosen a well mainly because of extended width of techniques, its process support, and its scripting facilities using a visual programming interface (Gersten, Wirth, & Arndt, 2013). Finally the tool used not only to observe at first sight but as well a constant check during the all process on analysis of the dataset created is the MS Office Excel. Some functions available were used and the filters were handled to uncover and discover knowledge by for instance just combining two different variables.

In this study, the first 3 phases (business understanding, data understanding and data preparation) were created with the help of R (*rminer*) and observations in excel/csv formats. The last 2 phases (modeling, evaluation) were done with the help of IBM SPSS Modeler.

Ratio cost benefit, risks and contingency plan

If the project is succeeded, this means that operational teams could have a prediction of the probability of exercise in a particular financial option at the maturity date. Depending on the period between forecast result and the event happening, some actions of preparation could be crucial in order to have a more controlled and expected behaviour. And this information is relevant allowing potential costs reduction in resources and enhancing efficiency in the process chain as soon as the markets closes. There are risks that need to be identified, studied and if possible mitigated. However, in certain cases, the risk cannot just disappear so a contingency plan must be put in practice in order to reduce as much as possible the damaged or just accept by describing in detail the most accurate response. The table 9 resume the risks and contingency plan found during the current phase.

Table 8 below lists the risks and contingency plans found during the business understanding phase. It is important to understand that this exercise should be done before further stages to make sure that the details does not affect the overall. The column "cat." represents the categories of risks related to the area affected by the risk (e.g. Business, IT, People & Organisation, External and Legal). The 4th column does a description of the risk that may occur in the project and its causes. What kind of problems will the risk result in and risk dependencies. The risks assessment columns are composed by "likelihood" (L), "impact" (I) and "risk response" (RL). The likelihood is a numeric value denoting the estimate of the probability that the risk will occur, and the impact is a numeric value denoting the severity of the impact of the risk (should it occur). The risk response is the product of the likelihood with the impact values (RL=L*I). The "risk response strategy" to deal with the identified risks are: Avoid (risk avoidance, working the project or project plan around those conditions or activities which introduce the risk); Reduce (risk mitigation or reduction through the proactive implementation of risk reduction activities); Accept (acceptance of the risk. In this cases, contingency plans can also be defined in case the risk occurs); Transfer/Share (transfer or share a risk with other entities, e.g. through insurances, sub-contracting etc.). Finally the "action details" traduces the reason of the risk and a contingency plan or call to action.

Table 8: Risks and contingency points found during the business understanding phase

Risk Identification and Description				***Risk Assessment***			***Risk Response***	
ID	***Cat.***	***Risk name***	***Risk description & details***	***Likelihood (L)***	***Impact (I)***	***Risk response (RL=L*I)***	***Risk Response Strategy***	***Action Details***
1	Business	Results impacts due to missing ratting	Observed during the business understanding stage: tables provided by Bloomberg don't suffice to be included in this study.	5	5	25	Accept	Known issue and one of the most alarmed risks. It must be taken into account during the discussion of results.
2	Business	Lack of attribute definition for daily financial news	Behaviour of stocks in markets, financial health could be important factors to include in the dataset, but too difficult to obtain in a reasonable time for this study.	2	5	10	Accept	It is a known issue and must be taken into account during the discussion of results, after the evaluation stage.
3	Business	Some attributes are empty	Excluded attributes, as they are not fulfilled. Reasons: not specific instructions of Operational teams that impact both manual and automatic processes.	3	2	10	Avoid	These attributes had to be removed, as the populated techniques could influence the result.
4	Business	Attributes aren't correctly introduced	The majority of attributes seem compliant with the expected, but others are identified as outliers.	1	2	2	Reduce	Techniques to remove outliers will be applied in the stage Data Preparation.
5	IT	Some attributes are empty	Some attributes were excluded, as not fulfilled due to not specific instructions of IT teams and application tools, affecting both manual and automatic processes.	5	5	25	Accept	These attributes had to be excluded as could influence the result. They must be taken into account during the results.
6	IT	Misunderstanding of business attributes	The meaning of certain business factors can be misleading during the interpretation of attributes correlation with others and the results	3	3	9	Reduce	Documentation, research in the life cycle of options and understanding each value is a constant concern of the researcher.
7	IT	Some tables weren't investigated	Among the big universe of databases tables, there is a chance to forget/misleading fields	1	5	5	Reduce	All databases present in the DBMS of the Organisation were scanned, however done before the interviews
8	IT	Misleading in anonymisation	There is a chance that the database selected for anonymisation could have saving data problems.	1	5	5	Accept	Organisation's responsibility. This problem doesn't often occur, though.

There are two risks that stood out from the others, with the major impact. The first business risk "results impacts due to missing rating" has the highest rate (RL=20) and it is a risk to be accepted. This means that the missing rating data related with the financial health of the most direct company's underlying asset is the major risk that this study has, as it is one of the most important factor to be taken into account for options owners, in order to decide rather to choose exercise or not. After several attempts to get this data without success is it an accepted risk, known issue and one of the most alarmed risks. It must be taken into account during the discussion of results. The risk nº5 is an IT risk named as "some attributes are empty" with a RL of 15. This risk alerts by the fact that some attributes haven't data and therefore are excluded from the dataset, as not fulfilled due to not specific instructions of IT teams and application tools, affecting both manual and automatic processes. As the risk nº1, the risk is accepted and the attributes had to be excluded to not influence the results, however they must be taken into account to understand the results.

Business understanding results

As experienced professionals and the researcher herself confirm, the Organisation could benefit of this work: identifying possible causes of abnormal volume; a detection of problems during rush hours; after the inventory of resources, update the contingency plan by saving some resources (servers) and teams (help desk, IT support, and human resources (operational teams) and support for new planning and more grounded decisions. The process around the maturity date in some teams could be more efficient if the objective proposed could be accomplished. As well, other results could trigger other recommendations. Internal factors like behavioural in the operational teams geographically dispersed by several countries of the world and are acting under different cultures and experiences cannot be collected and somehow evaluated, but other internal factors like successive go-live of new developments in a sort period of time and problems with servers needs to be taken into account.

Table 9 - Top 10 most popular tools in 2016 poll

Tool	2016 % Share	% Change	% Alone
R	49%	+4.5%	1.4%
Python	45.8%	+51%	0.1%
SQL	35.5%	+15%	0%
Excel	33.6%	+47%	0.2%
RapidMiner	32.6%	+3.5%	11.7%
Hadoop	22.1%	+20%	0%
Spark	21.6%	+91%	0.2%
Tableau	18.5%	+49%	0.2%
KNIME	18.0%	-10%	4.4%
scikit-learn	17.2%	+107%	0%

Adapted from (KDNuggets, 2016)[6]

[6] Note: The participation by region was: US/Canada (40%), Europe (39%), Asia (9.4%), Latin America (5.8%), Africa/MidEast (2.9%), Australia/NZ (2.2%).

The external factors are the external regulators for instance EMIR (European Commission, 2015), holidays where some markets are closed and trades suspension are the most important to note. As table 9 shows, the R project for statistical computing has been popular and widely (KDNuggets, 2016); as a GUI tool, SPSS Modeler will be server during the last phases of the study. Finally to complete this step, the project plan is now narrowed to: European markets MONEP and EUREX, without the help of external factors like the companies' rates as initially expected and the historic trades and orders of cleared options will be collected in the Organisation (after its consent). The tools are Oracle SQL developer, MS Excel sheets, and the analysis and further steps with statistical tools R and SPSS Modeler.

3.4.2 Data understanding

Collect initial data

The data is collected from the Organisation databases, in test environment and with some indications from the experience of the researcher. These databases are frequently refreshed from the production environment and being prepared (authorisation, authentication, anonymisation among others). Table 9 illustrates the nature of the information, their application tools and databases, filters associated in the script and critical attributes for the study. Figure 17 illustrates of the different components of the listed derivatives of EUREX and MONEP are connected.

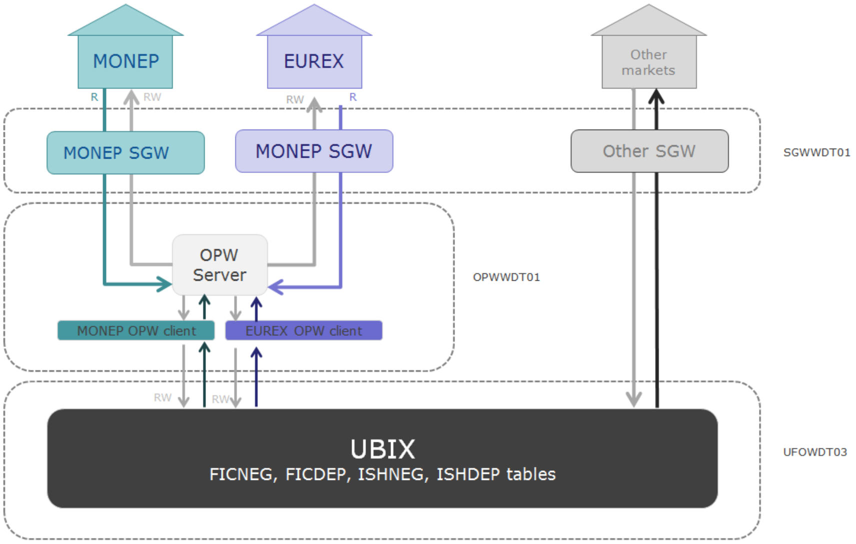


Figure 17: How different information systems are integrated in the Organisation

Adapted from (Organisation, 2013)

Each market has gateway (SGW) that integrates as read-only ("r" in the figure above) contracts to the option watch server (OPW server); their respective option watch client receives it then do changes/operations and interacts to the back-office (Ubix) by updating key tables, before at its turn sent back to the market with new information to be acknowledged ("rw" as read-write). Each of these three steps captures new information of three different databases: from the gateway (SGWWDT01), the option watch (OPWWDT01) and the Unix (UFOWDT03). Table 10 presents the provenance of the table collected in search of data quality for the study.

Table 10: Provenance of data collected

Tools	*Database*	*Filter*	*Tables*	*Attributes*
Stream Gateways	SGWWDD01	EUREX MONEP	Marketing messages Type =(AM, AW)(*)	Message code, market, data, contracts, and message content itself.
Option Watch	OPWWDD01	MONEP	Table of positions Table of errors	Maturity dates, market, contract type option, clearer member, status, last modification, exclusion errors, amounts in different states, internal codes
SRX	UFOWDD01	EUREX	Confirmations table	Maturity dates, market, contract type option, clearer member, status, last modification, exclusion errors, amounts in different states, internal codes
Ubix	UFOWDD01	EUREX MONEP	Current and historic Neg. table; current and historic dep. table; contacts table.	Price, strike price, quantity, direction of option (buy/sell), type of maturity, source codes, currency, operator, price; Detailed information of the underlying asset.

(*) AM, AW = Assignments and Exercise, respectively.

These tables came from different systems and therefore intermediated techniques are used in the database, in order to integrate the enumerated attributes. Then, the resulted query is executed and integrated month by month into an excel sheet, as the Organisation does not allow in bulk export system but also because the information is heavy. The integration of an attribute (or not) in the dataset is a process of analysis and as they are come from different databases it is an additional constraint either here or in a later data preparation phase (Chapman et. al, 2000).

Describe, explore and analyse attributes

Selected tables were evaluated and so their field/variables. Each of them were exhaustively analysed through the help of the R tool by using techniques like histogram, summary of statistics or just verified directly in the dataset, after converting it to a CSV file (CVS: comma-separated value). The appendix B shows more details like histogram charts and simple statistics get by R project of the 328.498 records. The next tables summarize data information and data quality assumption of each candidate attribute to become part of the dataset (or not). From the database extraction three type of dates related options stood out: the date of contract, the date of the maturity and the date of last modification, as describes Table 11 below.

Table 11: Dataset attributes and its respective analysis: dates

Attributes	*Analysis*
Date_C	Day of trading of the contract. It has the full absolute date format (DD/MM/YYYY HH:MM:SS but the time has value 00:00:00 in all rows), which is not convenient to be used by some tools. Days around the 19 are the most frequent, corresponding to the standard maturity (typically the 3rd Friday of the month), biggest months are each end of quarter (Mar, Jun, Sep, Dec), especially December and the best year is 2014, however the records are from mid 2012 until mid 2015.
Date_Mat	Standard maturity date of the contract (DAECA). As Date_C, it has the full absolute date format (DD/MM/YYYY HH:MM:SS but the time has value 00:00:00 in all rows), so it needs to be converted as relative format during the data preparation phase. It's visible that each end of quarter has a bigger activity, and despite a big end of year volume, the first half of the year is better (due to July and August poorer months).
	Non-standard maturity date of the contract (DAECB), it has the same purpose and same format as DAECA. Non-standard maturities can be per month, per week or even per day. This field is a candidate to be discarded, because it is a minority and it represents another type of option (Flex options, to be confirmed with the results of the interviews).
Date_Mod	The date of the last modification of a contract. It has the full absolute date format (DD/MM/YYYY HH:MM:SS) so has to be converted as relatives different variables in during the data preparation phase. It is curious that some values of this attribute correspond to the day after of the maturity date. This can be related with the last changes that affect the expiry/assignment or abandon events; sometimes these changes may take time to be done until one last confirmation by the operational team, as it needs the interaction of the operator, right after the market is closed. As well, some of these changes were performed by another time zone input (New Jersey (KOP), Lisbon and Paris (EMEA), Chennai, Mumbai and Singapore (APAC) are the cities where the Organisation has its teams). Analysing with the R function summary, the busiest hours of the day are early in the morning or late around 23h. As the European time zone reference for the Organisation is CET (Paris), it is possible that operational teams of APAC could help during early morning and KOP can help for later time and closings.

As seen by the above, the non-standard date could be removed from the dataset, and all dates should be unfolded into relative isolated numerical integer values to be easier to manipulate in R or in SPSS. Table 12 below describes the dependent variable (Var_Dep, also called "Exercise Y/N") as well as two new attributes created directly in the dataset (in Excel).

Table 12: Dataset attribute and its respective analysis: exercise/not exercise

Attributes	*Analysis*
Exercise Y/N	Variable dependent is a nominal categorical, which describes 3 possible values: exercise (E), abandon (B) and assignment (S). A simplified new attribute could be: exercised (Y) = 49,14%; not exercised (N) = 50,86%

Table 13 below shows the variables that help to characterize the option, its market, its clearinghouse and how it is integrated into the system. We can see that the intervenient code has too many different values and with the majority of them as "96152T" (113 394 records). Using the library *rminer* in R the attribute could have its number of levels reduced through the function `delevels`. However, after observing the results with a `plot` and even with a `threshold` of 100, the levels were still too high. The conclusion is that this attribute could be also discarded.

Table 13: Dataset attributes and its respective analysis: parties and other entities

Attributes	*Analysis*
Market	Market name are mostly MONEP (80,63%) than EUREX (19,37%).
Clearer	Legitimate entity to authorise the clearing (GCM[7]). Note: "MONEP" – the most common value – seems to be a value that is used by default, not representing a real clearer (MONEP represents a market and there is no clearer registered with a market name). Others like ARBIMULT and EUREXARB are the major values. The majority of manual operations are "MONEP". So, is this attribute relevant for the study? Not an interesting value to be used in the dataset due to its multiple values.
Cod_Orig	Integration origin code of the trade. With a proportion of 80,63% of OPTIONWATCH-O and 19,37% of SRX_EUREX, this code reflects the direction of the expiry is performed: the majority is through the Option Watch application, only used for MONEP market, which is correct as there is not process linked for the EUREX automated, yet. SRX_EUREX is the origin code for EUREX market.
User	Nominal categorical. Attribute that maps to a given trade is confirmed automatically or via a specialized manual operating. Not an interesting value to be used in the dataset due to its multiple values, so it can be discarded.
Cod_Interv	Intervener's trade code. This value can symbolise an individual or an entity and their values can be repeated. Biggest values are: 96152T (113394), 94923 (100512), BNAPAERX (52005), 99999 (26151) BFI.PARI (13063) and GSS (11362). As users codes have 5 and only digits and typically start with the digit 9, values "96152T" and "99999" seem to be outliers. BNAPAERX is a known entity and GSS is the technical name to designate EMEA (Europe, the Middle East and). Not an interesting value to be used in the dataset due to its multiple values.
Type_Op	Manual or electronic operation, most are "electro" (51,44%), other values are operational user codes.

Another important fact is that, filtering the market from MONEP and EUREX and looking to the dependent variable, it was found that all EUREX options were exercised. In the dataset during the modeling phase the market should be select uniquely by MONEP. Table 14 below describes the variables directly related with the option. Some codes seem to be shown too diversified; others like the Code_C identify the option and could be removed.

[7] GCM: Acronym described as a Global Clearing Member, in Annex A.

Table 14: Dataset attributes and its respective analysis: direct characteristics the option

Attributes	***Analysis***
Code_C	Contract code of contract, nominal categorical. The biggest values are OAEX (38928), PXA (37824), OESX (14853), and other minor codes, but there are still a lot of other differentiated values (206836). Is it a relevant for the study? Still in doubt if this attribute should be separated, or perhaps be clustered in smaller levels.
Call_Put	CALL (C) or PUT (P), nominal categorical. It's the direction of the operation, C (50,70%) P (49,30%).
Strike_Price	Option's rating during the moment of the acquisition. Despite of an average of 1104.5, the median is 368.0, because there are high strike prices (max: 13150.0). Are they outliers?
CSENS	V for "Vente" (sale) and A for "Achat" (buy). A new attribute could be created to facilitate the lecture: short for V and long A. Values are V (48,89%) and A (51,11%).
Quant_C	The quantity of a given option contract, with a median of 78 and a mean of 1455.
Price	The price of trade. According to its statistics, the price is lower than the strike price.
Price_val	Valued price of the trade. Unknown value and curiously bigger than the strike price but with a lower mean, and different than price too. Sometimes 10 times bigger than price, but it´s not always the case. Until now, it wasn't possible to know more, so still under investigation.
Type_Option	Nominal categorical Type of option: American (A) or European (E). As only the European options will be analysed, this attribute will be discarded.

Table 15 describes some characteristics around options. Abandon ATM seem not relevant to the study, as it just describes the consequence of the expiry event, after the occurrence.

Table 15: Dataset attributes and its respective analysis: characteristics around the option

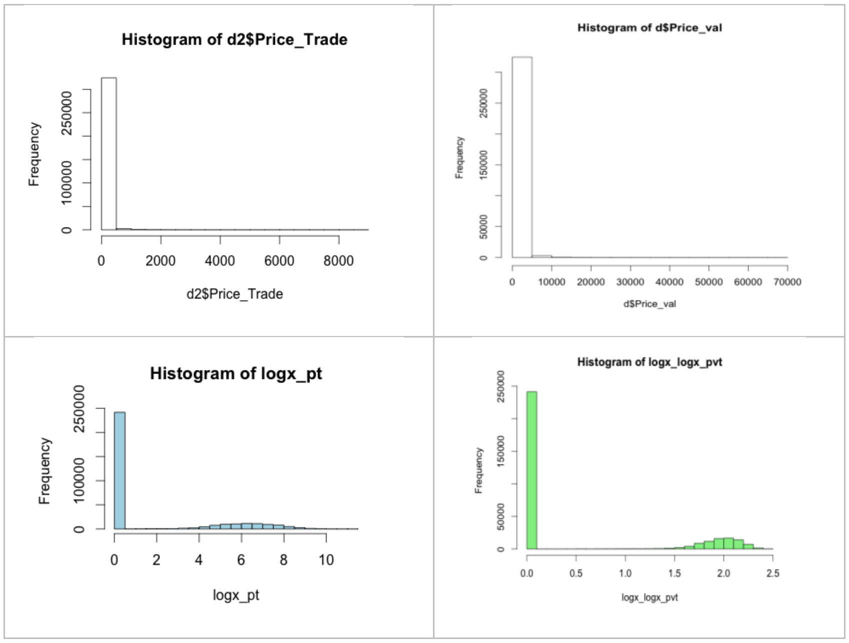
Attributes	***Analysis***
Currency	Different currencies of the trade were standardized into just one attribute due different currency in exchange, and removed obsolete currencies (e.g.: FRF = French francs). 99,98% of trades are in EUR.
Type_Option	Nominal categorical Type of option: American (A) or European (E). All records are European so this attribute will be discarded.
Cod_ Underlying	Nominal categorical. Underlying code (underlying asset) can be repeated over time. This value is different almost each row. However, there is no table (provided by the Organisation) that has such information of contracts. Candidate to be discarded
Abandon_ ATM	Abandonment decision option when "At- the- MONEP ", i.e. when it has the same value as its underlying. And that's why it's abandoned. Candidate to be discarded as it is a consequence of the Expiry event.

Regarding the **underlying code** attribute (Cod_Underlying), it was invested some effort in relate two important attributes, field CNACR (from options tables) and CAINI (ID code of the underlying asset table, without the suffix). Both values represent the underlying asset code of an options but CNACR has the suffix of its option in its name that could be 1 to 3 letters sometimes ending with an underscore. This nomenclature is used because of the process of daily updating of pricelists but the difficulty is to find a way to connect a specific option with its underlying asset. Some attempts were made to be able to have these values before the daily updating of pricelists but the values then didn't match, as there isn't any archive process with more than one week, so impossible to have them for the 3 years. For that reason, the underlying code is discarded.

Data from business and technical points of view

After analysing the dataset and as seen in table 16, 'Price' and 'Price_val' have a correlation of `0.793988` (function `cor`). The table below illustrates the comparison between both and the similarities are big. Even using the logarithm to show more results in the range of 0 and 2000, the values approximate and we can see that small number of records have a high Price and Price_val.

Table 16: Comparison between Price and Price_val in R. The histograms in blue and green are Logx



Using the linear model (function `LM`) for a single stratum analysis, the plot below shows the values continues to be quit similar. Values have similar values having Price_val most of time 100 bigger than 'price'. It means that apparently Price_val seems to be the same value of 'price' but without the decimals, ex: 'price=23,48', then 'Price_val=2348'. It wasn't possible to know more about this subject even with the help of professional operators, as they're not aware of this specific variable and didn't want to compromise; also, any other feedback from the provider was seen. Figure 19 illustrates a plot with correlation of both variables where the high values of Price_val seem to be the only difference.

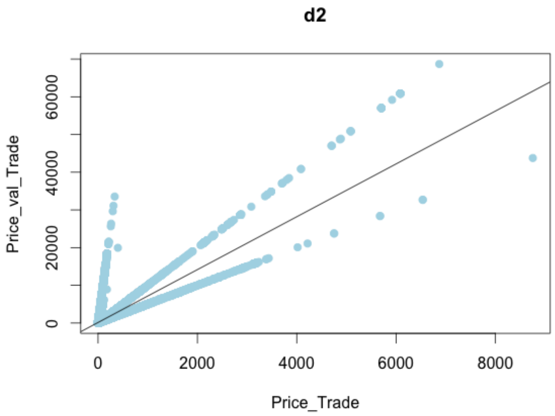


Figure 18: Plot by using `abline(LM$coefficients)`

From IT teams point of view, it is related with the projection of the theoretical price of price and what is the commercial value of a given option at a specific time. After analysing the value of them line by line, we can have the following types of behaviour: 0 < Price_val → this could mean that the value of price is not present of 0, and undervalued (as Price_val is positive); Price < Price_val → price is undervalued; price = Price_val → Price is well estimated; and finally price > Price_val → price is overvalued.

Could be each time a price valuation is bigger than the actually price, there are four possibilities depending on the type of option (call/put, short/long). So, the conclusion is that the 'Price_val' cannot be 100% taken into account because the meaning isn't totally understood with the tools authorised by the Organisation. However, it will be maintained in the list of attributes with a look of suspicion and attention. After analysing each attribute, it has been realized that, despite of a lot of fields available in tables for relevant **technical information** it was found that some important fields weren't been populated as thought, which was a surprise as a lot of effort were done to be able to have these data, at the time being.

The data understanding is completed if permission to do the interviews would not have been granted, though these took more time then expected.

Data analysis after interviews elements

The CRISP-DM approach evolves chronologically starting with the business understanding to the data understanding then, if justified, it could retreat back to the business understanding again for second thoughts. When the interviews were granted, the analysis of the data from the database was already done, so it needs to be reviewed from the business understanding's point of view if all of the analysis further is still in place. As described in the business understanding, two interviews were prepared in order to collect important information and understand which of that can be included in the dataset. The interviews could be done as soon as the authorisation came; just before completing data understanding step and important information was retrieved, as tables 17 and 18 are shown. Both complete interviews are available in the

appendix A. The first interview is with the head of IT Support of the Organisation. He has a more practical and technical profile, which is important to confirm the database information export process.

Table 17: Interview summary of the head of IT Support of the Organisation

Researcher	***Head of IT Support of the Organisation***
What is the global functioning of the OPW[8]?	There are two types of derivative options: standard (DAECA[9]) and flexible (DAECB[10]). **OPW aims to close positions** making the connection between the closing position requests and the settings (positions).
How the options appear in the Unix platform? What operations can be done in SGWs[11] and in the OPW?	SGWs are communication ports between market and its trading players. In the maturity date, the OPW server connects to the market through the corresponding market SGW, one per market. Here, the market intends to write by SGW to the database and the OPW server reads from the market, and is connected to OPW client, that informs to the market each time an E, S or B is communicated.
What is the normal flow of information?	A trade comes from the market station and enters to Ubix through the SGW and Lisa gateways. Between them, there are some operations including clearing, integration in Ubix system (*Tradefeed* or SRX), or the position confirmation making Exercise (by SGW of SRX or OPW).
What variable would you consider essential to an introductory information analysis?	The most important variables are the ones that somehow interconnect and tag each other and the database system. They are: CORIG, NUBIX and CNAOP. If we know the CORIG, we will be able to identify the origin of the trade / position. The NUBIX is the line code and CNAOP the line type.

The trader/investor has been directly involved in financial markets since 1998. Initially as manager of a bank's investment and trading portfolios, as well as running property funds and a pension fund. However his interest then moved to developing systematic trading systems for derivative (futures and options) and currency markets. In the latter role he became partner in a hedge fund before working on his own account.

[8] OPW described in the glossary, annex A.

[9] DAECA – Standard date of maturity

[10] DAECB – Non-standard date of maturity

[11] SGW: described in the glossary, annex A.

Table 18: Interview summary of the independent trader/investor

Researcher	***Trader / Investor***
Could you indicate what are the most sensitive period of European option life cycle, and the role of the trader and hedger?	The most sensitive time period for European options is in the run up to maturity, in the **day of maturity**. Rather than focusing purely on traders or hedgers, the important distinction is to look at who are **long** the options and who are **short** the options at maturity. (…)
During the maturity date (…), what are the critical factors of success for an option is exercised (In-The-Money)? What are the critical factors in case of abandoned (Out-Of-Money)?	On the maturity day, the chances that an option will be exercised really depend on where their **strike price is in relation to underlying price** and the maximum open interest strike price. If there is significant open interest and a low chance of a significant news item, then there is a high probability that the price of the underlying will be **pinned** near the nearby maximum open interest strike price. (…) There are other **external factors** whereby an option holder may not exercise an option even if it is “in-the-money”.
Besides external factors (…), which directly influences the prices of an option randomly?	Option pricing is based on a whole number of **factors** and **expectations**, of which the underlying price is just one. A crucial part of an option’s value derives from expected **volatility** and from **changes in market** sentiment.
Easily guess that periods of higher positions closing activity tend to be close to the closing time of the respective market, does it also in market opening?	If we ignore planned data announcements, from the US payrolls data to the ECB press conference by Mario Draghi, or unexpected sudden market moving events, you will find that market liquidity and volume is concentrated near market **open** and near market **close**.
We analysed the database having a large number of trades purchased during the day of maturity. What may be due this fact or not true? It is common to have a large volume of business done on the day of maturity?	With options, **yes**, you have a lot of volume at maturity as various players rebalance their books in trying to deal with the uncertainty of whether their long option positions will be in-the-money, or their short option positions will be exercised against. You don’t necessarily have to trade the underlying to hedge your position; you can trade the options, which on the day of maturity are more or less a substitute for the underlying.

The interviews have shown that the business understanding produced before the interviews is globally validated. After analysing the answers of the head of IT Support the information retrieved by the different databases is correct. What the manager didn’t expected is that some important information isn’t correctly saved in the databases. It means that or the data is not saved, or the data isn’t correctly been archived, as the target databases were from a test environment, the only with permissions to this study. Another subject mentioned is the fact that 90% of options have standard maturity date (DAECA) rather the non-standard (DAECB), meaning that, to facilitate the understanding of other options data characteristics and as DAECB is a minority, this attribute (DAECB = Mat_no_Std) will be discarded and the standard date will be saved (DAECA = Date_Std_DAECA).

Regarding the answers form the trader; important data has not been saved due to complications relatively to both security and availability of data (the Organisation, the Bloomberg helpdesk on-line archive process). Thus, the risk log of table 8 was updated for the "Results impacts due to missing ratting" business risk from 20 to 25, the "Lack of attribute definition for daily financial news" business risk from 8 to 10 and the "Some attributes are empty" IT risk from 15 to 25, in terms of risk response. Table 19 shows the information needed to do this research that we cannot have:

Table 19: Relevant information that influences the accuracy of the study

Attribute	***Reason***
Underlying asset and its rate	The critical factors of success for an option is exercised "really depend on where their strike price is in relation to **underlying price** and the maximum open interest strike price".
External factors	Impossible to predict some external factors like natural or economic disasters or foreign policy. This factor will be always part of all studies related with this subject, as the finance doesn't play as an isolated science.
Volatility	"A crucial part of an option's value derives from expected volatility". The volatility is possible to predict but again there is no data in the database to be used in order to calculate it.
Changes in the market sentiment.	Option pricing is based on **factors** and **expectations**. Traders and hedges go in opposite directions where the market tries to make their move as neutral.

It is clear that we don't have this info that they pointed, so the study will be damaged and we can foreseen that one of the lessons to be informed to the management of data is that it's not possible to predict with a good accuracy when an option is exercised or not. After the interviews, some attributes needs to be overlooked in order to understand which of them should be discarded. The list of attributes is presented in Table 20.

Table 20: List of attributes to be discarded and why

Attributes to be discarded	
Clearer	Legitimate entity to authorise the clearing. As the majority of manual operations are "MONEP", for MONEP options and few other values are different from MONEP, this attribute doesn't give added value to the characterisation of the option.
Code_C	Contract code of contract. As it has too many differentiated values, this attribute doesn't add interesting factors. Is it this attribute relevant for the study? Still in doubt if this attribute should be part of the dataset, perhaps be clustered could help.
Price_val	Valued price of the trade, (numerical real). Unknown value but curiously bigger than the strike price but with a lower mean, and different than price too. Sometimes 10 times bigger than price, but it´s not always the case. Until now, it wasn't possible to know more, so still under investigation. (Median: 0.0, mean: 333.2, 3rd Qu.: 39.0 and max: 68686.8).

Attributes to be discarded	
Currency	Different currencies of the trade (nominal categorical) were standardized into just one attribute due different currency in exchange, and removed obsolete currencies (ex: FRF = French francs). 99,98% of trades are in EUR.
Cod_Interv	Intervener's trade code. This value can symbolise an individual or an entity and their values can be repeated. Biggest values are: 96152T (113394), 94923 (100512), BNAPAERX (52005), 99999 (26151) BFI.PARI (13063) and GSS (11362). As users codes have 5 and only digits and typically start with the digit 9, values “96152T” and “99999” seem to be outliers. BNAPAERX is a known entity and GSS is the technical name to designate EMEA (Europe, the Middle East and).
Type_Op	Manual or electronic operation. Nominal category, created attribute: electro (51,44%) , manual (48,56%).

Data understanding results

The lack of technical data quality archived regarding the information systems in the Organisation is known. Therefore, it hampers to decision making in the resources (human, technical and time) in providing a better IT support for the operational teams. Reminding it, it was during the interview of Mohamed Tricky, where he says that they are aware of this gap but the cost of having these statistics takes more resources comparatively to the gain reward in doing it. Unless a new project is decided to provide this data quality and decisions on this direction is taken, the truth is that, if it doesn’t provide a balanced lucrative approach, the decision of having this change will be remote or even not considerate. Looking at the data collected and comparing to what the interview of the trader is different of the expected. The technical improvements are to be made to provide a better performance and support is at a first site to change properties to compulsory introduction of data by the operational team (or by computing new developments). Some crucial fields are empty where it was expecting to be fulfilled. This might damage the data mining results and affect the objective to accomplish.

From the perspective of an operator the highest peak of the day is in the morning and typically the last hour of the day. Interviewing the trader was important to find where is the knowledge of any news that might interfere with the business: Buy / sell, expiration date, volume, type of contracts, rise / fall in prices. (Bloomberg, Reuters, CMVM, among others) and what makes a customer withdrawing from exercising a contract. Usually, it oscillates all around the price. Soon the main dependence revolves around the speculation in relation to derivative released by the company, and this information can be found in credits rating like the SP&500. Theoretically, market prices must match the theoretical price that a given contract is (the underlying product, e.g.: exchange rates at any given time). If not, there is an arbitrage opportunity that takes advantage of the price difference and can leverage the rise or fall of it. However, these data are difficult to obtain due to the institution’s data security. In terms of data quality, the data must be anonymous (anonymisation data production[12]). Getting historical or archival data can be served to training and the amount of data collected was from the range of 2013-2015.

From these attributes, only 11 attributes continue pertinent for the study and the others as seen are not qualify for the purpose. In the next phase, data preparation, some of these qualified variables will suffer some alterations in order to fit for the modeling phase with the tool SPSS Modeler.

3.4.3 Data preparation

According to CRISP-DM process model, transformations regarding the quality of data in each of the selected attributes are looked more closely during the data preparation, and it's often considered one of the most important phases of the life cycle (McCormick, Khabaza, Abbott, Mutchler, & Brown, 2013). In this phase outliers, missing data and transformation of absolute variables in a relative one, are taking place for the 328 498 records. As seen in the previous phase, 11 attributes where selected. It is important look into their statistics (in R the function `summary()`) in order to find possible aspects to clean, format, construct or integrate (Chapman et. al, 2000).

Dates transformation

Dates of contract, maturity and modification have structured numeric absolute values, and the date of modification has as well the full timestamp format (DD/MM/YY hh:mm:ss). The year must be excluded from the dataset or in alternative use the year to help to range records of training and records of test during the modeling. Sad that, the dates as transformed in excel with the use of functions like DAY(DATRA), MONTH(DATRA), YEAR(DATRA) and WEEKDAY(DATRA). These new numeric integers are new created attributes, respectively:

Table 21: Date transformation from numerical structure into numerical integer types

Dates to discard	***Type***	***Transformed new attributes***	***New type***
Date_C	Numerical Structure	Day_C, Month_C, Year_C, Weekday_C.	Numeric integer
Date_Std_DAECA	Numerical Structure	Day_Mat, Month_Mat, Year_Mat Weekday_Mat, Days_Tra_Mat	Numeric integer
Date_Mod	Numerical Structure	Day_Mod, Month_Mod, Year_Mod, Weekday_Mod, Hour_Mod, Minute_Mod, Sec_Mod,	Numeric integer

[12] Anonymisation is explained in the glossary, annex A.

Table 22 displays the Acum_EUM, a new variable created to analyse the behaviour of Exercise Y/N for the last 30 days, and Peso_EUM was created to see its ratio. Both attributes showed that there not creating added value for the study as the percentage is in line with what the dates values tell: picks in end of quarter as the majority of trades are intraday, so these variables are discarded as well.

Table 22: Cumulated results of last 30 days of "Exercise Y/N" and its ratio

Acum_EUM	*A new attribute was created in Excel: cumulated results of last 30 days, as a numeric integer. The Excel function is: COUNTIFS(ExerciseY/N; "Y"; Date_C; ">=" & DATE(Year_C; Month_C; Dia_C); Date_C; "<"& DATE(Year_C; Month_C; Day_C)).*
Peso_EUM	A new attribute was created in Excel: % Of exercises occurred in the last 30 days, as a numerical real. The Excel function is: Acum_EUM / (Acum_EUM + COUNTIFS (ExerciseY/N; "N"; Date_C; ">=" & DATE(Year_C; Month_C1; Day_C); Date_C; "<" & DATE(Year_C; Month_C; Day_C))) * 100.

The annex D shows an exhaustive analysis of the attributes from data understanding with statistics all variables in study as well as the discarded ones, but the focus will remains in the 20 attributes.

Missing values and outliers

As said during the data understanding, some fields were completely empty. For those we cannot do anything. Regarding the resulted dataset, very few missing values were found (less than 10) and it was opted to discard them from the total of records. Some outliers were also seen, for instance the manual time of operation, the code of the operational was seen very other like "99999", as if it was a default value. For this it was opted to use the term "manual" and "electro". Some value of "Price_val" seemed to be outliers due to its very high value but nothing was done because this attribute was in the end discarded.

3.4.4 Modeling

The Modeling is the fourth phase of CRISP-DM. It is during this phase that algorithms, methods and techniques are tuned and adjusted to the dataset transformed in data preparation phase. In this study, the Modeling was performed in SPSS Modeler version 18 (http://www-01.ibm.com/support/docview.wss?uid=swg27046871) and it consists of a graphical interface where the user can chain and connect nodes in a stream desktop. Creating a modeling process is basically looking to three major group of nodes that tend to be chosen and interconnected: the input (source), the model (algorithm) and the output (result). Table 23 summarizes the parameters of common input nodes for the study, just before the application of the learning algorithms:

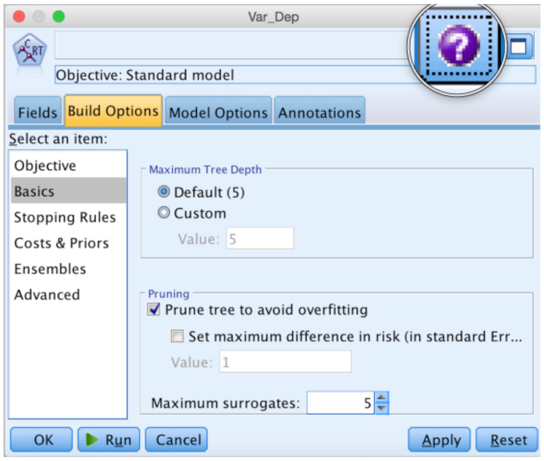
Table 23: Common parameters of nodes for the first modeling process

Common	***Properties***
Data	ExerciseY/N, Day_C, Month_C, Weekday_C, Market, Call_Put, Strike_Price, Long_Short, Quant_C, Price, Cod_Orig, Weekday_Mat, Day_Mat, Month_Mat, Days_Tra_Mat, Day_Mod, Month_Mod, Weekday_Mod, Hour_Mod, Minute_Mod, Second_Mod
Data audit	Outliers: 3,0, Extremes: 5,0
Generation	Distribution selection node: Market ='MONEP'
Partition	Training: 70, Test: 30

In the data audit node, basic statistics and graphs are displayed, such as the number of missing values. It indicates that none of the attributes selected have missing values, which confirms the process of the data preparation phase. As mentioned during the data-understanding phase, the attributes that were discarded are filtered in the node dataset, and the distribution graph node was as well attached to the dataset in order to include the market MONEP, by generating this selection. The partition node can now be connected to the generated node and selects the proportion of training and tests (holdout): the training partition size should be bigger than the test partition size (Larose, 2005) and by default it is 70/30, respectively (Caetano, 2013). The group of nodes that are needed for the selected DT chosen is now complete and each algorithm is attached to this group.

Select Modeling algorithms – 1st modeling test

A first modeling was performed with C&RT, C5.0, CHAID and QUEST algorithms attaching their nodes from the partition node, and selecting their proprieties (fully described in Annex D) as well as their results. Figure 20 shows the SPSS Modeler's Help, a crucial function to understand what is the best value and if it can be adjusted for a second test.



Var_Dep

Objective: Standard model

Select an item:

Maximum Tree Depth
● Default (5)
○ Custom
Value: 5

Pruning
☑ Prune tree to avoid overfitting
□ Set maximum difference in risk (in standard Err...
Value: 1

Maximum surrogates: 5

Figure 19: C&RT algorithm showing the help option (enhanced by the magnifying glass)

Font: SPSS Modeler

Table 24 resumes the main properties of each algorithm; disabled default properties are not mentioned. The C&RT node creates a single, standard model to explain relationships between fields. Standard models are easier to interpret and can be faster to score than combined methods such as boosting or bagging, thus this first iteration will be based on standard models. The maximum tree depth specifies the maximum number of levels below the root node (default frequency of splits is 5). The C5.0 algorithm does not have any "Prune tree to avoid overfitting" method, since it is only available in C&RT and QUEST models and consists of simplifying the tree by removing bottom-level splits that are not needed for the tree to be more accurate, hence this value that is selected.

Table 24: C&RT, C5.0, CHAID and QUEST algorithms default properties and techniques

Methods/properties	***C&RT***	***C5.0***	***CHAID***	***QUEST***
Use partitioned data	□		Y	Y
Calculate predictor imp.	Y	Y	Y	Y
Calculate raw propensity scores			Y	Y
Maximum tree depth: 5	Y		Y	Y
Prune tree	Y			Y
Max surrogates: 5	Y			Y
Overfit prevention set: 30%	Y			Y
Use global pruning		Y		
Pruning severity of 75; min child branch: 2		Y	Y	
Cross-validate folds: 10		Y		
Stopping (min. records) Parent: 2%, child: 1%	Y		Y	Y
Ensembles: cat. Voting; boosting, bagging: 10	Y		Y	Y
Alpha for Splitting and merging: 0.05			Y [13]	Y
Min impunity: 0.0001	Y			Y
Impurity by Gini	Y			
Chi-square for categorical: Pearson			Y	
Min change in expected: 0.001, Max conv.: 100			Y	
Epsilon For Convergence: 0.001			Y	
Maximum iterations for convergence: 100			Y	
Alpha for Splitting: 0.05				Y

Additionally, we could select a better value for the maximum surrogates in order to deal with missing values. As confirmed during the analysis of the data audit, the present dataset does not have any and for now the default value of 5 will stay unchanged. Stopping rules parameter controls how the tree is constructed and when to split into branches of the DT. For C&RT, CHAID and QUEST algorithms the default value is 2% for parent and 1% for child, which means that if the minimum number to be subdivided is 2% of records, than the parent node will not suffer more splits and the same is for child (1%). Increasing pruning severity level in C5.0 and CHAID algorithms to obtain a smaller and more concise tree is selected, and in this case we start with the value 75, with 2 records per child branch.

[13] For CHAID algorithm, this property is affected by using the Bonferroni method (Dursun, Cemil, & Ali, 2013).

The ensemble behaviours are towards boosting, bagging and very large datasets, consisting to combine the predicted values from the base models to compute the ensemble score value. As our target is categorical, predicted values can be combined using voting, highest probability, or highest mean probability and C&RT, CHAID and QUEST will use the default voting combining rule. Boosting method is used to improve the model accuracy rate and bagging or "bootstrap aggregating" is used to improve the stability of the model and avoid overfitting. Boosting and bagging stipulate the balance by enhancing the model between accuracy and stability criteria (for bagging it is the number of bootstrap samples and it should be above 0, default is 10). C5.0 algorithm has a special way to handle with accuracy: by cross-validation. Cross-validation is useful if the dataset is too small to split into traditional training and testing sets, which is not the current case, however this technique will be used, setting it with 10 folds (Larose, 2005). Since the cross-validation (also know as rotation estimation) in this version of SPSS Modeler is performed at the same time and cross-validation models are discarded after the accuracy estimate is calculated, this value can stay unchanged for now. Depending on the result, this parameter could be adjusted for a second iteration.

For out categorical target, a node is considered "pure" if 100% of cases fit into a specific category of the field and the goal of tree building is to minimize the impurity within each node. If the best split for a branch reduces the impurity by less than the specified amount, the split will not be made. The default is very low, 0.0001 and the impurity measure selected is the Gini, based on probabilities of category membership for the branch (other values could be "twoing" or "ordered"). The "Gini impurity" is specifically used in C&RT and it is related to incorrect and random probability of labeling being the faster metric of the available three. The "overfit prevention set" is set to 30%, used to track errors during the training and prevent variation in the data, for C&RT and QUEST models.

In CHAID, some advanced options enable to refine the process of building the tree. The splitting and merging have value 0.05. For low values of splitting it tends to produce trees with fewer nodes, and to prevent any merging of categories the value should be 1. This can be adjusted by using the Bonferroni method (method that tests category combinations of a predictor) based on the number of tests, which directly relates to the number of categories and measurement level of a predictor. This is generally desirable because it better controls the false-positive error rate. Disabling this option will increase the power of the analysis to find true differences, but at the cost of an increased false-positive rate. In particular, disabling this option may be recommended for small samples, which is not the current case. The alternative is to allow "re-splitting" of merged categories within a node, by attempting to merge categories in order to produce the simplest tree that describes the model. If this option is selected, it enables merged categories to be "re-split" if that results in a better solution (IBM, 2016). Additionally, specifying the method "Chi-square" (only for categorical targets) can be used to calculate the chi-square statistic: Pearson (faster calculations and the best option of our dataset) and Likelihood ratio (more robust but takes longer to calculate). Last advanced building option for CHAID is the minimum change in expected cell frequencies (default of 0.001) used to be adjusted if having problems with the algorithm not converging, by increasing this value or increase the maximum number of iterations until convergence occurs, with a default of 100 (IBM, 2016). After selecting carefully the methods displayed in each models represented by pentagons (e.g. the C&RT algorithm as the clear yellow pentagon highlighted in Figure 20), the model can be generated resulting a glowing diamond (in orange). This output can be displayed in different forms: trees (inside the diamond, tab view), chained tables (analysis represented by the “magnifying glass” node) and charts (the triangle nodes).

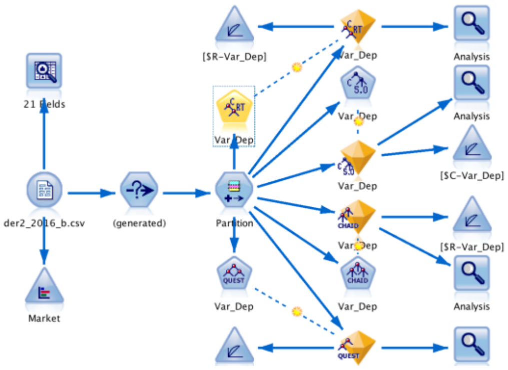


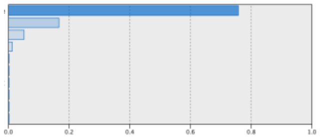
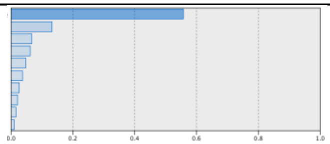
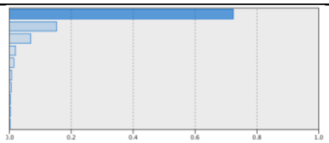
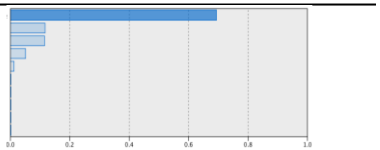
Figure 20: Stream created in SPSS Modeler

Observing Table 25 and as expected, each algorithm builds its tree differently, thus resulting in different data-driven models and predictions; however all had a good result by checking the overall accuracy (correct rows) from training and testing, presenting values above 80%. The fastest method is the QUEST model, requiring only 5 seconds of computational effort, a smaller value when compared with the 1 minute and 8 seconds required by the best performing model: C5.0 (with 99,2% accuracy on the test set). In effect, C5.0 is the most complex model including 15 attributes and 61 trees level of depth, probably due to its intrinsic booting and bagging method. CHAID and C&RT models have very similar percentages for the overall classification accuracy on test data, presenting 85,66% for the case of C&RT and 85,48% for CHAID. The lowest performance is achieved by QUEST, which presents an overall classification accuracy of 80.47%. Despite their differences, one thing is certain: according to the predictor importance – which is the relative importance of attributes as they relate to output variable (Exercise Y/N) plus the variance of predictive error (when an attribute is dropped)[14] (Dursun, Cemil, & Ali, 2013) – the attribute “Price” is the most important attribute in the dataset; after but with a great less expression is “Days_Tra_Mat” and “Strike_Price” attributes.

[14] These values represent the variable assessment that SPSS Modeler does by calculating each time a new variable importance V_{new} is handled. Then each of these new variables is multiplied with their weight in the DT and depending of the number of attributes: $V_{new} = \frac{V - V_{min}}{V_{max} - V_{min}}$ and $v_n(fused) = w_1 v_{1n} + \cdots + w_m + v_{mn}$, where v=var. importance, wi=weight, m=models number and n=attributes number.

The full description of each result is displayed in annex E. Table 25 summarizes below the information displayed in each model.

Table 25: C&RT, C5.0, CHAID and QUEST model observation results, 1st modeling

	C&RT model		***C5.0 model***		***CHAID model***		***QUEST model***	
Predictor importance								
	1. Price 2. Days_Tra_Mat 3. Strike_Price 4. Second_Mod		1. Price 2. Days_Tra_Mat 3. Strike_Price 4. Call_Put		1. Price 2. Days_Tra_Mat 3. Strike_Price 4. Call_Put		1. Price 2. Strike_Price 3. Call_Put 4. Long_Short	
Analysis	Tree depth: 4 levels		Tree depth: 61 levels Cross Validation mean: 98.4 Standard Error: 0.0 Number of records: 185,271 Analysis Accuracy: 99.186%		Tree depth: 5 levels Number of records: 185,271 Analysis Accuracy: 85.554%		Tree depth: 5 levels	
Input:	12 attributes		15 attributes		13 attributes		15 attributes	
Training summary	Algorithm: C&R Tree Model type: Classification Stream: Stream_Lidia.str User: home Date built: 10/18/16 2:33 PM Application: IBM® SPSS® Modeler 18 Elapsed time for model build: 0 hours, 0 mins, 13 secs		Algorithm: C5 Model type: Classification Stream: Stream_Lidia.str User: home Date built: 10/17/16 12:00 AM Application: IBM® SPSS® Modeler 18 Elapsed time for model build: 0 hours, 1 mins, 8 secs		Algorithm: CHAID Model type: Classification Stream: Stream_Lidia.str User: home Date built: 10/17/16 12:16 AM Application: IBM® SPSS® Modeler 18 Elapsed time for model build: 0 hours, 0 mins, 14 secs		Algorithm: QUEST Model type: Classification Stream: Stream_Lidia.str User: home Date built: 10/18/16 5:59 PM Application: IBM® SPSS® Modeler 18 Elapsed time for model build: 0 hours, 0 mins, 5 secs	
Training	**1_Training**		**1_Training**		**1_Training**		**1_Training**	
Correct	158,843	85.74%	183,762	99.19%	158,506	85.55%	148,893	80.36%
Wrong	26,428	14.26%	1,509	0,81%	26,765	14.45%	36,378	19.64%
Testing	**2_Testing**		**2_Testing**		**2_Testing**		**2_Testing**	
Correct	68,191	85.66%	78,970	99.2%	68,053	85.48%	64,063	80.47%
Wrong	11,419	14.34%	640	0.8%	11,557	14.52%	15,547	19.53%

Select Modeling algorithms – 2nd modeling test

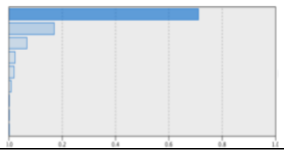
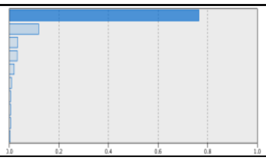
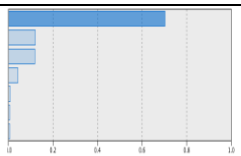
A second modelling iteration was performed with C&RT, CHAID and QUEST algorithms with the very same dataset. Their proprieties selected slightly changed and fully described in the annex D, section 2nd modelling test in page 99 as well as their results. Regarding the C5.0 model, it seems to be already a very good model so it will be discarded for this second round. Table 26 resumes the properties that have changed and affected to the particular model. For C&RT, the maximum number of surrogates has changed to 1, as we do not have any missing values in the dataset. The boosting is a technique "to convert weak learners to strong learners during the training" (Zhou, 2012, p. 23), so in order to boost a bit more the level of accuracy, the number of component models for boosting was raised from 10 to 20.

Table 26: C&RT, CHAID and QUEST algorithms parameters changed for the second modeling

Methods/properties	***C&RT***	***CHAID***	***QUEST***
Prune tree	N		Y
Max surrogates: 1	Y		Y
Ensembles: cat. Voting; boosting, bagging: 20	Y	Y	Y
Chi-square for categorical: Likelihood-ratio		Y	

Concerning CHAID's algorithm, why should we continue to adjust the model with significance value using Bonferroni's method? Because it adjusts the "significance values" when testing the combination of groups of a particular attribute with high preference (the possible predictor). Then, the values are adjusted depending of the number of tests, i.e., number of groups to detect the best predictor, in order to control the false positive error rate. If this method is disabled, the level of finding true positives (TP) increases but the difference of false positives (FP) can increase too. This option is encouraged to be discarded only if the dataset is small, which is not the case (IBM, 2016). In addition, the Likelihood ratio was selected instead of the Pearson method, seen as more robust then Pearson. For QUEST algorithm, the parameters that were changed are the ensemble value as well as the maximum of surrogates. As it has the worst result of the first modelling, the pruning tree was maintained. Table 27 summarizes below the information displayed in each model for the second modeling.

Table 27: C&RT, C5.0, CHAID and QUEST model observation results, 2nd modeling

	C&RT model	**CHAID model**	**QUEST model**
Predictor importance			
	1. Price 2. Days_Tra_Mat 3. Strike_Price 4. Long_Short	1. Price 2. Strike_Price 3. Days_Tra_Mat 4. Call_Put	1. Price 2. Strike_Price 3. Call_Put 4. Long_Short
Analysis	Tree depth: 5 levels Number of records: 185,271 Analysis Accuracy: 85.735%	Tree depth: 5 levels Number of records: 185,271 Analysis Accuracy: 85.554%	Tree depth: 5 levels

	C&RT model		**CHAID model**		**QUEST model**	
Input	9 attributes		14 attributes		7 attributes	
Training summary	Algorithm: C&R Tree Model type: Classification Stream_Lidia_2ndmodeling.str User: home Date built: 10/27/16 6:16 PM Application: IBM® SPSS® Modeler 18 Elapsed time for model build: 0 hours, 0 mins, 12 secs		Algorithm: CHAID Model type: Classification Stream_Lidia_2ndmodeling.str User: home Date built: 10/27/16 7:14 PM Application: IBM® SPSS® Modeler 18 Elapsed time for model build: 0 hours, 0 mins, 13 secs		Algorithm: QUEST Model type: Classification Stream_Lidia_2ndmodeling.str User: home Date built: 10/27/16 8:01 PM Application: IBM® SPSS® Modeler 18 Elapsed time for model build: 0 hours, 0 mins, 5 secs	
Training	**1_Training**		**1_Training**		**1_Training**	
Correct	158,843	85.74%	157,619	85.07%	148,893	80.36%
Wrong	26,428	14.26%	27,652	14.93%	36,378	19.64%
Testing	**2_Testing**		**2_Testing**		**2_Testing**	
Correct	68,191	85.66%	67,631	84.95%	64,063	80.47%
Wrong	11,419	14.34%	11,979	15.05%	15,547	19.53%

Table 27 summarizes the second modeling for C&RT, CHAID and QUEST. In a certain way the results seem to have decreased their percentage of accuracy comparatively to the first modeling attempt for the CHAID model. C&RT and QUEST have their respective same result comparatively to the first modeling, despite their boost tentative. An interesting factor that has changed is the calculation of the predictor importance, regarding specially the 4th predictor: from *Second_Mod* to *Long_Short* in the case of model C&RT, from CHAID the case has changed from the 2nd predictor however will not be considered due to a poorer result testing (from 85.48% to 84.95%), notwithstanding the same accuracy analysis 85.554% and just a second faster. Despite the second iteration of modelling and regarding the output evaluation metrics, the algorithm with best test results is still the C5.0. This method presents both AUC and Gini of 0.999 (corresponding to an almost perfect classification model). In second place comes the first iteration of modelling of CHAID, with an AUC of 0.915 and a Gini of 0.829, as Table 28 shows. Note that in the table "1_Training1" and "2_Testing1" represent the first modeling iteration, and "1_Training2" and "2_Testing2" represent the second modelling iteration.

Table 28: Evaluation results of all models, with two iterations of modeling, adapted from SPSS Modeler

	1_Training1		*2_Testing1*		*1_Training2*		*2_ Testing2*	
	AUC	Gini	AUC	Gini	AUC	Gini	AUC	Gini
C&RT	0.881	0.762	0.881	0.762	0.906	0.812	0.906	0.813
C5.0	0.999	0.999	0.999	0.999	-	-	-	-
CHAID	0.915	0.831	0.915	0.829	0.914	0.827	0.913	0.826
QUEST	0.852	0.703	0.853	0.706	0.852	0.703	0.853	0.706

Figures 21 and 22 represent Gains and ROC charts respectively, where we can observe that all DTs studied have their curves above the thick red diagonal, enhancing the good rates of C5.0 (thin red curve) and CHAID (blue curve). As seen in the previous tables, both training and test have similar results, thus we can assume that the testing performance is globally accurate and will be detailed next. In the Gains chart, when testing is at 20%, the results are very similar but after this threshold and particularly above 50%, the predictive performance of some methods, such as QUEST, become less valuable. This means that all DTs start to have a good acceleration performance in terms of accuracy. With less then 40%, C5.0 model achieves its maximum gain % and keeps it stable until the end of processing. The other models have a wider curve of progress, achieving their best at around 90% of processing. Figure 21 confirms that CHAID has a better performance when compared with C&RT and QUEST.

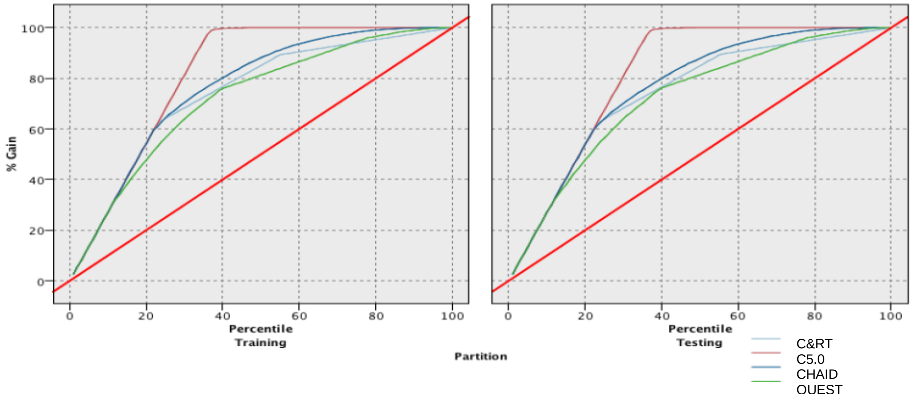


Figure 21: Gains chart where all DTs tested in the first iteration of modeling

Regarding the Receiving Operator Characteristic (ROC) curve analysis, the results are presented in Figure 22. In the figure, the y-axis denotes the True Positive (TP) rate, also known as sensitivity, while the x-axis represents the False Positive (FP) rate, which is equal to 1 - specificity. The larger the area of the ROC curve, the better is the discrimination, Overall, similar comparative prediction performances were observed, with C5.0 presenting the best performance (almost perfect), followed by CHAID.

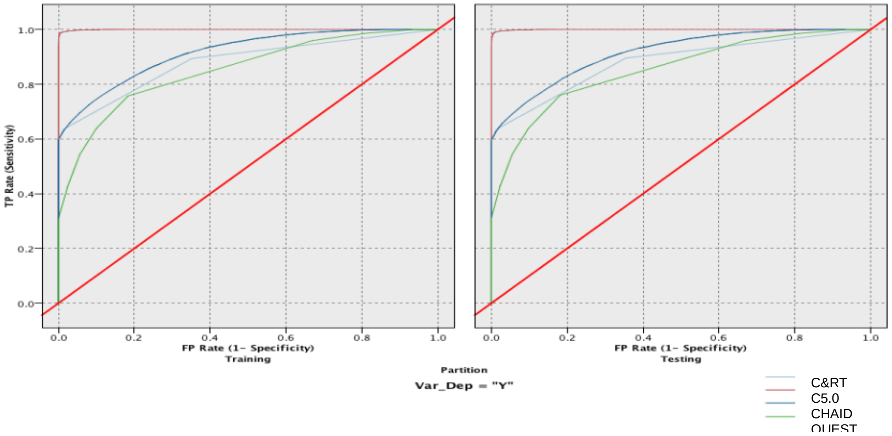


Figure 22: ROC chart type where all DTs tested in the first iteration of modeling

3.4.5 Evaluation

The evaluation phase consists to evaluate the models that were used. For that, a sensibility analysis is done with performance measures in order to better evaluate which of the 4 models should be applied to the problem. As the SPSS Modeler Help tool suggests in its glossary page, the Confusion Matrix is “the number of cases correctly and incorrectly assigned to each of the groups based on the discriminant analysis. Sometimes called the "Confusion Matrix."” (IBM, 2016). The performance measures chosen for this study are the measures that affect directly the results of the target, “exercise” (Y) or “not exercise” (N): accuracy, precision, recall, AUC and F-measure, and they will evaluate each tested model. The overall accuracy traduces the ratio between the correctly predicated cases with the total number of cases. In other words, the percentage of records correctly predicted by the model (Dursun, Cemil, & Ali, 2013). The correctly predicted cases are in fact the number of “true positive” cases that can be found in the result of each model. The concepts of true/false and positive/negatives are organized in the confusion matrix displayed in Table 29.

Table 29: Confusion matrix (or classification matrix) for financial performance

	Unsuccessfully predicted	*Successfully predicted*
No Exercise	*True Negative (TN)*	*False Positive (FP)*
Exercise	*False Negative (FN)*	*True Positive (TP)*

Font: adapted from (Dursun, Cemil, & Ali, 2013)

The columns represent the predicted results and the rows the actual values of the dataset to compare with. Normally this matrix is confronted with trained, tested or validated results of a model, having the possibility to discover how good or bad are the results. Thus, the overall **accuracy** is in fact the ratio of the number of true positive and negative cases by the total of cases. Another popular classification measure is the **precision**, which is the ratio between the true positive cases with all positives (true and false) cases. Figure 23 illustrates the difference of accuracy and precision, visualizing that it is important to have both high accuracy and precision if we want to make sure that a classification result is globally good.

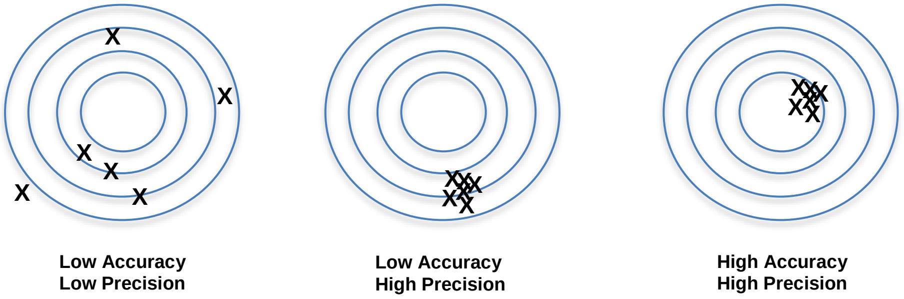


Figure 23: Accuracy and Precision meanings, adapted from (Research Gate, 2016)

The **sensitivity**, also known as recall is the rate of true positive TP cases (true positives over true positives and false negatives) and the **specificity** measures the rate between true negatives with the true negatives and false positives. Finally the **F-measure** condenses both precision and recall measures under a single computation. As described in previous phases of the CRISP-DM process model, the algorithms elected to answer to the objective proposed are decision tree models: C&RT, C5.0, CHAID and QUEST. In order to determine how well these models performed with real data, the dataset was split into two k folds 70% of training and 30% testing. In R, the function responsible is *`Holdout(df$Var_Dep,ratio= 2/3,mode="order").`*

3.4.6 Deployment

It was seen during the evaluation phase that the best method is the model C5.0, but this is not the end of the data mining process. Achieving the purpose of the objective 1 – predict if a particular financial derivative option will be exercised during the maturity date (so European options) – and after increasing knowledge from exclusively Organisation database data, the knowledge gained has to be organized and presented in a manner that it can be used. This assumption follows the CRISP-DM process methodology and now to next step is to design a plan to put in place.

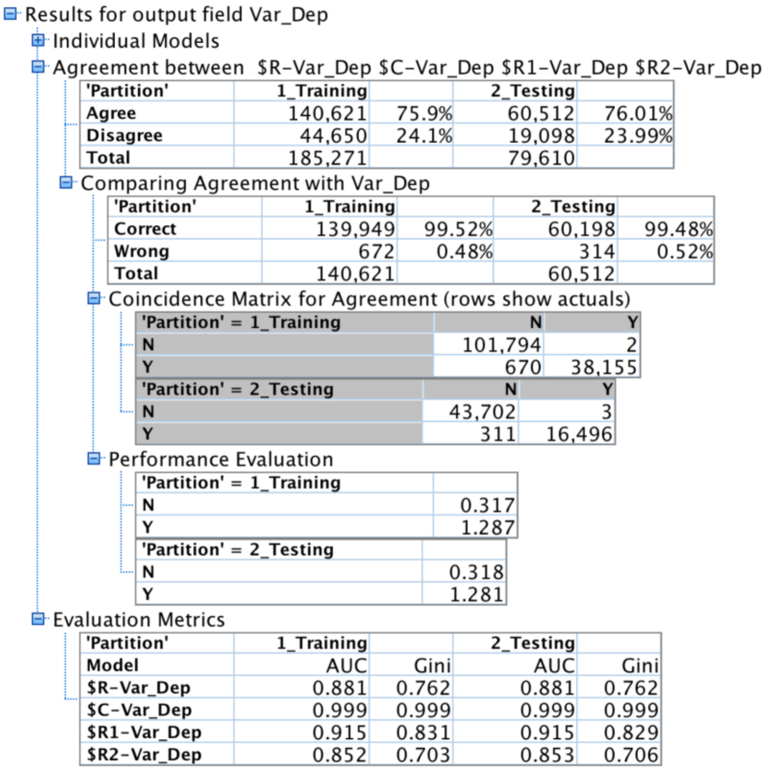
Due to temporal restrictions, the schedule set for this dissertation thesis does not allow to the execution of the implementation phase, which would require the execution of three major steps. The first step would be to inform the IT manager about the discovery knowledge made during this iteration and the importance of this stud. If the manager is willing to proceed, a second step would be to implement the model into the IT organizational system, via a support IT team. Finally, the third step would involve the real usage of the C5.0 model, preferentially during the day of each standard maturity date, to assist IT and managers in their decision making operations executed on a worldwide manner.

In particular, we highlight that the C5.0 model could be used primarily by the eastern team, based in APAC region that serving the very last iterations to help decision making for EMEA region. KOP region will only be used to European option only, also not possible as well for non-flex special options (non standard maturity date). During the C5.0 model usage, a data-driven maintenance procedure would be required, such that the C5.0 could be retrained as more data is collected.

4 Results and discussion

4.1 Results

The “Coincidence Matrix for Agreement” presented in Figure 24 shows the results for the output field Exercise Y/N. The individual models were already checked during the modeling phase (previous chapter). In this section, we will check if the matrices obtained from different classifiers are in the agreement. The results expressed that in 75,9% of training cases, all models agreed between them and the rate raised to 76,01% in the testing. This matrix is comparing with the best predictive performance (C5.0) confusion matrix for test data, where 99,48% of the cases are correct. This last confusion matrix shows that in only 3 FP are detected in the testing and 311 cases were incorrectly flagged as FN.



Results for output field Var_Dep

Individual Models

Agreement between $R-Var_Dep $C-Var_Dep $R1-Var_Dep $R2-Var_Dep

'Partition'	1_Training		2_Testing	
Agree	140,621	75.9%	60,512	76.01%
Disagree	44,650	24.1%	19,098	23.99%
Total	185,271		79,610	

Comparing Agreement with Var_Dep

'Partition'	1_Training		2_Testing	
Correct	139,949	99.52%	60,198	99.48%
Wrong	672	0.48%	314	0.52%
Total	140,621		60,512	

Coincidence Matrix for Agreement (rows show actuals)

'Partition' = 1_Training	N	Y
N	101,794	2
Y	670	38,155

'Partition' = 2_Testing	N	Y
N	43,702	3
Y	311	16,496

Performance Evaluation

'Partition' = 1_Training	
N	0.317
Y	1.287

'Partition' = 2_Testing	
N	0.318
Y	1.281

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$R-Var_Dep	0.881	0.762	0.881	0.762
$C-Var_Dep	0.999	0.999	0.999	0.999
$R1-Var_Dep	0.915	0.831	0.915	0.829
$R2-Var_Dep	0.852	0.703	0.853	0.706

Figure 24: The four models evaluation at the same time: Agreements, coincidence matrix for agreement

Legend: $R-Var-Dep (C&RT), $C-Var_Dep (C5.0), $R1-Var_Dep (CHAID) and &R2-Var_Dep (QUEST)

Figure 25 illustrates the first 2 nodes level of the C5.0 decision tree of MONEP, showing that Price is the predictor most importance variance with the separation into two internal child nodes. If the Price is inferior or equal to zero, then values records can be or not exercised, however if the Price is superior than zero, all the records are exercised (Y).

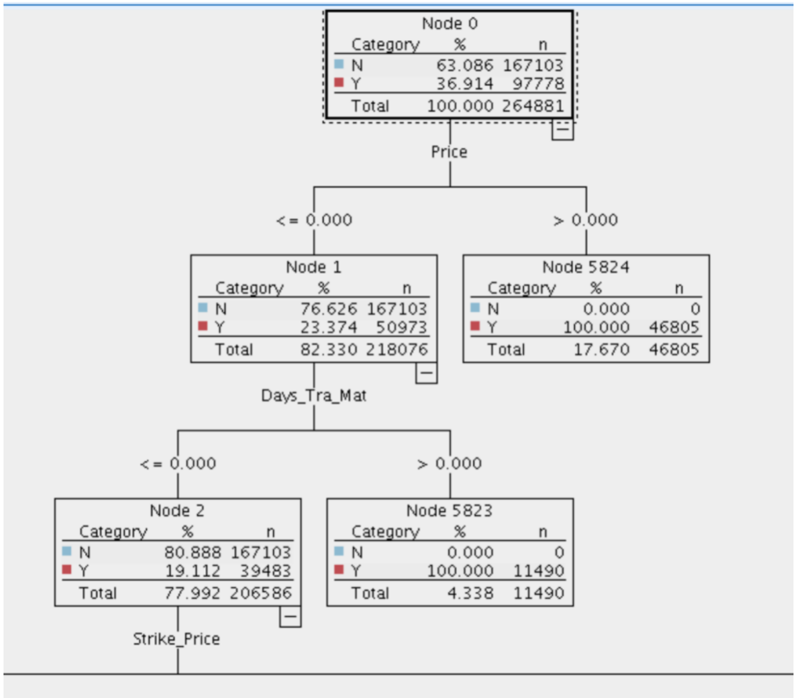


Figure 25: Resulted DT of the C5.0 model, first 2 of its 61 levels, from SPSS Modeler (1st model)

Regarding if records, which Price is zero, if the record was acquired during the day of maturity (Days_Tra_Mat <= 0), than the options could or could not be exercised whilst if the option was acquired days before its maturity date, the exercise will be affirmative. Once the third predictor is confronted, a great range of possibility made the tree very large (to 61 nodes).

4.2 Sensibility analysis and discussion

The analysis results were examined and summarized in Tables 30 and 31:

Table 30: Confusion (or coincidence) matrices of each DT model based on testing dataset

Model type	*True Condition*	*Predicted (N)*	*Predicted (Y)*		*2_Testing*
C&RT	No Exercise (N)	49449	660	Correct	68191
	Exercise (Y)	10759	18742	Wrong	11419
		60208	19402		79610
C5.0	No Exercise (N)	49876	233	Correct	78970
	Exercise (Y)	407	29094	Wrong	640
		50283	29327		79610
CHAID	No Exercise (N)	48672	1437	Correct	68053
	Exercise (Y)	10120	19381	Wrong	11557
		58792	20818		79610
QUEST	No Exercise (N)	45204	4905	Correct	64063
	Exercise (Y)	10642	18859	Wrong	15547
		55846	23764		79610

As shown by the confusion matrices of Table 30, C5.0 presents a remarkable number of correct predictions (correct: 78970), and as such stands out in terms of the overall accuracy rate (Table 31). In effect, C5.0 model obtains the highest accuracy level (99,196%), while C&RT model presents the second highest accuracy level (85,656%), although very close to the accuracy level of CHAID (85,483%). All four methods present a very good accuracy, higher than 80%.

Table 31: Prediction results for exercising MONEP derivative options at a standard maturity date

Model type	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*FP (%)*	*FN (%)*	*Precision (%)*	*F-measure*	*AUC*	*Gini*
C&RT	*85,656*	*63,530*	*98,683*	*3,402*	*17,870*	*96,598*	*254,120*	*0,881*	*0,762*
C5.0	*99,196*	*98,620*	*99,535*	*0,794*	*0,809*	*99,206*	*394,481*	*0,999*	*0,999*
CHAID	*85,483*	*65,696*	*97,132*	*6,903*	*17,213*	*93,097*	*262,784*	*0,915*	*0,829*
QUEST	*80,471*	*63,927*	*90,211*	*20,640*	*19,056*	*79,359*	*255,707*	*0,853*	*0,706*

Font: Measures calculated by the researcher.

When analyzing Table 31, it is clear that C5.0 is the best classifier, presenting the best performance values for all measures. For the comparison among other methods, we note that several of the presented measures are related with one possible trade-off between sensitivity and specificity. In other words, a predicted class is accepted as positive if its probability is higher than a threshold that by default is set to 0.5. This corresponds to one point of the ROC curve. Since different sensitivity-specificity trade-offs can be selected by changing the threshold value, it is better to compare classification models using more global measures, such as the AUC. When analyzing this measure (and also the Gini index), CHAID is ranked as the second best method, followed by C&RT and then QUEST.

According to several evaluation metrics of predictive capacity in the testing data (e.g. overall accuracy, F-measure, AUC area of the ROC curve), the C5.0 model obtained the best results. For instance, this model has an almost perfect AUC (0.999) corresponding to a high level of discrimination, 8.4 percentage points (pp) better than CHAID method and 11.8 pp better than C&RT method. In the second position are the CHAID and C&RT methods. Figure 26 illustrates the ROC curves, where C&RT and CHAID have very similar results, helping to visualize better this comparison.

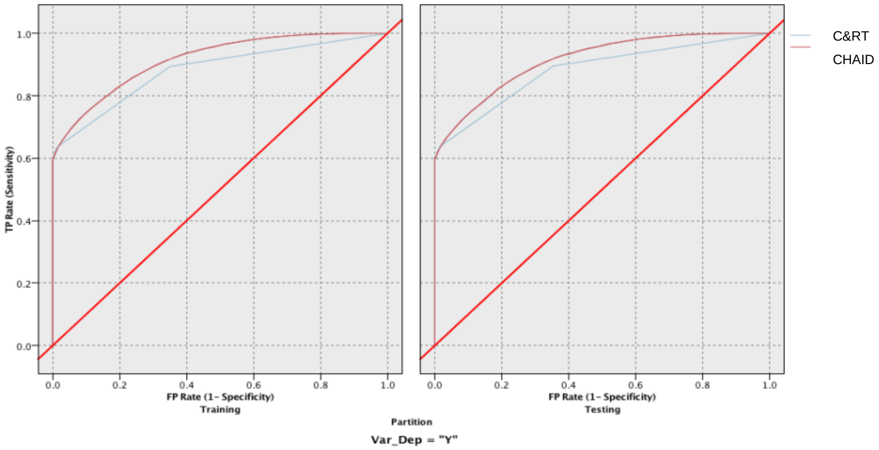


Figure 26: ROC chart type where C&RT and CHAID have very similar results

Both C&RT and CHAID have very similar performances in terms of overall accuracy, rounding the 85,5% in both models. However, it is important to emphasize that the overall accuracy corresponds only to one single coordinate point of the ROC curve for one coordinate point of the 0.5 intersection (as previously mentioned), being that in terms of AUC (and as well the F-measure), the CHAID model presents the second highest performance (e.g. AUC of 0.915 for CHAID comparing to 0.881 for C&RT). Thus, we consider CHAID as the second best classifier in terms of the predictive capacity in testing data. In Figure 25, we can see that CHAID (in red) has a slightly better AUC when compared with C&RT (in blue). Ultimately, the QUEST appears in last position with a smaller AUC value when compared with the previous methods (0.853) but still considered a very good discrimination result.

Additionally to these results, two graphs were created to analyse the influence of one relevant predictor, Call_Put in order to understand the nature of the predicted attribute. Figure 27 shows the influence of Call_Put on C5.0 model, while Figure 28 presents the same influence predictor for the CHAID model.

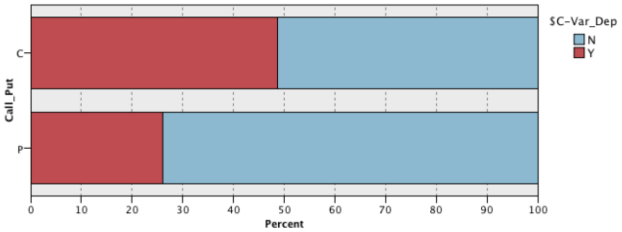


Figure 27: Call_Put, a categorical predictor of C5.0 model, adapted from SPSS Modeler

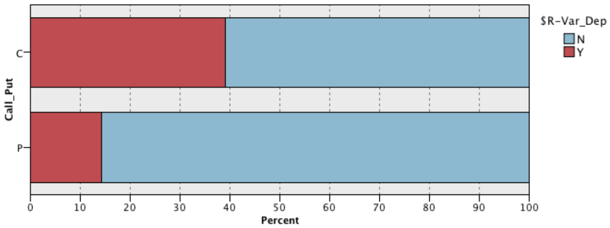


Figure 28: Call_Put, a categorical predictor of CHAID model, adapted from SPSS Modeler

After analyzing all these results and writing the thesis, we found that the continuous attribute Strike_Price had very low values for Y. In particular, we inspected with more detail the C5.0 output result by exporting it to an Excel sheet. We found that Strike_Price is zero when the 2_Testing partition is used in combination with the dependent variable filtered as "Y" but the model predict as "N" (FP, for 233 of cases) or vice-versa (FN, for 407 of cases). This fact could explain why the C5.0 model presents such good predictive result. As such, in future work, a new iteration of the CRISP-DM methodology should be conducted, in which the Strike_Price variable is discarded, in order to check in this case if the C5.0 predictive performance is still very good or not.

5 Conclusions

This investigation problem is relevant because it opens the discussion of how the exercise works and how it is handled in the Organisation. As well, it is appropriate to absorb the acquired knowledge created for instance what is needed to fully understand the drivers of an investor to wanting to exercise (Augen, 2009). If so, the IT support teams could provide a better quality of service in special stressed maturity days. Some evidence in the literature provided new knowledge from the data science insights in general (Chapman et. al, 2000), with practical examples (Caetano, 2013); from financial derivatives options (Avellaneda, Kasyan, & Lipkin, 2011) and (Augen, 2009), and ultimately with the objective to acquire more knowledge from data science techniques applied to financial derivatives options. After looking deeper, few studies were found to the exact problem in hands thus this study made clear that more investigations should be addressed to this topic.

Thanks to the business understanding investigations, it was demonstrated that the major drivers to acquire efficient knowledge was from three requirements gathering. The first one is related with the details of the underlying asset of the derivative option, according to the interviews. The trader of the interview informed as well that the information of the company owner of the derivative option (finance health, rating, news) must be taken into account. A second aspect is related with the availability source of this information throughout time, as it was demonstrated with the attempts made with the help of the Bloomberg helpdesk online and during the interview with the head of IT Support teams of the Organisation. The third aspect is to have the acceptance of the IT manager of the Organisation in willing to allow providing more useful data. Starting to implement this model could be beginning to collect it, at least to be able to understand better in the Organisation how the exercise process work and how the information flows, as decision maker different stakeholders of the derivative options.

Focusing on increasing the quality of knowledge, the results of the model C5.0 during the testing demonstrated that it worth to be considered as it has a very good result (0.999) to be implemented. Should be noted that the methodology chosen was very important to do this study. The CRISP-DM process model is an old but consistent, accepted by scientific community and widely used methodology; moreover it focus a special attention to the business understanding, therefore it was used for this study. The first three stages of this process model with the use of R were determinant due to its intuitive functions and its statistics and methods transformations along the data preparation. Finally, the SPSS Modeler was also important due to the facility of using the DT as classification models and its parameters and analysis of results, which avoided to resort of other DM techniques such as neural networks, SVM, among others.

The challenge of this study is how to acquire useful knowledge of derivatives options in the maturity date, and how the systems could contribute to this increase of knowledge. The knowledge extracted from this work provided the necessary drivers to improve IT services in supporting the business team of derivatives options in the maturity date throughout the globe.

5.1 Contributions

This was the first time that the exercise process was been analysed in the Organisation. Before that, only the operational team and traders have it, and it was only in a little business perspective having the IT support team being unable to help when, for instance, a big amount of requests to be exercised appear. Thanks to this study, the managers have a better idea to help to prioritize projects regarding this new factor: in the maturity date, an European financial derivative option will be exercised depending on its price and its strike price, from how many day the investor booked it until the maturity date and if it is a Call or a Put option. In terms of business success, the result of this study is an additional point in favour to “tender” the decision makers in enhancing IT data storage for a future better data quality. According this study, the C5.0 model is the best DT which reinforce rumours of its popularity for instance reflected by Dursun et. al in 2013 when studied decision trees for measuring firm performance using financial ratios.

5.2 Limitations and research opportunities

Focusing on the knowledge discovered from this study, this investigation contributed to realize that there is still a lot to improve regarding the overall process around the exercise of European financial derivatives options, from IT perspective and in a specific organisation. An important aspect is related with the few knowledge of the investigator and in the business context. In order to bridge the already gained IT knowledge with this complex business and fast pace world of financial derivatives, a course of three months dedicated to this subject was assisted just before starting this project. Although it was an important help to get into details, it was nevertheless a remained challenge to understand data and it took considerable time to figure out such narrow field of derivative options (the exercise). Also, it is important to emphasize the low quality of data, as it was not expected to have so many missing values, some of data were incorrectly introduced (manual input should be avoided as much as possible) and difficult to detect.

There is a strong lack of data not saved or not directly related subjects that have a direct impact in the interpretation of such important decision-making. So, in future work, a new iteration could be addressed by using neural networks or SVN. Also, we intend to collect the data that was focused during the conclusion (chapter 5), the three most important aspects to fully understand an investor to wanting to exercise are: discover all the relations between derivative options underlying asset and its company financial health information progress (post-exercise data difficult to acquire after 3 months); the suspicions of pinning in markets raised by the trader during his interview, that it was not possible to deepen due to lack of time; and the third is related to the limitation of permissions of data regarding the political securities services of the Organisation, as well as the lack of IT performance measures collection. If these characteristics could be collected in a reasonable time frame for this study, it would be possible to certainly discover more and be more thorough and conformed.

Bibliography

Augen, J. (2009). *Trading Options at Expiration - Strategies and models for winning the endgame.* new Jersey: Pearson Education, Inc.

Avellaneda, M., Kasyan, G., & Lipkin, D. M. (2011). Mathematical models for Stock Pinning Near Option Expiration Dates.

Benoit, G. (2002). *Annual Review of Information Science and Technology.* Data mining.

Bernstein, S. (2009). *Understand Derivatibes in a Day* (2ª edição ed.). Kent, EUA: Global Professional Publishing.

Black, F., & Scholes, M. (1973). The Pricing of options and Corporate Liabilities. *Chicago Journals*, (pp. vol.81, nº3). Chicago.

Bloomberg Finance L.P. (2016). *Bloomber L.P.* Retrieved 06 11, 2016 from Bloomberg Anywhere: https://bba.bloomberg.net

Bloomberg L.P. (2014). Deutsche Boerse's EUREX to start offering weekly equity options. *Bloomberg News* , http://www.bloomberg.com/apps/news?pid=newsarchive&sid=acy.havgJ6R0.

Breiman, L. (1994). *Bagging Predictors.* Technical, University of California at Berkeley, Department of Statistics, Berkeley.

Caetano, N. (2013, September). Previsão de tempos de internamento de pacientes via técnicas de Data Mining. Lisbon, Lisbon, POrtugal.

Chapman et. al. (2000). *CRISP-DM 1.0.* SPSS, Inc.

Chen, S. (2016). Detection of fraudulent financial statements using the hybrid data mining approach.

Cortez, P. (2015). *A tutorial on the rminer R package for data mining tasks .* Teaching, University of Minho, Department of Information Systems, ALGORITMI Research Centre, Engineering School , Guimarães.

Dhar, V. (2013). Communications of the ACM. In V. Dhar, *Data Science and Prediction Vol. 56 No. 12* (pp. 64-73). New York .

Dursun, D., Cemil, K., & Ali, U. (2013). Measuring firm performance using finantial ratios: a decision tree approach. *Expert Systems with Applications* , 3970-3983.

Estelle Cantillony, P.-L. Y. (2008, July). Competition between Exchanges: Lessons from the Battle of the Bund.

Eurex. (2014, December 1). *Exchange info of Eurexchange*. Retrieved December 6, 2014 from Eurexchange: http://www.eurexchange.com/exchange-en

European Commission. (2015, 01 11). *European Commission Banking and Finance*. Retrieved 01 18, 2015 from European Commission Web site: http://ec.europa.eu/finance/financial-markets/derivatives/index_en.htm

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Retrieved 12 21, 2014, from http://www.csd.uwo.ca: http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf

Gary Miner, J. E. (2014). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications.* USA, OK: Academic Press.

Gersten, W., Wirth, R., & Arndt, D. (2013). Predictive Modeling in Automotive Direct Marketing: Tools, Experiences and Open Issues. *Data Mining Solutions FT3/AD* , 398-406.

Hull, J. C. (2014). *Fundamentals of Futures and Options Markets* (8th edition ed.). University of Toronto: Pearson Education International.

IBM. (2015, 10 16). *Have you seen ASUM-DM?* Retrieved 5 16, 2016 from https://developer.ibm.com: https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/

IBM. (2016, January 1). *ibm.spss.modeler.help/clementine/.* From IBM SPSS Modeler Help Clementine: http://127.0.0.1:55759/help/index.jsp?topic=/com.ibm.spss.modeler.help/clementine/modelingnode_analyzetab.htm

Investopedia. (2014). *Investopedia*. Retrieved 12 6, 2015 from Investopedi: http://www.investopedia.com/terms/e/exerciseprice.asp

IRMA. (2013). *Data Mining: Concepts, Methodologies, Tools, and Applications (4 Volumes).* USA: ACM Digital Library.

Jeannin, M., Iori, G., & Samuel, D. (2007). *Modeling Stock Pinning.* London, London.

Jeong, D., Yoo, M., & Kim, J. (2016). Accurate and Efficient Computations of the Greeks for Options Near Expiry Using the Black-Scholes Equations. (H. P. Corporation, Ed.) *Discrete Dynamics in Nature and Society* .

KDNuggets. (2014, Outubro 28). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. Retrieved sepembro 02, 2015 from http://www.kdnuggets.com/: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

KDNuggets. (2016, June 1). *R Python top analytics data mining data science software*. Retrieved July 22, 2016 from http://www.kdnuggets.com: http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

Larose, D. T. (2005). *Discovering Knowledge in Data.* Connecticut: Wiley.

Library of Congress. (2015, 01 05). *Library of Congress Prints & Photographs Online Catalog*. Retrieved 01 25, 2015 from Library of Congress web site: http://loc.gov/pictures/resource/ppmsca.03199/

Lothian, J. (2015, 01 22). *John Lothian News* . Retrieved 01 25, 2015 from John Lothian News Web site: http://www.johnlothiannews.com/2015/01/just-floored-says-jeff-sprecher-hates-open-outcry/#.VMUsFXCsX3o

McCormick, K., Khabaza, T., Abbott, D., Mutchler, S. R., & Brown, M. S. (2013). *IBM SPSS Modeller Cookbook.* Birmingham, Mumbai: Packt Enterprise.

Morales, v. M., Horrein, P.-H., Baghdadi, A., Hochapfel, E., & Vaton, S. (2014). Energy-Efficient FPGA Implementation for Binomial.

Moro , S. M. (2011, September). Optimização da Gestão de Contactos via Técnicas de Business Intelligence: aplicação na banca. Lisbon, Lisbon, Portugal.

NYSE. (2014, December 1). *Index of NYSE*. Retrieved Decmber 6, 2014 from https://www.nyse.com: https://www.nyse.com/index

Operational Team MONEP of Organisation. (2013, 05 1). Lited Derivatives: a Functional Overview. Paris, Paris, France.

Oracle. (2016, July 02). *R technologies from Oracle*. Retrieved July 02, 2016 from http://www.oracle.com: http://www.oracle.com/technetwork/database/database-technologies/r/r-technologies/overview/index.html

Organisation. (2013, March 1). Business architecture of listed derivatives. Paris, Paris, France.

Organisation. (2015, march 1). EXE, ASS and NAS image of Option Watch database, in test environment. *Option Watch database, in test environment* . Lisbon, Lisbon, Lisbon: The researcher.

Quivy, R. (1992). *Manual de Investigação em Ciências Sociais.* Lisboa: Gravida.

Research Gate. (2016). *Accuracy and precision*. Retrieved October 23, 2016 from https://www.researchgate.net: https://www.researchgate.net/figure/259581184_fig1_Fig-1-Accuracy-indicates-proximity-of-measurement-results-to-the-true-target-value

Standard & Poor's. (2010). Standard & Poor's Ratings Definitions. *RatingsDirect on the Global Credit Portal* . New York, USA: The McGraw-Hill Companies, Inc.

Wei-Yin Loh, Y.-S. S. (1997). SPLIT SELECTION METHODS FOR CLASSIFICATION TREES. *Statistica Sinica 7* , 815-840.

Zhou, Z.-H. (2012). 2. Boosting. In Z.-H. Zhou, & C. &. Hall/CRC (Ed.), *Ensemble Methods: Foundations and Algorithms* (pp. 23-47). New York: CRC Press.

Zvi Bodie, A. K. (2008). *Investments.* Boston University: McGraw-Hill/Irwin.

Appendix A - Interviews

Interviewing operators to help to find where the information comes from is a delicate procedure. Investors do not compromise where the acknowledgement of certain news take place, because that might interfere with the decision making of the business: buying or selling, with which maturity date, volume, type of contracts, rise / fall in prices, just to name a few factors. Who to interview and what kind of questions to do? Conclusions are:

- However planned, Operators weren’t interviewed. Operators from EUREX and MONEP markets work in different units throughout 3 continents. However their knowledge seems to be prominent, it will have a similar result as the expected with the trader interview. Their contribution would only be possible after a special authorisation of each of their head of unit, so the researcher give up to interview them, and because the head of IT Support will also be interviewed;
- A set of questions was created for two different interviews: one in-person to the head of IT Support, and another via email to an investor/trader;
- The main interest of these interviews is to know what makes a customer withdraws from exercising their right in an option agreement, and how the Organisation would be prepared for eventual picks of real-time data during exercise periods. Usually oscillates all around the price, so it is expected that the main dependence turn around the speculation in relation to derivative released by the owner of option, and how to find this information;
- There may be abnormal earnings? Theoretically, the market price must match the theoretical price that a given contract is (the underlying product, e.g.: exchange rates at any given time). If not, there is an arbitrage opportunity that takes advantage of the price difference and can leverage the rise or fall of it. It is intended to confirm with the operator if there are meters, calculations show that these possible discrepancies. However, these data are difficult to obtain (by the bank's data security);
- Analysis of behaviours like: "Who buys a sort option A, also purchase an option type B?"
- What is the minimum time an exercise of law and how we can speed this up? You can optimize (databases performance, market access and customer confirmation)?

Interviews are oral conversations in-person preference to people carefully selected. This way of getting information has to be very well thought out and has to meet the criteria of conduct and writing so as not to influence the interviewed.

Interview with head of IT Support team[15]

It is an introductory and open character interview, to determine what's the most relevant data to take into account for the act of the Exercise. A datasheet was created with dependent variables on the options when the maturity date approaches.

Mohamed Triki, 3 years of development experience in SunGARD Tunisia (market leader company in developing applications in investment banking) in the area and functional analyst in Listed Derivatives team for about six years, is considered the best person to perform a technical interview about how technically accomplished confirmation Exercises at maturity. Former developer in SunGARD supplier he has a vast knowledge of Option Watch, Stream gateway and Ubix in general, as well as the type of tasks that operators perform daily. Noting also that he is of the most knowledgeable person for specific requests of EUREX market - considered the most important market at this point in the Organisation.

As the research worked with him in a daily basis, the whole interview was a simple conversation, conducted in French. An initial question was asked and the interview turned into an explanation around a paper sketch where each agent was added as soon as mentioned in the conversation.

Researcher: what is the global functioning of the Option Watch application?

Head of IT Support: There are two types of derivative options: a standard (which in Ubix tables to point to the DAECA[16] column) and where 90% of the business is done in this way, i.e., respecting the timing of maturities. This calendar is no more than an indication of the 3rd Friday of each month. If, for some reason, that date is a public holiday, the maturity date is transferred to the previous working day. The second type of option is flexible non-standard maturity date (DAECB[17]) where you can choose the date of maturity. These are European options, and American options can be exercised until the date of maturity and therefore don't need of schedule.

Regardless of the nature of the option, OPW application aims closing position making the connection between the closing position request (requests) and the settings (positions). A pre-structured file, done by operators during the maturity days, integrates requests.

Researcher: How the options appear in the Unix platform? What operations can be done in SGWs and in the OPW?

[15] Ex-SunGARD and the person with the biggest technical and functional knowledge of the tools Ubix, Stream Gateway e Option Watch

[16] DAECA – Standard date of maturity

[17] DAECB – Non-standard date of maturity

Head of IT Support: The stream gateways (SGWs) are communication ports between a given market and its inbound and outbound trading players. The OPW server connects to the market through the corresponding market SGW (one per market). Here, the market intends to writing by SGW and in turn, the OPW server has read permission for the market. The OPW server, which is connected to OPWDT01 database, has the OPW client interface that makes getting through a database link the various automated communications E, S and B. The OPW interface has the write permission on Ubix database UFOWDT03 and reading positions is sent to the Ubix OPW interface, typically via file.

E = Exercise. Business locking intent (option / confirmed sale);

S = aSsignment notification to the underlying asset holder (obligation to purchase / sell confirmation);

B = aBandon option of abandonment.

Researcher: What is the normal flow of information?

Head of IT Support: A trade comes from the market station and enters the Ubix back-office system through the SGW and Lisa gateways. You can identify for example their nature by CORIG variable (recovery source code lines of business that can be trades or positions, the latter if they are cleared). Between the SGW and Ubix, there are some operations including clearing (via Lisa or SRX), integration in ubix system (via *Tradefeed* or also from SRX), or the position confirmation making Exercise (by SGW of SRX or OPW).

Researcher: What variable would you consider essential to an introductory information analysis?

Head of IT Support:

The most important variables are the ones that somehow interconnect and tag each other and the database system. They are: CORIG, NUBIX and CNAOP. If we know the CORIG, we will be able to identify the origin of the trade / position. The NUBIX is the line code and CNAOP the line type. Furthermore, this identification of the last is aided by the SGW of each market, and may be 3 groups: ‘AS’ for standard, ‘AW’ for assigned and ‘AM’ exercised.

Interview with trader

Interviewer: Good afternoon, what is your profession?

Trader: Good afternoon, I am a trader/investor by profession and have been directly involved in financial markets since 1998. Initially I managed a bank's investment and trading portfolios, as well as running property funds and a pension fund. However my interest then moved to developing systematic trading systems for derivative (futures and options) and currency markets. In the latter role I became partner in a hedge fund before working on my own account.

Researcher: In trader quality, the concept of financial options derived from MONEP and EUREX markets is familiar to him. May indicate that the most sensitive period of the life cycle of a European option, and the role of the trader and hedger?

Trader: The most sensitive time period for European options is in the run up to maturity, especially the day of maturity. By then, any time value in the option has disappeared and its value purely comes done to whether the underlying price on the last day will move for the option to expire in-the-money or be worthless. As to the second part of your question, rather than focusing purely on traders or hedgers, I think the important distinction is to look at who are long the options and who are short the options at maturity. Classically, traders will in general be smaller speculators who will be either net long calls/puts in expressing a directional view on the market price. Larger sophisticated traders will more often be involved in collecting option premium on one leg of a multi-legged strategy in order to offset premiums paid on other options, a simple example being a risk reversal. Additionally large sophisticated traders who do take net positions in the options market are often those who have an underlying exposure that they wish to hedge against. Finally we have the market makers who are the most active hedgers due to fact that they are constantly short options given their business of supplying liquidity. Market makers are constantly adjusting their books to be neutral, though with a special focus on Delta and Vega hedging. It's worth noting that the majority of options expire worthless, inflicting "maximum pain" on the maximum number of market participants.

Researcher: During the maturity date (in 90% of cases, the third Friday of each month), what are the critical factors of success for an option is exercised (In-The-Money)? What are the critical factors in case of abandoned (Out-Of-Money)?

Trader: On the maturity day, the chances that an option will be exercised really depend on where their strike price is in relation to underlying price and the maximum open interest strike price. If there is significant open interest and we can anticipate that there is a low chance of a significant news item (scheduled ECB/Fed speech etc.), then there is a high probability that the price of the underlying will be pinned near the nearby maximum open interest strike price. Why does this happen? Well, to answer this we have to look at the role of the market makers. When call or puts are purchased, there has to be a seller. The seller is probably a market maker. Being short the options carries risk and in order to hedge that risk the market maker will either short or buy the underlying in order to maintain a delta neutral position. As the underlying fluctuates in price the market maker will continue to vary the hedge in order to remain neutral. As expiration approaches and options begin closing out in higher frequency, the constant rebalancing of the hedge pressures the underlying to a certain price point or “pins” it. This of course assumes that no event occurs during the day that will cause a sudden new supply or demand of the underlying that will pressure price away from the pin. Therefore if I am long an option and my strike is on the “right” side of the pin, then I will probably be fortunate in being able to exercise my option “in-the-money”.

There are other external factors whereby an option holder may not exercise an option even if it is in-the-money. A clear example is transaction/clearing costs. I may be long a call, which is marginally in-the-money at maturity. If my costs for exercising the option and taking possession of the underling are greater than my expected payoff, then I will not exercise the option. If we look at the example of the Euro-Bund Future options on EUREX, then my exercise of an option will result in my having a position in the next Euro-Bund Future contract, which implies putting up significant margin. Additional uncertainty is created by the fact that on this options contract, the close of trading on the last trading day is at 17:15 CET, whilst the exercise period continues till 18:00 CET. This means that prior to 17:15, my option may be worthless, so I can’t sell it, and then in the next 45 minutes, the contract moves marginally into the money. Do I exercise or not? Additionally as you refer, most maturity dates are Fridays, so if I exercise an option, I will be taking delivery of the underlying over the weekend, exposing me to potentially significant “gap” risk. If you ever want to really see that prices are not continuous, but discrete, hold a position over the weekend and see what happens.

Researcher: Besides external factors such as foreign policy, or economic disasters, which directly influences the rise / fall in prices of a derivative option randomly?

Trader: As you appreciate, option pricing is based on a whole number of factors and expectations, of which the underlying price is just one. A crucial part of an option’s value derives from expected volatility and from changes in market sentiment. The best way to demonstrate this is by example. In the run-up to the first Gulf war, the price of crude oil “went through the roof”, accompanied by a gigantic jump in volatility. It is known that some producers tried to lock in higher oil prices by using strategies that involved buying puts. However the night the bombing started represented a top in both the market price and in volatility. Although the hedges became delta positive, the collapse in volatility and subsequent theta burn meant that some of these strategies actually lost money even though they were “right” in terms of directional movement.

Researcher: Next date of maturity, it has been proven by Marc Jeannin (Jeannin, Iori, & Samuel, 2007) the thesis of *Pinning*[18] by itself advised. The *Pinning* focuses only on the price of the underlying or there are more factors inherent in this?

Trader: Well, the paper that you refer to looks at a number of models, and shows that discrete hedging (non-perfect hedging as in the real world) and liquidity (which is a function of market volatility generated by market orders themselves) have an impact on theoretical pinning effects. An additional fundamental factor that will determine whether or not we can anticipate a pinning effect ex-ante is the amount of open interest that exists in the likely pin strike levels versus the volume traded in the underlying, i.e. how much “skin is in the game”.

Researcher: Easily guess that periods of higher positions closing activity tend to be close to the closing time of the respective markets, we analyse the timeline and for the possibility of price fluctuation until the end of the day. It will nevertheless exist or higher intensity periods, such as, in market opening? There will be more price stability after such period or not?

Trader: If we ignore planned data announcements, from the US payrolls data to the ECB press conference by Mario Draghi, or unexpected sudden market moving events, you will find that market liquidity and volume is concentrated near market open and near market close. In between you will find that markets will “in general” be calmer and offer less directional movement.

Researcher: After doing an analysis with the help of helpdesk team of Bloomberg[19] on the rating of the most relevant companies in the options market study market, the organization warned of the need to maintain the “anonymised” data. There was talk also on the influence of price on the acquisition of the option strike price versus the underlying asset. Victor is aware of some other factor additionally relevant to this study?

Trader: The rating of the clearing members of the exchange would really be the only important parties that would need their ratings to be tracked by the exchange. Long option positions require the payment of a premium. Brokers will ensure that their clients have enough funds to pay the premium in their accounts.

[18] *Pinning* – When the market tends to converge the Exercise price of the underlying asset price strike (Jeannin, Iori, & Samuel, 2007).

[19] *Bloomberg* – It was possible to get in touch with this team through a terminal available at ISCTE library. Thanks to this help, can create a rating table of the most important companies in the European markets, where it can be refreshed whenever possible with real-time data. Bloomberg is an operational financial news agency worldwide.

Option sellers, especially the market makers, however have to post collateral and fund margin variations due to their potentially limitless risk given movements in the underlying. As to the relationship between *premium* as a percentage of the strike price, I guess that is of some relevance to unsophisticated investors who try to get "maximum" leverage when they buy their deeply out-of-money lottery tickets. However as I mentioned above, the two crucial elements for pin risk is the amount of open interest at the pinned strike, and a generally "calm" day where price discovery is not swamped by some significant news item.

Researcher: We have analysed the database that we have a large number of trades that have been purchased on the day of maturity. What may be due this fact or not true? It is common to have a large volume of business done on the day of maturity?

Trader: With options, yes, you have a lot of volume at maturity as various players rebalance their books in trying to deal with the uncertainty of whether their long option positions will be in-the-money, or their short option positions will be exercised against. You don't necessarily have to trade the underlying to hedge your position; you can trade the options, which on the day of maturity are more or less a substitute for the underlying.

Researcher: I would add a comment, observation?

Trader: Never forget that markets are driven by the interaction of human behaviour. A market is a collection of participants that range from the ignorant and ill informed to the sophisticated and well informed. Option markets, much like the Futures markets are more or less a zero sum game, which operate as a wealth transference mechanism from the ignorant majority to the savvy few. If you can understand that, then you will begin to understand price movement.

Researcher: Thank you.

Trader: Your welcome, I hope my comments have been of some use and good luck with your studies.