Study of Market Influence on Tender Performance

Bruno Miguel Repolho Pires

Thesis to obtain the Master of Science Degree in

Mathematics and Applications

Supervisor(s):	Prof. Cláudia Nunes Philippart
	Prof. Igor Kravchenko

Examination Committee

Chairperson:	Prof. António Manuel Pacheco Pires
Supervisor:	Prof. Igor Kravchenko
Member of the Committee:	Prof. Maria da Conceição Esperança Amado
	Eng. Hugo Branquinho

July 2021

Acknowledgments

I would like to express my gratitude to my thesis advisors, Professors Cláudia Philippart and Igor Kravchenko, for the guidance in both, the choice and development of this project.

This project was developed in partnership with EQS Global, and thus, I would also like to express my gratitude to EQS Global, and especially to Hugo Branquinho, for the opportunity provided. Without their support and collaboration this project could not have been possible.

I would also like to address a special thank you to my family, and in particular, my parents, for all their support and hard work that motivated and allowed me to take part in this project.

I would also like to leave a word of appreciation to everyone not mentioned and that, in some way, may have given a contribution for this work.

Finally, I acknowledge the support from the Portuguese Science Foundation (FCT) through the project PTDC/EGE-ECO/30535/2017 - (BL32/2021-IST-ID).

Resumo

Empresas nos ramos da indústria e construção participam em contra licitações de forma a obter contratos com clientes. Nestas licitações o cliente tem um serviço que pretende que lhe seja realizado e as empresas interessadas fazem propostas para realizar esse serviço. Após rever as propostas, o cliente escolhe a proposta mais adequada consoante os seus critérios.

É possível identificar facilmente dois setores onde melhorias podem ser feitas de forma a ganhar contra-licitações sem reduzir desnecessariamente as receitas. Mais concretamente, extração de informação útil de contra-licitações passadas e, como integrar fatores externos no desenvolvimento de novas propostas. Este projeto tem dois objetivos principais: estudar a influência do mercado na performance da empresa nos concursos; criar uma ferramenta de apoio à decisão que aumente a probabilidade de sucesso em concursos futuros.

De forma a estudar a influência do mercado, considerámos duas abordagens, uma explícita onde criámos um índice económico específico para o mercado em questão, e uma implícita onde usámos ”Hidden Markov Models” para obter os estados ocultos do mercado.

Após considerar diversas abordagens para prever a probabilidade de sucesso em concursos futuros, considerámos modelos de Regressão Logística, onde desenvolvemos modelos Frequencistas e Bayesianos. Estes modelos foram incorporados numa ferramenta de apoio à decisão que os gestores da empresa podem utilizar.

Após comparar as performances dos modelos desenvolvidos concluímos que os modelos Bayesianos que incorporam o estudo implícito do mercado produzem os melhores resultados e são os que melhor capturam a relação existente entre o mercado e a performance em concursos.

Palavras-chave: Hidden Markov Models, Regressão Logística, Estatística Bayesiana, Categorização de Clientes

Abstract

Companies in construction and industrial fields participate in reverse tenders in order to carry out contracts with clients. In these tenders, a client requires a certain service to be performed and then a pool of companies propose offers for the given project from which the client must choose the best offer.

To improve performance, two different lines of improvement can be identified. Namely, how to extract additional useful information from the previous auctions and how to integrate the external factors into the consideration. This project has two main objectives: to study the influence of the state of the market on the company's performance; to create a decision support tool for the company managers to optimize the performance in tenders.

In order to study the market influence, we consider two approaches: one explicit, where we create an economic index specific to the market; and one implicit, where we use those index values and, using Hidden Markov Models (HMM), we compute the hidden states of the market.

After considering various approaches for predicting the success probability of future tenders, we consider a Logistic Regression approach where we develop Frequentist and Bayesian models. These models are then incorporated in a decision tool that the company managers can use.

After comparing the performance of all the models, we conclude that the Bayesian models with the market states from the implicit study of the market influence yield the best results and are the best at capturing the relationship between the market and tender performance.

Keywords: Hidden Markov Models, Logistic Regression, Bayesian Statistics, Client Categorization

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

EQS Global is a service provider for the highly demanding industries that operates on construction and industrial markets. On these markets it is common for companies to participate in reverse auctions/tenders or RFQ (Request for Quotation). In a regular auction, i.e. a forward auction, the seller offers a product or service that is in demand or is being requested by several buyers. These buyers participate in a bidding competition, where the highest bidder, after a pre-determined time interval, wins the auction and acquires the goods and/or services at play by the bidding price. In reverse auctions, as the name indicates, the process is reversed, i.e. the buyer is the one that chooses which seller is presenting the best value offer.

Usually this process has multiple phases: first, the client sends the set of services for which the price is asked, then the suppliers answer the request with a price proposal, and finally the client provides a feedback. This can range from immediately accepting the offer, immediately rejecting the offer or giving another opportunity to a subset of the buyers with the best-value propositions so that they can, once again, compete among themselves and present more competitive offers.

In the present paper, due to the provided data, we will look only at one phase auctions. It is important to mention that, although the price is the main criteria, it is not the unique criteria for the selection of the buyers. As the data shows, the relationship between the buyer and seller plays an important role in the outcome.

We are considering the scenario that the partner company, (further denominated by company) is participating in such auctions. Each auction is characterized by multiple variables: the internal information of the proposal provided by the company and the external information relative to the state of the market, that is not provided by the company and needs to be collected. The internal information of each proposal provided by the company consists of: a global margin for the proposal, denominated by K, the date at which the auction took place, an indicator of which client the company engaged in business with, ClientID, and an indicator of whether or not the auction was won by the company. The global margin K corresponds to the percentage increase of the proposal over the base cost of the service being provided and accounts for the following factors: the profit margin of the proposal (P) , overheads (O), which corresponds to the cost associated with managing the project and the human resources contribution, and the risk factor (R), which accounts for the risk associated with the proposal:

$$K = P + O + R \tag{1.1}$$

Each one of these factors is given as a percentage of the baseline cost of the service/product. For the considered industry K typically ranges from 20% to 60%, corresponding to around 15% of overheads (O) and 5% of risk factor (R).

The data provides information for the time periods between 2018 and 2020. As mentioned before, we consider only tenders with one phase (and therefore the winner is decided after the first round of proposals). Moreover, we use only the values proposed by the company. In particular, when the proposal from the company is not the chosen one, we do not have information regarding the winning bid. When the company has the winning bid, we do not have information regarding the other competitors. Hence, we only have partial information regarding each tender.

This project presents an unique opportunity because, having real data provided by EQS differentiates this study from most other studies, such as [1] and [2]. In those cases, the data used for analysis consisted of data from public tenders, where the researchers only had at their disposal information about winning bids. On the other hand, in our study, the provided data also includes information where the partner company was not successful in the tender which, in turn, constitutes a new type of data that has not been studied in the previously mentioned works. This allows for a new look at the market impact on the proposals. Furthermore, the close proximity with actual managers working in the company allows for constant feedback regarding the obtained results and the real world implications of some of the relationships discovered in this study.

This study also provides a way of extracting useful (practical) information from the data collected for the company. When constructing a proposal the managers typically focus on the industry standard variable, the global margin K, defined in equation (1.1), that corresponds to the amount added to the proposal over the base cost. This is one of the most important parameters to be used in the development of a new proposal. Others being the market conditions at the time of the proposal and the relationship the company has with the client. These factors are calibrated by the company managers based on their past experience and expertise. There is a consensus among the managers that a more data driven approach, that could quantify or categorize the company-client relationship and the market state, would most certainly help managers improve the company's performance in future tenders.

1.2 Objectives

The project has two main objectives. One is to create a decision support tool for the company managers to optimize the performance in tenders. The other is to study the influence of the sate of the market on the company's performance.

Although the company recognizes the state of the Market as a possible relevant variable in the outcome of the tenders, the shared data does not provide any information regarding this subject. For this reason, and in a joint collaboration with the company, the first step is to define possible financial and economic variables that can be useful in the definition of a new variable, that we call ”Market Index”, that can summarize the state of the market at any given quarter, this will constitute the explicit approach of the study of the market influence.

Additionally, and for reasons that will become clearer later on, we also consider an implicit approach to the influence of the market conditions, using a Hidden Markov Model (HMM). The idea about this model is to assume that there is an underlying process modelled by a Markov Chain, whose states are not directly observable. In addition, it is assumed that there is another process, observable, whose behavior ”depends” on the state of the Markov Chain. In our setting, the observable process is the Market Index changes every quarter, and the states of the Markov Chain describe the state of the economy. We call the ”explicit approach” when we use the Market Index as explanatory variable, and whenever we use the HMM states, we call it the ”implicit approach”.

For this study, the company gathered the results from past tenders of a specific type of industrial service they provide. The data consists of tender results from 2018 until the second quarter of 2020 with several variables concerning the parameters of the proposals, the clients in question and whether or not the company was successful in each case. Then, we construct the Market Index variable to measure the state of the market with constant feedback from the company managers.

The next step in the project concerns the creation of a decision tool that can predict the success probability of a future tender, given the specific information of the tender, namely, the global margin and the client in question, as well as the state of the market at the desired time. In our case, after testing multiple models from Neural Networks, Random Forests, XGBoost algorithm and Logistic Regression, we decide to follow the Logistic Regression approach. When considering the Logistic Regression models we decide to follow two separate approaches, a frequentist and a Bayesian one. For both these two approaches we develop models considering the explicit and the implicit study of the market influence on the tenders.

After computing all the relevant models, we compare their performances using the AUC metric, area under the Receiver Operating Characteristic curve, for several train/test partitions and we are able to derive some major conclusions from the final results.

1.3 Outline

In Chapter 2 we start studying the influence of the market with the explicit approach, by creating an economic index in Section 2.1, and then developing algorithms to predict its future values in Section 2.2. In Chapter 3 we analyse the data, paying special attention to the client-company relationship in Section 3.1, where we develop a process to categorize the relationship between the company and its clients. Then we construct the Logistic Regression models in Chapter 4, using two different approaches, a Frequentist and a Bayesian one, considering the explicit study of the market influence only, i.e. using the MI variable. In Chapter 5 we study the market influence but now with the implicit approach where we use Hidden Markov Models to determine the market's hidden states. Then, we compute similar models to the ones in Chapter 4 that, in this case, incorporate the market states instead of the market index. The analysis of performance results from all the created models and conclusions are presented in Chapter 6, as well as, a description of the decision support tool implemented in Python that the company managers can use in day-to-day operations.

Chapter 2

Market Index

2.1 Index Construction

In the explicit approach to the study of the market influence on tender performance we construct a Market Index (MI) and then define models to predict its future values. The sector in which the company, EQS Global, operates is a very specific market which can be considered as a combination of several fields. This means that it is essential to construct a market index that can reflect the environment in which the company operates. After several discussions with our industry partners it was decided to use four indexes from the EuroStat [3] public platform, that gathers economic and social indexes regarding the European region. These indexes are described more thoroughly in the appendix. Namely, the Producer Price in Industry Index (Pr. Price), Producer Price in Construction Index (Co. Price), Turnover Index (Turn) and Labour Input in Construction Index (Lab. Inp.), that can be found in [3] and are updated every quarter. The index is then constructed by using Principal Component Analysis (PCA), as in [4]. By recurring to the **sklearn** Python package [5] and using the routine PCA, we can compute the principal components given the data from the indexes with values from 2006 until the second quarter of 2020. The obtained 4 principal components explain 86.2%, 9.4%, 4.2% and 0.2% of the total variation in the data, respectively. Therefore, considering the first principal component, it explains over 86% of the total variation in the data and thus, it is a good representative of the indexes considered. The market index (MI) is then defined as the first principal component obtained and has the following coefficients:

Table 2.1: Market Index coefficients

Index	Pr. Price	Co. Price	Turn	Lab. Inp.
Coef.	0.263	0.449	0.537	0.664

The four indexes are all subject to the same scaling rules. By definition, the EuroStat platform applies the same formula when defining all the gathered indexes. The year of 2015 is considered the base year for the calculation of the index values. For example. considering the Producer Price in Industry Index, then, the mean of the quarterly values of the Production Price for the year of 2015 is set to be 100 and the values of the Producer Price in Industry Index in every quarter are calculated taking into consideration that the mean of the 2015 values equals 100. This way, all the indexes considered have the same scale.

Thus, we can analyse the first principal component, and consequently the market index as a weighted average of the four components. We see that, in the market index the biggest factors influencing the index are Labour Input in Construction, Turn Over Rate, Producer Price in Construction and then Producer Price in Industry, respectively.

2.2 Index Forecast

We have MI values until the second quarter of 2020. Once we need to make predictions on tenders in the first and second quarters of 2021, it is then necessary to develop a method to estimate future values for the MI. In the literature, estimation and prediction of such market indexes that involves the construction industry are very closely related to estimation of Tender Price Index (TPI), as in [6]. By definition, Tender Price Index measures the movement of prices in tenders for building contracts in the public sector in a respective region. It doesn't, however, include contracts for housing, engineering and maintenance works. In the literature various methods to predict Tender Price Index values are applied, from Regression Analysis(RA) [6], [7], [8], Time Series(TS) [1], [9], [10], Vector Error Correction(VEC) [8], [11], Fuzzy Sets(FS) [12], Structural Equations(SE) [13] and Neural Networks(NN) [14]. The general consensus among the researchers is that an integrated model, as presented by [6], is the best approach and the most reliable alternative, which we decide to follow. Taking a closer look at the example in [6] one can see that the final presented model corresponds to a combination of a time series Autoregressive Integrated Moving Average (ARIMA) [15] model with a Regression model, where macroeconomic and other construction based variables are used. The two models are combined through an affine linear combination.

Regression Model

For the Regression Analysis we gather several other macroeconomic variables from the EuroStat platform [3]. As the impact of certain variables on the MI might be delayed for a few quarters we need also to consider, in the pool of possibly relevant variables, the one, two and three quarters lagged variants of the variables already gathered. We then adopt an automated stepwise procedure in order to eliminate those variables with negligible impact on the MI. The selected variables are chosen based on the p-values of each feature.

Table 2.2 summarizes the stepwise procedure of the multivariate Regression Analysis. Variables are added or removed from the regression model step by step. The variables selected to incorporate the model are the following:

- Production in Service Index (PSI)
- Producer Price in Service Index (PPSI)

- Labour Input in Industry Index with 3 quarters lag (LII3)
- Interest Rates with 3 quarters lag (IR3)
- Production in Construction Index (PCI)
- GDP Index (GDPI)

In Table 2.2 we can see the 6 variables that were added to the model (there were no variables removed), as well as, the R-Squared and Adjusted R-Squared [16]. We also have the values for the Bayesian Information Criterion (BIC) [17]. At last, we have the p-value from the t-test for each variable.

Table 2.2: Summary table of the stepwise procedure of multivariate regression

Step	Variable Added	R^2	R^2_{adj}	BIC	p-value
1	Prod_Serv_Index	0.9604	0.9597	266.68	7.647e-39
2	Prod_Prc_Serv_Index	0.9669	0.9656	260.86	2.407e-3
3	Lab_Inp_Index_3	0.9788	0.9776	240.31	2.046e-6
4	Int_Rates_3	0.9869	0.9858	218.06	1.157e-6
5	Prod_Cons_Index	0.9892	0.9880	211.49	2.254e-3
6	GDP_Index	0.9909	0.9897	205.92	4.040e-3

From the selected variables, only the Interest Rates variable has a different scale. The remaining variables have the same scale as the variables used for the construction of the MI, i.e. the mean of the 2015 values is set to 100 and then the values are recalculated accordingly. On the other hand, the Interest Rates variable contains the absolute values of the interest rates registered in that time period, that range from -0.5% to 5%.

Thus, the resulting model is given by the following expression:

$$\begin{aligned}\widehat{Y}_{MI} = & -243.4089 + 0.6425X_{PSI} + 3.3905X_{PPSI} - 0.3808X_{LII3} \\ & + 1.6982X_{IR3} - 0.2423X_{PCI} + 0.9369X_{GDPI}\end{aligned} \tag{2.1}$$

with X_{PSI} denoting the values for the PSI and similarly for the other variables, and $\widehat{Y}_{MI}$ denoting the estimate of the MI.

Then, considering the task of determining the future values of the explanatory variables of the model in equation (2.1), taking into consideration the problems identified by Yuu in [18], we improve on the method in [6]. In [6] the prediction of future values for the explanatory variables is derived by the growth rate of the historic periods of each variable and then extrapolated for the next two quarters.

We improve on this method by using a stochastic time series modelling technique known as Auto Regressive Integrated Moving Average (ARIMA) [15]. An Auto-Regressive (AR) model expresses the present value of a process as a linear combination of past values plus a random stochastic term representing uncorrelated forces acting on the system (white noise). $(X_t)_{t\in\mathbf{N}_0}$ is a stationary autoregressive process of order p (abbreviated AR(p)) if

$$X_t = \psi_1 X_{t-1} + ... + \psi_p X_{t-p} + Z_t \tag{2.2}$$

where $\psi_1, ..., \psi_p$ are constants and $(Z_t)_{t\in\mathbf{N}_0} \sim N(0, \sigma^2)$, is usually called white noise process. The elements of the series, $(Z_t)_{t\in\mathbf{N}_0}$, are independent and identically distributed, with mean zero, finite variance σ^2 and uncorrelated.

Using the backward shift operator denoted by B and defined as $B^d X_t = X_{t-d}$, we can rewrite equation (2.2) as,

$$\Psi(B)X_t = Z_t \qquad \text{with} \qquad \Psi(B) = 1 - \sum_{i=1}^{p} \psi_i B^i \tag{2.3}$$

A Moving Average (MA) model expresses the present value of a process as a linear combination of white noise variables. The process $(X_t)_{t\in\mathbf{N}_0}$ is a stationary moving average process of order q (abbreviated (MA(q)) if

$$\begin{aligned} &X_t = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-p} \qquad \text{or, using the backshift notation} \\ &X_t = \Theta(B)Z_t \qquad \text{with} \qquad \Theta(B) = 1 + \sum_{j=1}^{q} \theta_j B^j, \end{aligned} \tag{2.4}$$

where $\theta_1, ..., \theta_q$ are constants.

A stationary autoregressive moving average (ARMA) process, $(X_t)_{t\in\mathbf{N}_0}$, of order p and q, abbreviated as ARMA(p,q), can then be defined as follows:

$$\begin{aligned} &X_t = \psi_1 X_{t-1} + ... + \psi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-p} \qquad \text{or,} \\ &\Psi(B)X_t = \Theta(B)Z_t \quad \text{with } \Psi(B) = 1 - \sum_{i=1}^{p} \psi_i B^i, \text{ and} \quad \Theta(B) = 1 + \sum_{j=1}^{q} \theta_j B^j \end{aligned} \tag{2.5}$$

where $\psi_p \neq 0, \theta_q \neq 0$.

If the process in question is non stationary then we can attempt to remove the trend component by differencing the series until the transformed observations resemble a realization of some stationary time series. To this extent, the difference operator is defined as

$$\nabla^d Y_t = (1 - B)^d Y_t \tag{2.6}$$

Then, $(Y_t)_{t\in\mathbf{N}_0}$ is an ARIMA(p,d,q) process if

$$X_t = (1 - B)^d Y_t \tag{2.7}$$

is an ARMA(p,q) process (with d a non-negative integer that indicates the number of differencing steps). This definition means that $(Y_t)_{t \in \mathbf{N}_0}$ satisfies a difference equation of the form

$$\Psi(B)(1-B)^d Y_t = \Theta(B) Z_t \tag{2.8}$$

where $\Psi(z) \neq 0$ for $|z| \leq 1$, which guarantees causality in the ARMA process.

In order to obtain the best fit ARIMA model for each feature we first use the Augmented Dickey-Fuller test [19] to check if the given series is stationary. When stationarity is not satisfied we perform differencing to the series. In cases where differencing is not enough (as it is the case for the Labour Input Index with 3 quarter lag), we apply a logarithm transformation to the series and then apply differencing to achieve stationarity. Then, having the new stationary series we use a GridSearch approach to find the best fit ARMA model for each series using as selection criteria the Akaike Information Criterion (AIC) [20].

The methods applied to each variable, as well as the p-value from the Augmented Dickey-Fuller test for the transformed series can be seen in Table 2.3. At 10% significance level all the transformed series can be considered stationary.

Table 2.3: Transformations to explanatory variables and p-value of AD-Fuller test of stationarity

Variable	Transformation	p-value
PSI	None	0.047
PPSI	Differencing(1)	0
LII3	Log + Differencing(1)	0.071
IR3	None	0
PCI	Differencing(1)	0
GDPI	None	0.003

In Table 2.3 the variables are named after the nomenclature used for the regression model equation, in equation (2.1). The transformations applied to the series range from none, to differencing in one quarter indicated by ”Differencing(1)” or to differencing in one quarter after being applied the logarithm transformation to the series, indicated by ”Log + Differencing(1)”.

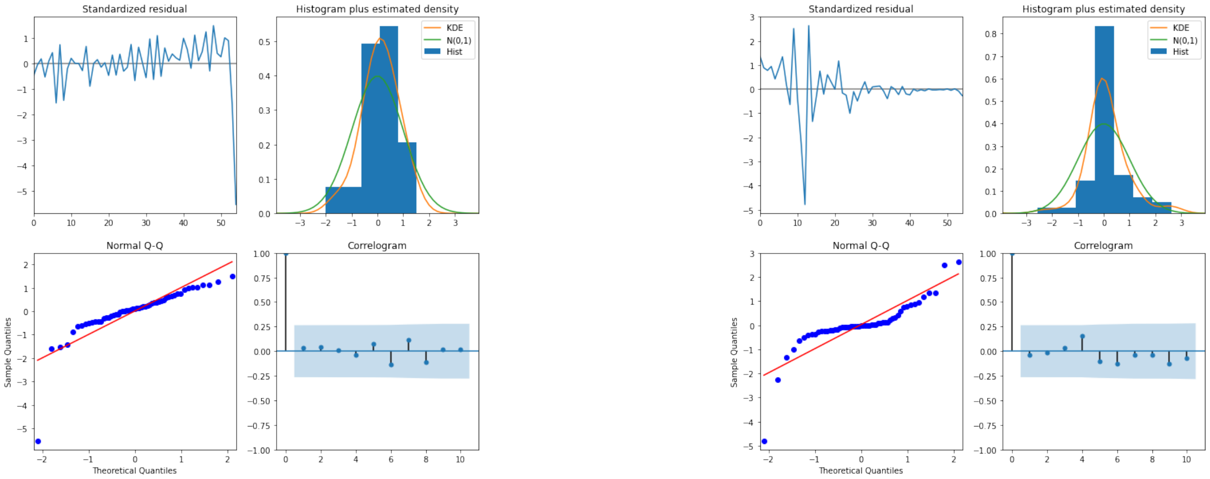
Afterwards, considering now the stationary transformed series, we can find the best fit ARMA model to each series using a GridSearch approach. The implemented algorithm cycles through multiple combinations of the model parameters, i.e., it computes ARMA(p,q) models for the series with p and q varying from 0 to 4. In each iteration it computes the AIC score for the model. The final model corresponds to the one with the lowest AIC score.

The obtained estimated ARMA models' p and q parameters can be seen in Table 2.4. After obtaining the models we still need to do some diagnostic checking to see if the models are indeed a good fit for the data. One way to perform the diagnosis is by analysing the residuals. The errors from an ideal model would resemble a realization of a white noise and they would have no autocorrelation. The Ljung-Box test is used to assess if the residuals have no autocorrelation. The Jarque-Bera test is used to check if the residuals resemble a Gaussian distribution. In Table 2.4 we have the p-values for both the Ljung-Box and the Jarque-Berta test for each models' residuals.

Table 2.4: ARMA models and Ljung-Box and Jarque-Bera tests results on the residuals of the explanatory variables

Var		PSI	PPSI	LII3	IR3	PCI	GDPI
ARMA(p,q)		(2,1)	(3,1)	(1,3)	(1,2)	(0,0)	(2,2)
Ljung - Box	Lag 1	0.765	0.991	0.698	0.675	0.817	0.923
	Lag 2	0.931	0.997	0.894	0.865	0.954	0.988
	Lag 3	0.967	0.877	0.952	0.899	0.991	0.998
	Lag 4	0.989	0.915	0.968	0.887	0.997	0.999
Jarque-Bera		0	0	0.582	0	0	0

In all the models and for all the lags from 1 to 4 (quarters), corresponding to lags up to one year, the p-values of the Ljung-Box test are larger than 0.1 and thus we accept the null hypothesis that the residuals are not autocorrelated. For the Jarque-Bera test we see that only for the model of the Labour Input Index with 3 quarters lag do we get a p-value above 0.1. Thus, only in that case do we not reject the null hypothesis that the residuals resemble a normal distribution. In all the other cases, we reject the null hypothesis and cannot assume that the residuals are normally distributed. Albeit important, the lack of normality in the residuals distribution does not have much impact in our work as the main consequence of this result is that we need to be more careful with making inference and interpreting the p-values of the models' parameters. However that interpretation is not a major concern for our study.



(a) Plots for PSI residuals

(b) Plots for LII3 residuals

Figure 2.1: Diagnostic plots of the residuals for two of the explanatory variables series

For illustrative purposes, in addition to the Ljung-Box and Jarque-Bera tests, we can also use density plots, histograms and QQ plots to check normality, and autocorrelation plots to check for autocorrelation in the residuals. In Figure 2.1 we have two examples of those plots for two different series. The plots on the left correspond to the PSI residuals, while the plots on the right are from the LII3 residuals. From left to right and top to bottom we have the following plots: Standardized residuals over time; Histogram of the residuals plus estimated density of standardized residuals, along with a Normal(0,1) density plotted for reference; Normal Q-Q plot, with Normal reference line; ACF plot (autocorrelation plot).

We see that, in both cases, the ACF plot corroborates the results from the Ljung-Box test because there is no evidence of autocorrelation in these residuals. Regarding the Jarque-Bera test we see that, for the PSI residuals, these appear to have a negative mean, the Normal Q-Q plot shows some very distant points from the reference line and the histogram shows that the residuals have a distribution that does not resemble a Gaussian distribution. These plots corroborate the Jarque-Bera test which indicates that these residuals are not normally distributed. When looking at the residuals from LII3 we see that they do not seem to be skewed, the histogram and the estimated density distribution closely resemble a Gaussian distribution. On the other hand the Q-Q plot does not offer strong indications that the residuals are normally distributed. Nevertheless, the overall analysis of these plots leads to the conclusion that these residuals are normally distributed, corroborating the results from the Jarque-Bera test for this series.

Therefore, we are able to develop methods for predicting future values for each of the explanatory variables of the regression model defined in equation (2.1).

Time Series Model

For the time series model we perform a similar process to the one described for the prediction of future values of the explanatory variables of the regression model. First, we perform Augmented Dickey-Fuller test [19] to check if the series is stationary. When the stationary is not satisfied we perform differencing and logarithm transformation to the data in order to achieve the condition. The final transformed series is obtained after differencing 2 quarters on the logarithm of the initial series values. For this transformed series the Augmented Dickey-Fuller test yields the p-value 0.001502 and thus we accept the stationarity of this series.

Afterwards, we find the best fit ARMA model for the transformed series through the same GridSearch approach on the p and q parameters of the ARMA model with the decision criterion being the AIC score. The best fit model obtained is the ARMA(2,1) model. With the Ljung-Box test we assess that the residuals have no autocorrelation and with the Jarque-Bera test we cannot assess normality in the residuals distribution. These results can also be seen in Table 2.5

Table 2.5: ARMA model and Ljung-Box and Jarque-Bera test results on the residuals of the MI time seires

Var	ARMA	Ljung-Box				Jarque
	(p,q)	Lag 1	Lag 2	Lag 3	Lag 4	Bera
MI	p=2,q=1	0.856	0.973	0.934	0.950	0

The ARMA(2,1) model obtained is defined in the following equation (using back shift notation)

$$(1 - 1.5873B + 0.7688B^2)Y_t = 0.0013 + (1 - 0.9631B)Z_t \tag{2.9}$$

where Y_t is the value of the transformed MI series in current time period t, B is the back shift operator, and $Z_t \sim N(0, 0.0003)$. In Table 2.6 we have the summary of the model's coefficients. On the one hand, all the p-values indicate that all the coefficients are statistically different from zero, on the other hand the Jarque-Bera test yields a p-value of 0 and thus we cannot assume that the residuals are normally distributed. As in the regression model, the latter does not have an impact on the remaining of the analysis.

Table 2.6: Summary of the coefficients of the ARMA(2,1) model for the transformed MI time series

Coef	Value	Std Err	z	P>$\|z\|$
intercept	0.0013	0.001	2.306	0.021
ar.L1	1.5873	0.107	14.777	0.000
ar.L2	-0.7688	0100	-7.717	0.000
ma.L1	-0.9631	0.260	-3.700	0.000
sigma2	0.0003	8.73e-5	3.295	0.001

We also compute the relevant plots of the residuals (Figure 2.3), similar to Figure 2.1. For the MI time series we see that there seems to be some negative skewness in the residuals. Additionally, both the histogram and the Q-Q plot do not show evidence of the data being normally distributed. Finally, the ACF plot shows that there is no autocorrelation in the residuals. Both these analysis are corroborated by the Ljung-Box and Jarque-Bera tests' results from Table 2.5.

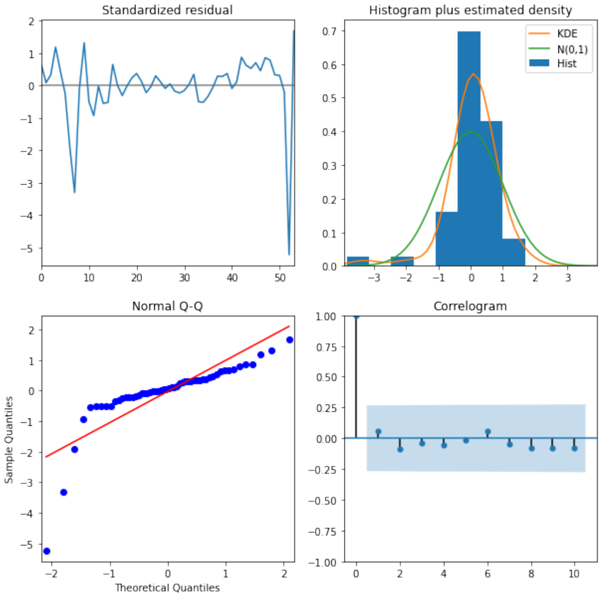


Figure 2.2: Plots for MI residuals

Integrated Model

The Regression, (RA), and Time Series, (TS), models are integrated by linear combination by considering the forecasts made by both models using the same algorithm used in [6]. The coefficient for the linear combination is obtained such that it minimizes the Root Mean Square Error (RMSE) of the forecasts of the integrated model. For that, the algorithm presented in [6] is used. First, the data is partitioned into train and test datasets with the training set containing the values until 2015Q3 and the testing set containing the remaining values until 2020Q2 (with 19 observations). Then the RA and TS models are fitted to the training data. One-step predictions are performed with each model and the errors of those predictions are computed.

The RMSE for the Regression model and the Time Series model are defined, respectively as:

$$\mathcal{R} = \sqrt{\frac{\sum_{i=1}^{19}(R_i)^2}{19}} \quad \text{and} \quad \mathcal{T} = \sqrt{\frac{\sum_{i=1}^{19}(T_i)^2}{19}} \tag{2.10}$$

where R_i corresponds to the error of the i^{th} prediction of the Regression model and T_i corresponds to the error of the i^{th} prediction of the Time Series model. The RMSE of the Integrated model, denoted by $\mathcal{IM}$ is given by $\mathcal{IM} = \beta \times \mathcal{T} + (1 - \beta) \times \mathcal{R}$ with $\beta \in [0, 1]$.

The algorithm for determining the coefficient β that minimizes $\mathcal{IM}$ is constructed by successive decimal approximations:

- Compute the RMSE of the Integrated model, $\mathcal{IM}$, for the following values of β: $\{0, 0.1, 0.2, ..., 0.8, 0.9, 1\}$ and define β_m as the minimizer within that subset of β values;

- Compute the RMSE of the Integrated model, $\mathcal{IM}$, for the following values of β: $\{\beta_m - 0.1, \beta_m - 0.09, \beta_m - 0.08, ..., \beta_m + 0.09, \beta_m + 0.1\}$ and update β_m with the minimizer within that subset of β values;

- Compute the RMSE of the Integrated model, $\mathcal{IM}$, for the following values of β: $\{\beta_m - 0.01, \beta_m - 0.009, \beta_m - 0.008, ..., \beta_m + 0.009, \beta_m + 0.01\}$ and update β_m with the minimizer within that subset of β values;

In each step of the algorithm we explore a smaller range of possible values for the coefficient of the linear combination that minimizes the RMSE of the integrated model's predictions in the test set.

After performing the algorithm, the obtained coefficient is $\beta_m = 0.545$ which means that the predictions of the integrated model have a slightly larger influence of the Time Series model (0.545) than the Regression model (0.455).

In Figure 2.3, it is possible to compare the real values for the market index (dashed line) with the predictions for the integrated model (dotted line), since 2015Q3 until 2020Q2.

We see that the values obtained with the integrated model closely resemble the actual values, especially if we consider the period from 2015Q3 to 2019Q4. Obviously the last three quarters compared, 2019Q4 to 2020Q2, are directly impacted by the COVID-19 pandemic. Therefore it is difficult to predict the substantial drop in the market index during these three quarters. Nevertheless, one can see that a portion of the drop is, indeed captured by the prediction model.

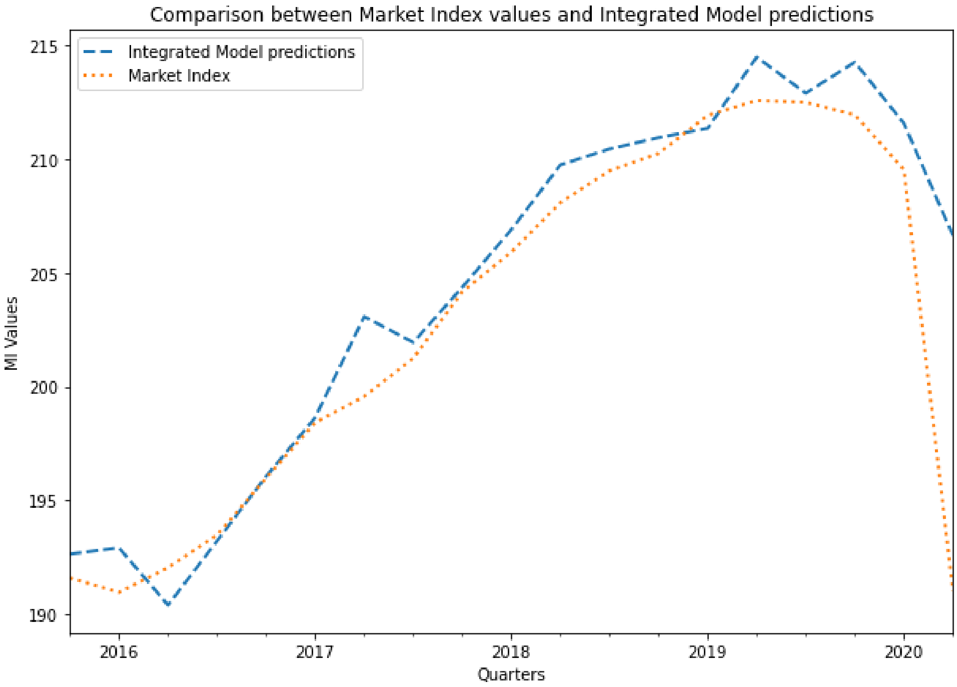


Figure 2.3: Market Index predictions

Chapter 3

Data Analysis

After contextualizing the real world environment in which the company operates and after successfully creating and estimating an adequate market index, it is important to properly analyse the data itself. Although the main objective is to study the impact of the state of the market and to develop a strategy that improves future performance, the given data not only provides some insights to each proposal's characteristics but also information about the clients. It constitutes additional information external to the market state that can influence the performance of the proposals that needs to be taken into consideration and integrated into the final model.

The provided data includes:

- Time at which the auctions took place (between 2018 and the last quarter of 2020);
- The global margin K;
- The three variables that form the global margin (profit P, overheads O and risk R);
- The identification of each individual client (ClientID);
- A binary variable, indicating if the auction was successful (Adjudicated);

In Figure 3.1 we see that in most quarters there are more unsuccessful tenders than successful ones, except for the first quarter of 2019 and the last two quarters of 2020. Additionally, we also note that there are considerably fewer observations from 2020 compared to the previous 2 years, with 23 observations for 2020, 53 for 2019 and 41 for 2018. The lack of observations from 2020 is mainly due to the fact that the data was gathered in December of 2020 and thus not only wasn't the quarter over but also some of the information from completed tenders wasn't yet available for this study.

Additionally, we see that the MI rises steadily until the last quarter of 2019 when it suddenly collapses, due to the COVID-19 pandemic. The index starts to recover on the third quarter of 2020, according to the integrated model's predictions.

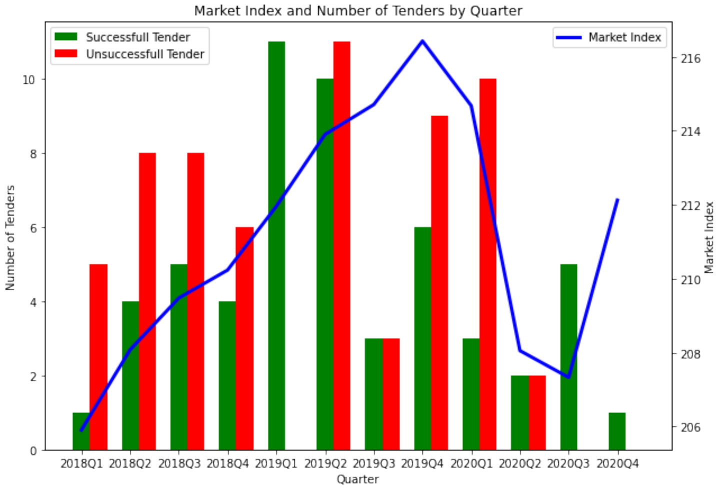


Figure 3.1: Market Index and Number of Tenders by Quarter

The proposals are defined by the value of the global margin K that, as mentioned before, corresponds to the sum of three other internal variables: profit P, overheads O and risk R. These three variables play an important role for the internal analysis performed by the company. The buyer does not have access to these variables, and the final price is characterized exclusively by the global margin, K.

In Figure 3.2 we plot the global margin as a function of time, including also, information about the outcome of the tender (successful or unsuccessful).

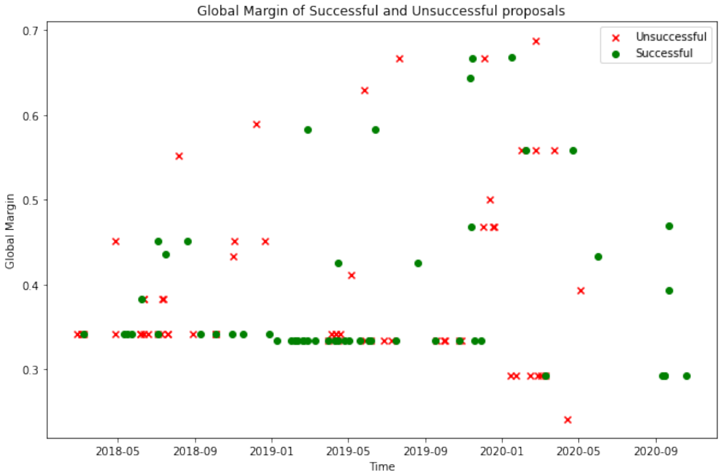


Figure 3.2: Global Margin of Successful and Unsuccessful tenders

We see that many tenders have a K value in the range $30 - 35\%$. Furthermore, we can see that the successful tenders tend to have lower global margins and also that a majority of the tenders with global margin above 50% are unsuccessful. However, there are still a considerable amount of unsuccessful tenders with low levels of K and also some successful tenders with high levels of K. This suggests that, if there is some correlation between K and whether or not a tender is successful then it is certainly negative but it is not clear if that correlation is significant or not. We can perform a point biseral correlation test on those two variables to confirm our suspicions:

Table 3.1: Results of Point Biseral Correlation between variables Adjudicated and K

Score	P-Value
-0.0404	0.6655

Then we conclude that it is not significant, showing that the dependency of the outcome of the tender and the margin is not straightforward, and further analysis is necessary.

3.1 Client Categorization

The remaining variable of interest is the ClientID variable that has a unique value for each client with which the company has engaged in business since 2018. Some of the clients have several interactions (number of tenders), shown in Table 3.2.

Table 3.2: Number of tenders by client

client1	1	client12	1	client23	1	client34	1	client44	1	client54	1
client2	2	client13	2	client24	1	client35	1	client45	1	client55	1
client3	1	client14	1	client25	1	client36	2	client46	2	client56	1
client4	5	client15	5	client26	1	client37	1	client47	1	client57	1
client5	6	client16	2	client27	1	client38	1	client48	1	client58	3
client6	1	client17	1	client28	2	client39	1	client49	1	client59	1
client7	1	client18	2	client29	1	client40	1	client50	1	client60	1
client8	15	client19	7	client30	1	client41	1	client51	2	client61	1
client9	1	client20	1	client31	1	client42	1	client52	1	client62	1
client10	2	client21	3	client32	1	client43	1	client53	2	client63	1
client11	1	client22	1	client33	1						

From analysing Table 3.2 and the values of K for each client's proposals we immediately identify several interesting properties:

1. Each client appears in the dataset an irregular number of times.
2. A large percentage of the clients only appear once in the dataset.

Additionally, when looking at the dates of the proposals we realise that there are multiple cases of identical proposals in the same quarter, whose single differentiating factor is the client in question. These remarks justify considering clusters in the data. Due to the small number of observations, (117), the number of clusters to consider is, obviously, very limited. In the experiments, various clustering techniques are performed with the objective of partitioning the clients into two, three or four categories. The careful choice of the categorization has to be made to incorporate the insights from data and create the categories that can be used on the day-to-day basis by the managers. Partitioning the clients using the k-means method [21], not only does it not improve the performance of future models but also, the partitions created are not clear or intuitive. One other alternative is to divide the clients by the number of occurrences in the dataset. Although this seems the natural approach to follow, simply dividing the clients by the number of occurrences would create too many clusters, 7. Therefore, we decide to create a new variable, denoted *Sympathy*, that describes the level of sympathy that the clients have with the company in the following way:

1. Category 0 - Unfriendly: corresponds to those clients with which the company tried to engage in business more than once and was unsuccessful every time;
2. Category 1 - Friendly: corresponds to those clients with which the company tried to engage in business more than once and was successful every time;
3. Category 2 - New: corresponds to those clients from which there isn't enough information about the state of their relationship with the company, because they only engaged with each other once (independently of the fact that they won or lost the respective auction with);
4. Category 3 - Regular: corresponds to those clients with which the company tried to engage in business multiple times and the overall outcome is uncertain, i.e. there are some cases when the company wins the auction, and in other it loses;

This is the final and best performing categorization due to several factors. In the first place, it is an understandable categorization, that can be easily described to the company managers and has an inherent basis in the business environment and business relationships. This partition of the clients is an intuitive way for the company managers to categorize their clients based on past experience and personal relationships. It is also beneficial that this categorization does not create too many categories. This categorization allows the subsequent models to have realistic relationships between the overall profit margin of a proposal and the success probability, which will be explained in further detail in the next sections.

With this partition the data is then divided in the following way:

Table 3.3: Number of observations in each client category

New	Regular	Friendly	Unfriendly
46	43	18	10

3.2 Correlation Analysis

After defining the Sympathy variable we can then proceed with the analysis of correlation between variables. Interpreting such correlations may provide some insights into the company's behaviour in some situations. Before doing so it is important to understand the type of data we are dealing with, more precisely, regarding the Sympathy variable. As it is defined, this is a categorical variable with values ranging from 1 to 4, where 1 indicates New clients, 2 indicates Regular clients, 3 indicates Friendly clients and 4 indicates Unfriendly clients. We can see that this is not an ordinal variable in the sense that the categories do not have a specific order. Thus we can create 4 dummy variables, one for each category of clients and use those variables to analyse relations between the MI and K for certain groups of clients. These four binary variables are mutually exclusive, i.e. each client can only belong to one of the Sympathy categories, which means that if, for example, the binary variable indicating the New clients has a value of 1 then the remaining three binary variables must be 0. This property needs to be taken into consideration when analysing and interpreting relations between variables.

In our case, it is important to study the relations between the continuous variables MI and K, and the binary variable indicating the success/failure of the tenders. When computing the correlation between MI and Adjudicated or K and Adjudicated we are dealing with one continuous and one binary variable. Thus we have to calculate the point biseral correlation coefficient [22], which corresponds to the value of the Pearson's product moment correlation when one of the variables is dichotomous and the other variable is metric (which is exactly our case). Considering the entire data we get the following results:

Table 3.4: Point Biseral Correlation between MI and Adjudicated, and K and Adjudicated

	Score	P-Value
K	-0.0404	0.6655
MI	-0.0199	0.8315

We see that both correlations are negative but non-significant. Although there does not seem to exist any significant correlation between these variables at first glance we can compute the correlations but now conditioning the set of observations according to the type of client. When considering all the observations we might have correlations with opposite forces for different clients that cancel each other resulting in an overall correlation coefficient that is non significant.

Table 3.5: Point Biseral correlation between Adjudicated and MI, and Adjudicated and K conditioned on different client groups

		MI	K
New	Score	0.080	-0.046
	P-Value	0.599	0.761
Regular	Score	0.184	-0.026
	P-Value	0.238	0.869
New + Regular	Score	0.121	-0.034
	P-Value	0.260	0.754

Therefore we compute the point biseral correlation coefficients between the MI and Adjudicated, and the K and Adjudicated variables conditioned on the clients categories whose values can be seen in Table 3.5. We see that, although we are able to get some correlations with p-values closer to acceptable margins, at the usual 1%, 5% or 10% significance levels, there is no evidence that the correlations are statistically significant. These results possibly indicate that there is no evidence for linear correlation between the MI or K variables and the Adjudicated variable.

Regarding the correlation between the New and Regular classes of clients and Adjudicated, the results can be seen in Table 3.6, where we see that, once again, we cannot find a significant correlation between these variables. Nevertheless, we see that the correlation value for the Regular clients is more positive than the value for New clients suggesting that the company has better performance with clients with whom a prior relationship has already been established.

Table 3.6: Correlation between the Adjudicated variable and the different client groups

Category	Score	P-Value
New	-0.1271	0.1722
Regular	-0.0431	0.6444

The absence of significant correlation between variables, as we will see, indicates the existence of non-linear relations between variables.

Chapter 4

Models

After analysing all the variables we can now proceed to the creation of models that predict the success probability of future tenders. We have, at our disposal to include in the model, several explanatory variables such as the global margin K, the MI variable, the Sympathy variable and the binary indicators for each client category. In many other works, such as [23], some other variables related to geographical region and specific characteristics of each tender were also added. However, in our case, all the tenders are from the Euro region, which was taken into consideration by gathering only European economic and social indexes when constructing the MI. The tenders are all from clients that operate in similar markets, and no additional significant information is provided by the company. Similar studies include in the models a pool of macroeconomic variables. In our case, this role is played by the MI.

For this situation, several types of models are considered, such as Neural Networks [24], Random Forests [25], XGBoost algorithm [25] and Logistic Regression [23]. The method we choose to model the probability of success of the tenders is Logistic Regression. When fitting the different models to the data, Neural Network, Random Forests and XGBoost models have poor performances. Additionally, all of these models return success probabilities between 45% and 55% for New and Regular clients while the Logistic Regression models yield probability curves that span across a wider range of values. With such a narrow difference between proposals, the former models prove to be not suitable for our case because they do not provide enough distinction between the proposals in their predictions. Therefore, we focus on the Logistic Regression approach.

4.1 Logistic Regression

For the Logistic Regression model construction, it is important to consider two different approaches. On one hand, a frequentist model, similar to [2], [23] and [26] with the additional component of robustness considered [27]. On the other hand, a Bayesian approach, where prior knowledge of the field and the company can be taken into consideration in the model. The Bayesian approach can be implemented using the PyMC3 Python package [28]. There is not an extensive literature catalog but there are still some articles that approach this topic, such as [29]. In both approaches, the base model we first consider consists of a logistic regression model given by the following expression:

$$P(Success) = \pi(X) = \frac{exp(\beta_0 + \beta_K X_K + \beta_{MI} X_{MI} + \beta_N X_N + \beta_R X_R + \beta_F X_F + \beta_U X_U)}{1 + exp(\beta_0 + \beta_K X_K + \beta_{MI} X_{MI} + \beta_N X_N + \beta_R X_R + \beta_F X_F + \beta_U X_U)} \quad (4.1)$$

where the β_i's are the coefficients associated with the explanatory variables X_i's which, in this case, correspond to the global margin variable, K, the market index variable, MI, and the sympathy indicators for each client category, New, N, Regular, R, Friendly, F and Unfriendly, U.

After the definition of the model, we then consider two different approaches in the remaining steps of the model conception, frequentist and Bayesian approaches.

4.1.1 Frequentist Approach

In the frequentist case we perform 1000 iterations of the train/test split of the data, where, for each iteration we consider a Logistic Regression model with the formula from equation (4.1) and then, use the Logistic Regression Python routine [5] to estimate the parameters of the model, via maximum likelihood.

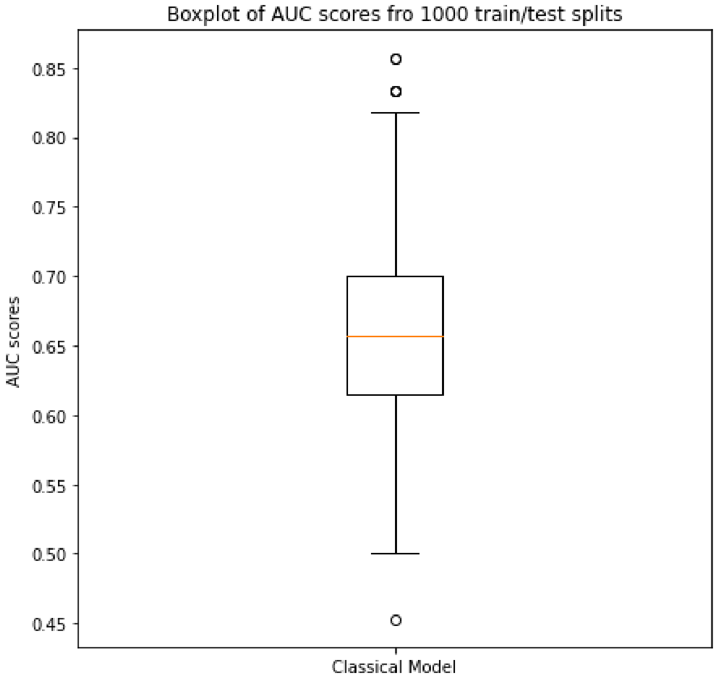


Figure 4.1: Boxplot of AUC scores for classical logistic regression models

This maximization, for each one of the train partitions, that constitutes 70% of the overall dataset, can be adjusted and calibrated with several parameters, namely in terms of the weights associated with the classes of the dependent variable (*class weight*), where we can adjust the importance of having a successful/unsuccessful tender. The algorithm used in the optimization problem (*solver*), where we can consider multiple algorithms from a coordinate descent (CD) algorithm [30], Stochastic Average Gradient descent [31] or the Broyden–Fletcher–Goldfarb–Shanno algorithm [32]. The norm used in the penalization (*penalty*) and the regularization strength parameter (*C*), where *C* denotes the inverse of the regularization strength and, similar to Support Vector Machines, smaller values specify stronger regularization. The best parameters are obtained through a GridSearch approach using the AIC as criteria for choosing the best model.

Looking at the obtained results for 1000 different dataset partitions, we realise that there is a large variance in the overall AUC score, indicating that the dataset partition has a huge impact on the final score.

One of the possible reasons for such disparity in the results lays in the presence of outliers in the training set. We recall that, in particular, the data concerning the 2020 period follows a different pattern from the previous years, and therefore its presence in the training set may have an impact on the results.

In view of this, we propose the use of a robust version of the previous algorithm. For that purpose, we use the redescending M-estimators method [33], which can be summarized as follows.

4.1.2 Robust Logistic Regression model with Redescending M-estimators method

According to equation (4.1), it follows that

$$log\frac{P(success)}{1-P(success)} = \beta'\vec{x}$$

where $\beta' \in \mathbb{R}^7$ (with $\beta' = (\beta_0, \beta_K, \beta_{MI}, \beta_N, \beta_R, \beta_F, \beta_U)$) and $\vec{x} \in \mathbb{R}^7$. Let $F(y) = \frac{e^y}{1+e^y}$, $y \in \mathbb{R}$ and $\vec{x} = (1, x_K, x_{MI}, x_N, x_R, x_F, x_U)$ (where x_k represents the variable k, and similarly for the other variables).

Then the dataset may be represented by the sample $(\vec{x_1}, y_1), ..., (\vec{x_n}, y_n)$, where y_i takes the value 1 (0) if the i^{th} observation corresponds to winning (loosing) the auction. Moreover let $p_i(\beta) = F(\beta'\vec{x_i})$.

The log-likelihood function, $L(\beta)$, based on the sample $(\vec{x_1}, y_1), ..., (\vec{x_n}, y_n)$, is given by:

$$L(\beta) = \sum_{i=1}^{n}[y_i log\big(p_i(\beta)\big) + (1-y_i) log\big(1-p_i(\beta)\big)]$$

Then it follows that the maximum likelihood estimative (MLE) of β, based on the data, is solution of the following equation

$$\sum_{i=1}^{n} \frac{y_i - p_i(\beta)}{p_i(\beta)(1-p_i(\beta))} p_i'(\beta)\vec{x_i} = 0 \qquad (4.2)$$

Where p' is the derivative of p with respect to β.

The non robust approach can be constructed using deviance. Let

$$D(\beta) = \sum_{i=1}^{n} d^2(p_i(\beta), y_i),$$

where $d(u, y) = \{-2[ylog(u) + (1-y)log(1-u))]\}^{1/2} sgn(y-u)$, with

$$d(u,y) = \begin{cases} 0 & \text{if } u = y \\ -\infty & \text{if } u = 1, y = 0 \\ \infty & \text{if } u = 0, y = 1 \end{cases}$$

In the logistic model, the values $d(p_i(\beta), y_i)$ are called deviance residuals, and they measure the discrepancies between the probabilities fitted using the regression coefficients β and the observed values. Bianco and Yohai [34], proposed robust M-estimators for the logistic model based on minimizing

$$M(\beta) = \sum_{i=1}^{n} [\rho(d^2(p_i(\beta), y_i)) + q(p_i(\beta))] \tag{4.3}$$

with weight ρ, a non decreasing and bounded function, in the family given by:

$$\rho(u) = \begin{cases} u & \text{if } u \leq c \\ 2(uc)^{1/2} - c & \text{if } u > c \end{cases}$$

and $q(u) = v(u) + v(1-u)$, with $v(u) = 2\int_0^u \rho'\big(-2log(t)\big)dt$.

Using this technique, we can compute a new, robust, logistic regression model for the same partitions considered in the classical case. After computing each model for 20 different train/test partitions, the AUC scores distributions for both cases can be seen in Figure 4.2. Contrary to the previous case where we perform 1000 different partitions and obtain 1000 AUC scores, here we just calculate the scores for 20 different partitions because too much computational time is required to obtain the robust β coefficients, and thus it is not feasible to try and perform the same number of partitions. We can see that the robust models have less variance in the AUC scores, while the 25% and 75% quantiles remaining relatively similar with the classical case, which is an indication that the robust model has overall better performance than its classical counterpart.

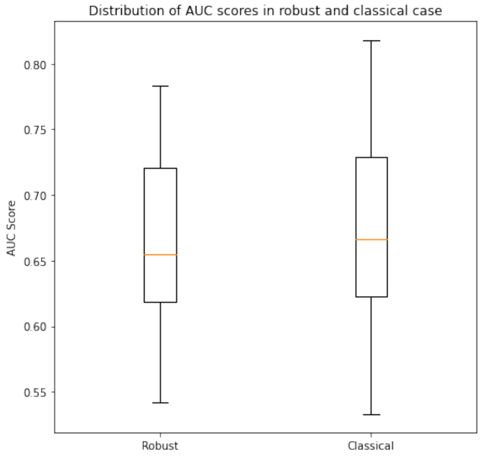


Figure 4.2: Boxplot for the AUC results of the robust and classical logistic regression models for the same 20 train/test partitions

In some of the partitions, there is no significant difference between the classical and robust model, suggesting that in those cases either there are no outliers present in the partition or their impact is not relevant. But overall the performance of the robust methods is better and hence from this point on, whenever we present results from the frequentist approach, we use their robust version.

In Figure 4.3 we plot the estimated curves for the success probability as a function of the global margin, for each class of clients, for the first quarter of 2021, assuming an MI value of 202.305 (for illustration purposes), which corresponds to the estimated value for MI in the first quarter of 2021. The coefficients of the model whose curves are presented in Figure 4.3 correspond to the mean estimated values using the values obtained in the 20 partitions considered in the robust approach.

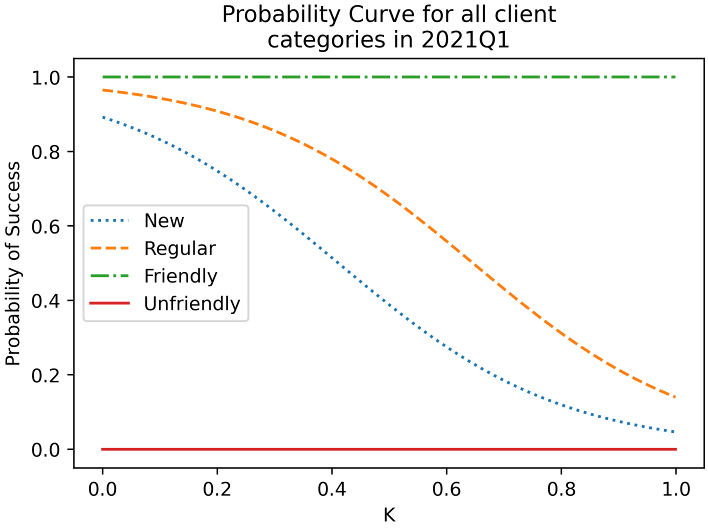


Figure 4.3: Estimated probability curves for each client category using the frequentist model for 2021Q1

First, the estimated curves for the Friendly and Unfriendly clients presented in Figure 4.3 are redundant because the outcome of the tenders for these two categories is completely deterministic. The two categories of clients that can be analysed are the New and Regular. As expected, the probability curve for clients in the Regular category is above the curve for the New one. This result confirms the prior rationale when analysing the correlation coefficients for each variable because the company is expected to have better performance in auctions for clients with which there is an established prior relationship when compared to clients with no significant prior relationship.

Although the presented plot contains values from 0 to 100% global margins, after confirmation from the company managers, the day-to-day interval of operation is in the range $30 - 60\%$ profit margins.

In Table 4.1 we can see the estimated values for each one of the β_i coefficients. Once again it is important to reiterate that the analysis of these coefficients is not straightforward because the client categories binary variables of the model are mutually exclusive. In Table 4.1 we also have the p-values from the Likelihood Ratio Test (LRT) [35]. The LRT tests the hypothesis $H_0 : \beta_i = 0$ against the hypothesis $H_1 : \beta_i \neq 0$. The test statistics is given by $LR = -2(l(\widehat{\beta}|H_0) - l(\widehat{\beta}|H_1)$, where l denotes the log-likelihood. When we test an hypothesis for just one coefficient, $LR \sim \chi^2_1$.

Table 4.1: Mean coefficients and LRT p-values for the robust model

Coefficient	Value	LRT p-value
β_0	10.563	0
β_K	-5.140	0
β_{MI}	-0.049	0
β_N	1.468	0.0001
β_R	2.674	0.0081
β_F	19.208	0
β_U	-15.031	0.0953

At a 10% significance level we reject the null hypothesis in all the cases. At a 1% significance level, all except β_U are statistically significant.

The β's coefficients for the client categories quantify the strength of each client category compared to the remaining categories with the condition that these categories are mutually exclusive. Nevertheless, we can compare the β values between categories in order to find interesting properties of each client type. First, the Friendly and Unfriendly coefficients have large (absolute) values, with the Friendly coefficient being positive and the Unfriendly coefficient being negative, as expected. Regarding the New and Regular clients, we see that the coefficient for Regular clients is larger than the one for New clients, suggesting that the company is expected to have better performance with clients with whom they have already an established relationship when compared with clients without any pre-existing relationship.

Additionally, according to the model, the company's performance and the MI variable have a negative relation in the sense that the company is expected to perform better when the market is more fragile compared to when the market is healthier. This unexpected situation has some interesting interpretations. When talking with company managers one idea that was corroborated was that when the market is more unstable and uncertain the company managers, taking notice of this dire external situations, adjust their proposals and end up with lower margin proposals in order to be more competitive. As a result, the company has better performance in uncertain times due to a higher willingness in offering lower margin proposals, while when the market is in a healthier state that incentive is gone and thus the company's performance ends up falling. In other words, in uncertain markets the company prioritizes the capture of the business, lowering the margins, and in healthier markets the company prioritizes the profit, increasing the margins. Regarding the global margin, K: its coefficient is negative, which was to be expected, because a proposal with higher profit margins is more expensive and thus less appealing to the clients.

4.1.3 Bayesian Approach

The other approach for model creation consists on taking advantage of Bayesian statistics theory.

For estimation of this model we use the Python package PyMC3 [28]. In this package, one can construct Bayesian Logistic Regression models and define specific prior distributions for the coefficients. This Python package has a unique modeling process that generally follows the following steps:

1. Encode a probability model by defining the following:
   (a) The prior distributions that quantify knowledge and uncertainty about the β parameters.
   (b) The likelihood function that combines the parameters with the data according to the specification of the logistic regression.
2. Analyze the posterior by sampling from the posterior using Markov Chain Monte Carlo (MCMC) methods [36].
3. Check the model using various diagnostic tools.
4. Generate predictions.

The resulting model can be used for inference to gain detailed insights into parameter values as well as to predict outcomes for new data points.

For logistic regression, the likelihood contribution from the i^{th} observation is binomial and given by:

$$L(y_i|x_i,\beta) = \pi(x_i)^{yi}(1-\pi(x_i))^{1-y_i} \tag{4.4}$$

with $\pi(x_i) = P(y_i = 1|x_i, \beta)$ given by equation (4.1). Then the likelihood function over the data set of n observations is:

$$\begin{aligned}\prod_{i=1}^{n}\Bigg[\left(\frac{e^{\beta_0+\beta_K x_{K_i}+\beta_{MI}x_{MI_i}+\beta_N x_{N_i}+\beta_R x_{R_i}+\beta_F x_{F_i}+\beta_U x_{U_i}}}{1+e^{\beta_0+\beta_K x_{K_i}+\beta_{MI}x_{MI_i}+\beta_N x_{N_i}+\beta_R x_{R_i}+\beta_F x_{F_i}+\beta_U x_{U_i}}}\right)^{y_i}\times \\ \times\left(1-\frac{e^{\beta_0+\beta_K x_{K_i}+\beta_{MI}x_{MI_i}+\beta_N x_{N_i}+\beta_R x_{R_i}+\beta_F x_{F_i}+\beta_U x_{U_i}}}{1+e^{\beta_0+\beta_K x_{K_i}+\beta_{MI}x_{MI_i}+\beta_N x_{N_i}+\beta_R x_{R_i}+\beta_F x_{F_i}+\beta_U x_{U_i}}}\right)^{1-y_i}\Bigg]\end{aligned} \tag{4.5}$$

Although there are many possible options for the prior distributions of the unknown parameters β_j, for simplification purposes let's consider one of the most popular choices, the normal distribution:

$$\beta_j \sim N(\mu, \sigma_j^2) \tag{4.6}$$

where $\mu = 0$ and σ is usually chosen to be large enough to be considered as non-informative. Common choices being in the range from $\sigma = 10$ to $\sigma = 100$.

Then the posterior distribution is derived by multiplying the prior distribution over all the parameters by the full likelihood function, so that:

$$\begin{aligned}p(\beta|Y,X) = \prod_{i=1}^{n}\Bigg[\left(\frac{e^{\beta_0+\beta_K x_{K_i}+\beta_{MI}x_{MI_i}+\beta_N x_{N_i}+\beta_R x_{R_i}+\beta_F x_{F_i}+\beta_U x_{U_i}}}{1+e^{\beta_0+\beta_K x_{K_i}+\beta_{MI}x_{MI_i}+\beta_N x_{N_i}+\beta_R x_{R_i}+\beta_F x_{F_i}+\beta_U x_{U_i}}}\right)^{y_i}\times \\ \times\left(1-\frac{e^{\beta_0+\beta_K x_{K_i}+\beta_{MI}x_{MI_i}+\beta_N x_{N_i}+\beta_R x_{R_i}+\beta_F x_{F_i}+\beta_U x_{U_i}}}{1+e^{\beta_0+\beta_K x_{K_i}+\beta_{MI}x_{MI_i}+\beta_N x_{N_i}+\beta_R x_{R_i}+\beta_F x_{F_i}+\beta_U x_{U_i}}}\right)^{1-y_i}\Bigg]\times\prod_{j=1}^{7}\frac{1}{\sqrt{2\pi}\sigma_j}e^{-\frac{(\beta_j-\mu_j)^2}{2\sigma_j^2}}\end{aligned} \tag{4.7}$$

The above expression is expensive for computations and multiple integrations have to be performed to obtain the marginal distribution for each coefficient.To avoid this MCMC (Markov Chain Monte Carlo) [36] methods are used.

MCMC simulation [36] is a general method based on drawing values of β from approximate distributions and then correcting those draws to better approximate the posterior distribution, $p(\beta|y, x)$. The sampling is done sequentially, with the distribution of the sampled draws depending on the last values drawn.

Markov Chain simulation is used when it is not possible (or not computationally efficient) to sample β directly from $p(\beta|Y, X)$. Instead we sample iteratively in such a way that at each step of the process we expect to draw from a distribution that becomes closer to $p(\beta|Y, X)$.

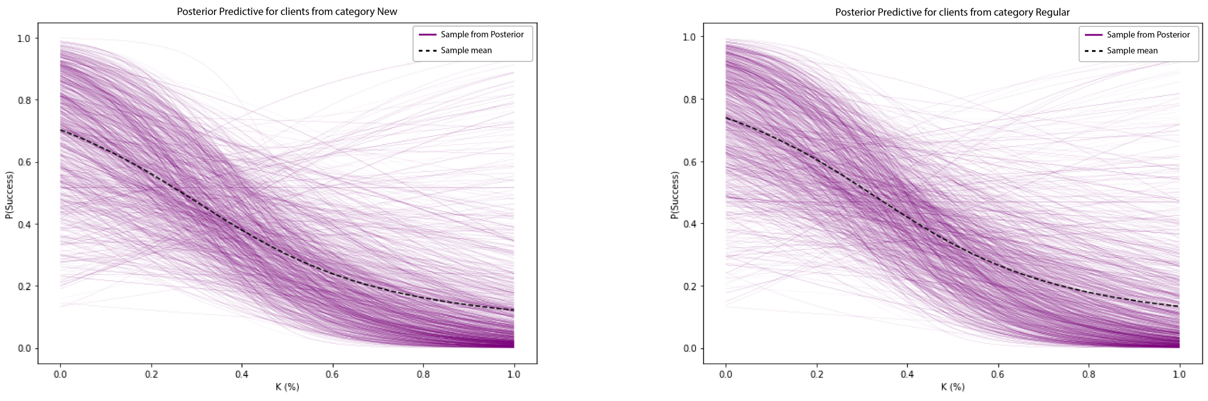
The key to Markov Chain simulation is to create a Markov process whose stationary distribution is the specified $p(\beta|Y, X)$ and to run the simulation long enough so that the distribution of the current draws is close enough to this stationary distribution.

The algorithm used in the PyMC3 package [28] is denoted NUTS (No-U-Turn Sample) [37]. Although the detailed description of this algorithm is not the main focus of this work, the algorithm can still be, very roughly, described as an improvement over the Hamilton Monte Carlo (HMC) [38] algorithm, by removing the need to set a number of steps parameter L. In turn, the HMC algorithm can be described as an improvement over the simpler Gibbs Sampler [39] and Metropolis-Hastings [40] algorithms, by avoiding the random walk behaviour present in the former. This is achieved by borrowing an idea from physics that allows the algorithm to move much more rapidly through the target distribution with the introduction of a momentum component.

The calibration phase consists of experimenting with different configurations for the prior distributions of the coefficients. The calibration can go from completely information-less priors, given by uniform distributions with large variance, to specific distributions to each coefficient. This can be done by analysing the posterior distributions of each sampled coefficient or by incorporating knowledge acquired from the company's managers experience. Additionally, one can consider the prior distributions of the coefficients as generalized t-Student whose long-tail property reduces the effect outliers can have on the posterior predictive distribution.

Then we can sample from the posterior predictive distribution and create density plots similar to the ones from Figure 4.4. Here in Figure 4.4, the plots represent the posterior predictive distributions for the New and Regular clients' categories.

Figure 4.4(a) represents 2000 samples (given by the purple lines) from the posterior predictive distribution for the New clients in 2021Q1 and the mean curve of those 2000 sample curves (given by the black dashed line) while Figure 4.4(b) represents 2000 samples for Regular clients in the same time frame. The means of the two categories appear to be relatively similar and by analysing the density of the 2000 curves in each case one can say that there is quite some volatility and uncertainty in both cases. The definitive proof of high volatility is the fact that some of the curves have positive slope which indicates that the sample value for the $\widehat{\beta}_K$ is sometimes positive, for those samples.



(a) Posterior Predictive for New clients

(b) Posterior Predictive for Regular clients

Figure 4.4: Posterior Predictive plots for the New and Regular clients on 2021Q1

The results presented on Figure 4.4 are obtained assuming that the prior distribution is a generalized t-Student,

$$f(x|\nu,\mu,\sigma) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\sigma}\left(1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{\frac{\nu+1}{2}} \tag{4.8}$$

where ν is the degrees of freedom, μ is the location parameter and σ is the scale parameter. When $\mu = 0$, we get a standard (central) t-Student.

Regarding the hyperparameters of the prior distribution (in particular μ and σ), we consider:

- For μ: we take the mean of the coefficients obtained in the 20 different train/test split iterations, i.e. for the location parameter we use the values in Table 4.1.
- For σ: we take $\sigma = 10$ in order to provide some freedom and uncertainty to the model.

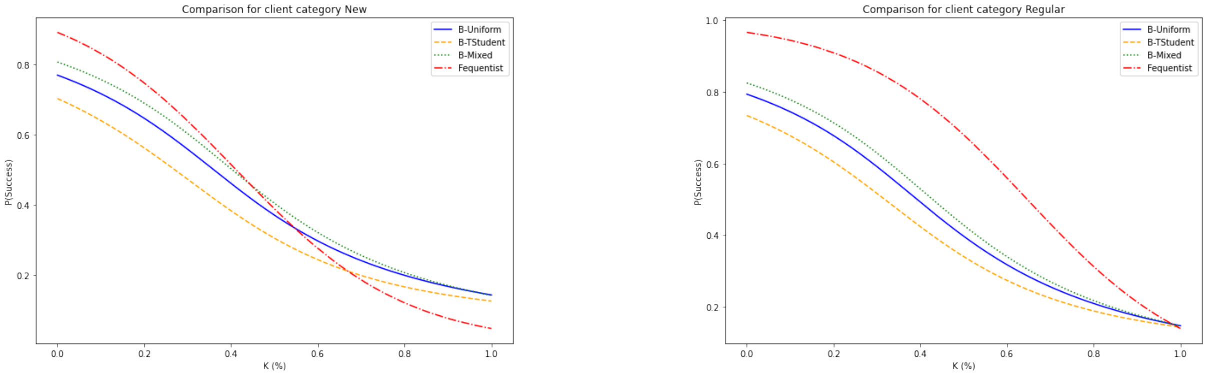
Other distributions are considered and tested for the priors, such as Uniform distributions centered on the values used for the location parameter in the case of the generalized t-Student, as well as, a combination of Uniform, t-Student and Chi-Squared distributions for some of the coefficients.

In the Uniform case, the prior distributions are all Uniform distributions centered on the coefficient values used for the frequentist model and range 10, i.e. taking the example of the prior distribution for β_{MI}, it is $Uniform(\widehat{\beta}_{MI} - 10, \widehat{\beta}_{MI} + 10)$, where $\widehat{\beta}_{MI}$ corresponds to the value of β_{MI} in the frequentist model. The same methodology is applied to the other coefficients. In the t-Student case, the prior distributions are all generalized t-Student distributions with location parameter $\mu = \widehat{\beta}_i$, where $\widehat{\beta}_i$ is the value used in the frequentist model for the coefficient β_i, scale parameter $\sigma = 10$ and degrees of freedom $\nu = n - 1$, where n is the size of the training set. For the Mixed case, the prior distribution for the coefficients correspond to a mixture of generalized t-Student, Chi-Squared and Uniform distributions, namely, β_K, β_F and β_U have Uniform priors identical to the ones used in the Uniform case, β_{MI}, β_N and β_R have generalized t-Student priors such as in the t-Student case, and β_0 has a $\chi^2_{(5)}$ distribution.

In all those cases, we have multidimensional distributions $X = (X_0, X_1, X_2, X_3, X_4, X_5, X_6)$ where the X_i's are uncorrelated. However, when comparing the performances of all those models, we find that the AUC scores, for the New clients, were relatively similar, whereas for the Regular clients the model with generalized t-Student prior distributions have slightly better results than the remaining models.

In Figure 4.5 we present the results for the frequentist model as well as for the Bayesian models (for 3 types of prior distributions). As the name suggests, B-Uniform means that we consider uniform prior distributions, B-TStudent a generalized t-Student prior distributions and B-Mixed is related with a combination of different priors, as previously explained.

The curves corresponding to the frequentist case are the same as the ones presented in Figure 4.3. The curves for the Bayesian approach correspond to the mean results from 2000 sampled curves of the posterior predictive distribution of each case.



(a)

(b)

Figure 4.5: Estimated probability curves for various tested models for 2021Q1 (The Bayesian curves correspond to the mean curve from 2000 sampled curves of the posterior predictive distribution)

For the Regular clients, (Figure 4.5(b)), the curve obtained in the frequentist approach dominates the corresponding Bayesian ones, which means that the results obtained with the frequentist model are more optimistic (the probability of success of the auction is larger than in the Bayesian approach). The Bayesian curves are quite similar, and hence, it seems that the prior distribution does not impact significantly the results.

For the category of New client (Figure 4.5(a)), the conclusions are similar although in this case, for large values of K, the frequentist model is no longer optimistic (but instead is the most pessimistic). But in the most relevant range for K (typically 30-60%), all models for this category lead to similar results.

After the discussion of these results with the company managers they suggested that the slope of the Bayesian curves is closer to the reality. In other words, the more conservative approach of the Bayesian models appears to better reflect the reality of the daily operations.

Chapter 5

Market States

After constructing two different models that estimate the success probability of proposals, we address next the study of the market influence through an implicit approach.

In this chapter we propose the use of an Hidden Markov Model (HMM) to model the market states. The motivation behind this approach comes from the fact that the models previously used for estimation of the success probability of the proposals used the variable Market Index (MI) as an explanatory variable. In the frequentist case, the estimated value of the coefficient for this variable ($\widehat{\beta}_{MI}$) is -0.049 and, when performing a likelihood ratio test [35] to check the significance of the coefficient we get that, at a 10% significance level we reject the null hypothesis that $\beta_{MI} = 0$. In the Bayesian approach, the posterior distribution of β_{MI} has a non-zero mean.

According to the experts from the company, the state of the market should influence the decision about the success/failure of the bids. This idea is vastly explored in the economical theories. For example, in the research dedicated to the real estate market, it is long established the importance of not only the global economy/business cycles but also property cycles. The economy/business cycles are described as intervals of either expansion or recession of economical activity. The property cycles on the other hand are divided into Boom, Slump and Recovery. The relationship between these cycles is still a matter of debate between scholars. The seminal work on property cycles is attributed to Homer Hoyt in his doctoral dissertation [41]. Here, we will try to create the similar notion for our particular market.

So, instead of using the MI as explanatory variable, we use a variable that describes the trend or state of the market, and we check if its explanatory power is larger than the MI, improving, this way, the quality of the proposed models. There is a vast literature on Hidden Markov Models [42], [43], [44] and it has been used in several fields such as Speech Recognition [42], Information Retrieval [43] and Gene prediction [44].

5.1 Hidden Markov Model

We consider a time discretization, with the time instances denoted by t, and the state of the system denoted by Q_t, with $q_t \in \{S_1, S_2, ..., S_N\}$. We say that $Q = (Q_t, t \in \mathbb{N})$ is a Markov chain if

$$P(Q_t = S_j | Q_{t-1} = S_i, Q_{t-2} = S_k, ...) = P(Q_t = S_j | Q_{t-1} = S_i), \quad \forall t, i, j, k \tag{5.1}$$

If, in addition, this probability does not depend on time t, then Q is an homogeneous Markov chain and we define:

$$a_{ij} = P(Q_t = S_j | Q_{t-1} = S_i) \quad \forall i, j, t \tag{5.2}$$

The matrix $A = \{a_{ij}\}_{1 \leq i,j \leq N}$ is called transition probability matrix.

In case the states sequence (Q_t) is not observable but can only be observed through another stochastic process ($\theta_t, t \in \mathbb{N}$), taking values in $\{V_1, V_2, ..., V_M\}$, then we say that (Q_t) is an Hidden Markov process. In that case, besides the transition probability matrix A, we also need the following probability matrix of observations:

$$B = (b_{ju})_{j \in \{1,2,...,N\}, u \in \{1,2,...,M\}}, \qquad \text{with} \qquad b_{ju} = P(\theta_t = V_u | Q_t = S_j) \tag{5.3}$$

Finally, as for all Markov processes, we also need to specify the distribution of the initial state of $Q : \pi = (\pi_i)_{i \in \{1,2,...,N\}}$, where $\pi_i = P(Q_0 = S_i)$. Hence a HMM is characterized by the tuple (A, B, π).

An inference regarding HMM regards the estimation of A, B and π, such that

$$(\widehat{A}, \widehat{B}, \widehat{\pi}) = \text{arg max } \{P(\theta | A, B, \pi)\} \tag{5.4}$$

where $\theta = \theta_1 \theta_2 ... \theta_T$ is the sequence of T observations of the stochastic process θ.

This optimization problem is usually solved using the Braum-Welch algorithm [45]. In addition, once the parameters A, B and π are estimated, we can then obtain the sequence of the most likely states of the Markov Chain, Q. For this, the Viterbi algorithm [46] is used.

Returning to our problem, we consider that the HMM has two states, with time discretization corresponding to one quarter. One of the states corresponds to ”Stable Market”, whereas the other corresponds to ”Turbulent Market” (in the sense that it is highly volatile). The observable process is the 1-quarter percentage change in the MI.

5.2 Python Implemmentation

We use the `hmmlearn` package in Python for the estimation of the HMM parameters (A, B, π). For the model it is sufficient to indicate the number of hidden states to be estimated and then, by calling the `fit()` function the HMM model can be trained. The inferred optimal hidden states can be obtained by calling the `predict()` function.

We decide to train the HMM with the new variable, Fluctuation, that corresponds to the percentage change between the MI from consecutive quarters. We test several possibilities for the number of states: two, three and four. The four and three states models end up with artificial data separations without a clear economic meaning.

For the HMM with two hidden states we obtain the following results:

Table 5.1: Market States obtained from the HMM

State	0	1
Mean	-0.189	0.68
Variance	7.81	0.353
Name	Turbulent	Stable

and for the estimated transition matrix A, we get:

Table 5.2: Estimated transition matrix obtained from the HMM

States	Turbulent	Stable
Turbulent	0.871	0.129
Stable	0.049	0.951

In Figure 5.1 we present the values of the Fluctuation variable and the most likely states of the HMM, colored by green (red) for the Stable (Turbulent) states.

The HMM prediction/classification for each quarter since 2006 can be compared to the Fluctuation value for that quarter, which can be seen in Figure 5.1 below.

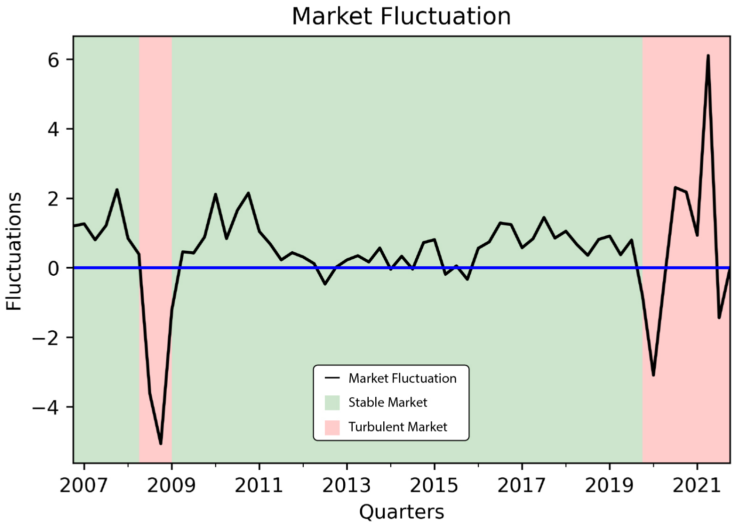


Figure 5.1: Market Fluctuation values and hidden Market States

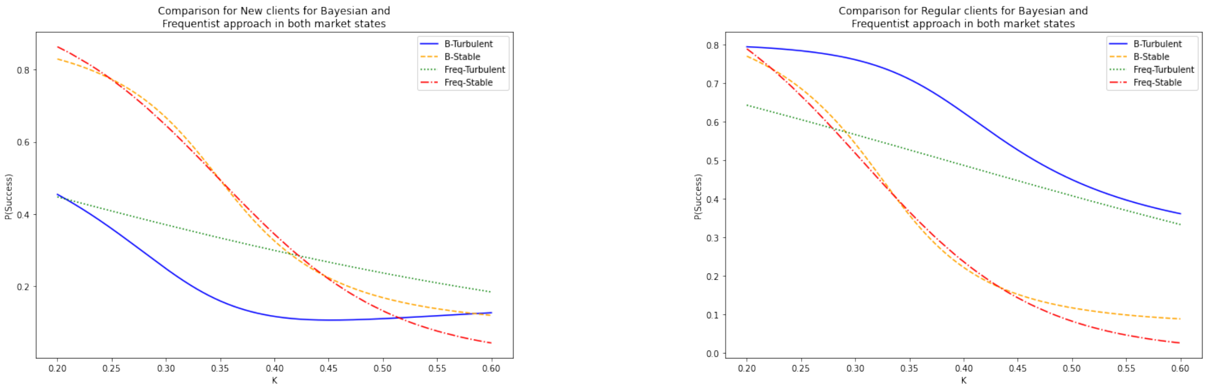
The identified states clearly divide the market into two categories, a more optimistic, stable and healthy state (Stable Market) and a more pessimistic, volatile one (Turbulent Market). The designation of Stable and Turbulent markets of each state is done based on the variance of each state and is later confirmed by observing the plot in Figure 5.1. We can clearly see that the market fluctuations, when the market in Stable, are smaller in percentage and usually do not vary drastically between positive and negative changes. On the other hand, in the Turbulent state we see that the percentage market changes are much higher in absolute value and it is common to have a positive MI change followed by a negative change indicating volatility and instability in the market at the time.

The Stable Market phases correspond to the 2006-2008Q1 and 2009-2019Q3 periods where the index registered relative stability and continuous quarter over quarter growth. The two periods of Turbulent Market state are 2008Q2-2008Q4 and 2019Q4-2021Q3 periods which represent the 2008-09 financial crisis and the 2020 COVID-19 epidemic where the index experienced high volatility and sudden decreases in value.

5.3 Models with the Market States

We designate by Market States models the group of logistic regression models, with both frequentist and Bayesian approaches, where instead of having the market index as explanatory variable, we consider the influence of the Stable/Turbulent market states. The goal is to obtain better performances with these new models which would mean that the market influence would have been captured by the market states constructed through HMM.

In Figure 5.2 we have the probability curves for the frequentist model and the mean curve obtained from 2000 samples of the posterior predictive distributions from the Bayesian model with generalized t-Student prior distributions.


Comparison for New clients for Bayesian and Frequentist approach in both market states

Comparison for Regular clients for Bayesian and Frequentist approach in both market states

(a)

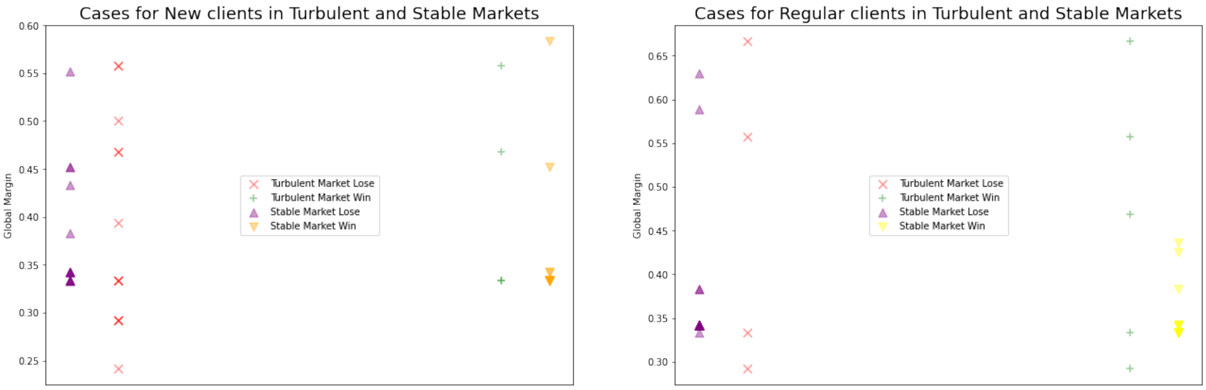
(b)

Figure 5.2: Estimated probability curves for the frequentist and Bayesian models in Stable and Turbulent Market states for New and Regular clients

For the clients in the New category the relation between Stable and Turbulent Market curves seem to be as expected, in the sense that the higher probability curves correspond to the Stable Market state while the lower probability curves correspond to the Turbulent Market state. Then the company is expected to have better performance in auctions in times of Stable Market than in times of Turbulent Market with clients with which no special relation has been created. Additionally, the frequentist model is, in general, more pessimistic than the Bayesian one, contrary to what happened with the models with the MI variable as explanatory.

On the other hand, for clients in the Regular category, the Stable/Turbulent Market relationship is the opposite, i.e., for both the frequentist model and the Bayesian model the company is expected to perform better in Turbulent Market situations than in Stable Market situations with clients. At first glance this result seems counter intuitive and quite odd so some further analysis of this situation is required.

In Figure 5.3 we can see every case for New clients 5.3(a) and Regular clients 5.3(b), where the purple and orange triangles indicate the profit margins for the observations in Stable Market that were Successful or Unsuccessful, respectively. The red cross and green plus signs indicate the profit margins for the observations in Turbulent Market that were Successful or Unsuccessful, respectively.



(a)

(b)

Figure 5.3: Case analysis in Stable and Turbulent markets for each observation using the Bayesian model

Each point represents one observation from the provided 117 observations dataset by EQS. We see that, with the New clients, for both Turbulent and Stable markets there are observations with high and low profit margins in the Win and Lose categories. For clients in the Regular category, the Stable Market observations usually have lower profit margins while the Turbulent Market observations have some instances with high profit margins. This translates into a more optimistic curve in Turbulent Market states compared to Stable Market states for the clients in the Regular category because there are some high profit successful proposals in Turbulent quarters while the same does not happen in Stable quarters.

Upon further consideration about this situation it is possible to deduce a reasoning for this relationship. In dire times the Regular clients tend to give preferential treatment to the company due to their prior relationship. On the other hand, when the market is Stable the Regular clients do not have too much pressure to choose a safer, well known company and can afford to take some risks by considering other companies. This translates into the company having better performance in Turbulent states rather than in Stable states with this set of clients affected by their prior relationship and the fact that they constitute a safe harbour in more volatile times for this clients.

Considering the New clients, the prior rationale cannot be applied. When the market is in Stable states the company is expected to perform better with these clients whereas when the market is volatile these clients are more apprehensive in engaging in business with companies with whom they do not have an already established relationship.

These important market relationships are captured by Market States models and ignored by the MI models. The market states of the HMM transmit a relation much closer to the property cycles also described in [41]. Albeit, the property cycles mentioned have three phases, Boom, Slump and Recovery, here we limit the number of states to two, Stable and Turbulent. This classification provides the models with a different relationship when compared with the MI variable. We can look at the HMM states as a filter for MI. It removes the excess noise that is not relevant for our problem, exposing thus useful relationship.

Chapter 6

Conclusions

At last, we can then develop models for all the different approaches mentioned previously. In order to compare all the models we need first to standardize the testing process. In each testing iteration the full dataset is divided into 4 different partitions in every turn corresponding to the following: $Train\_S$ containing the training observations in the Stable market category; $Test\_S$ containing the testing observations in the Stable market category; $Train\_T$ containing the training observations in the Turbulent market category; $Test\_T$ containing the testing observations in the Turbulent market category;

The partition is done in such a way that the number of observations in each partition remains constant. The models with the MI or Fluctuation variables are trained using the combined dataset $(Train\_S + Train\_T)$ and for testing $(Test\_S + Test\_T)$, while the models that take advantage of the market state classification are trained separately, i.e. two models are created, one for the Stable observations and other for the Turbulent observations and the prediction results are gathered and combined in the end in order to compare those results with the remaining models.

The frequentist model (Freq) results correspond to the results obtained with the model whose coefficients are estimated using the redescending M-estimators method described previously. Then, the (Freq F) corresponds to the frequentist model similar to the one just described but now considering the variable with the fluctuation values instead of the absolute MI values. At last, the remaining frequentist model (Freq-States) results correspond to the aggregate of the results from two frequentist models constructed each considering the Stable market observations or the Turbulent market observations, also with the robust algorithm.

Additionally, we have several Bayesian models with different configurations. In the Uniform case, the prior distributions are all Uniform distributions centered on the coefficient values used for the frequentist model and range 10, i.e. taking the example of the prior distribution for β_{MI}, it is $Uniform(\widehat{\beta}_{MI} - 10, \widehat{\beta}_{MI} + 10)$, where $\widehat{\beta}_{MI}$ corresponds to the value of β_{MI} in the frequentist model. The same methodology is applied to the other coefficients. In the t-Student case, the prior distributions are all generalized t-Student distributions with location parameter $\mu = \widehat{\beta}_i$, where $\widehat{\beta}_i$ is the value used in the frequentist model for the coefficient β_i, scale parameter $\sigma = 10$ and degrees of freedom $\nu = n - 1$, where n is the size of the training set. For the Mixed case, the prior distribution for the coefficients correspond to a mixture of generalized t-Student, Chi-Squared and Uniform distributions, namely, β_K, β_F and β_U have Uniform priors identical to the ones used in the Uniform case, β_{MI}, β_N and β_R have generalized t-Student priors such as in the t-Student case, and β_0 has a $\chi^2_{(5)}$ distribution. Finally, we have the results from the Bayesian model without the MI or the Fluctuation variables, denoted B-States, where two Bayesian models are constructed, one for each partition of the data into the Stable and Turbulent market states. These two Bayesian models have, as priors, the same distributions for $\beta_0, \beta_K, \beta_N, \beta_R, \beta_F, \beta_U$ as the ones considered in the Bayesian model with generalized t-Student prior. The AUC scores correspond to the scores obtained after grouping the predictions of both models for the respective test partition and then computing the AUC score with all the predictions.

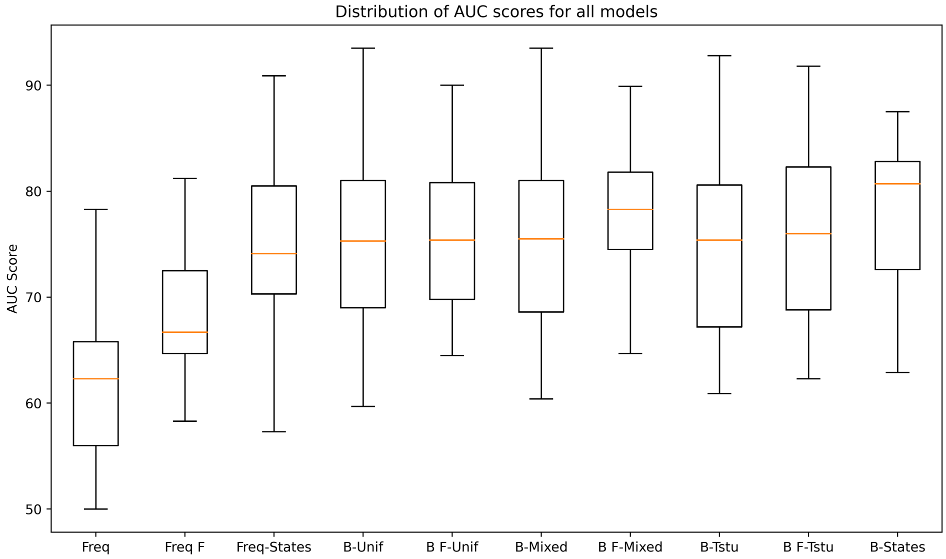


Figure 6.1: Boxplot for AUC scores of each model constructed

Before properly analysing the results it is important to reiterate that, across this study, the metric used to evaluate the performance of a given model is the AUC score [47], as it is scale invariant and classification threshold invariant, i.e. as we do not have a clear threshold for determining whether an observations should be classified as 'successful auction' or 'unsuccessful auction', it is advantageous to consider a metric that weights various probability thresholds. If we consider the Accuracy metric, the presence of a single threshold for classifying positive and negative observations makes it too restrictive, especially considering that there is no clear consensus on what the probability threshold for classifying observations should be.

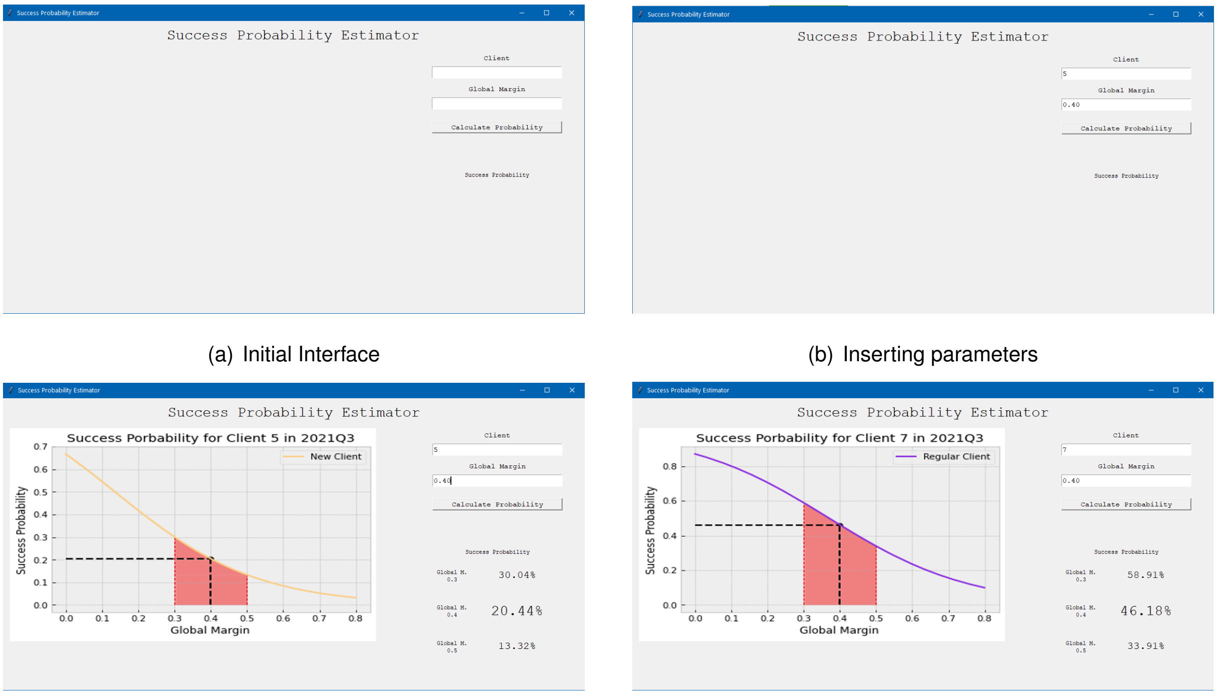
Analysing Figure 6.1, there are several conclusions that can be drawn. When looking at the frequentist models, we see that the first frequentist model, i.e. the robust logistic regression model with the MI variable is the worst performing model of all the models tested. Then, we immediately see a clear improvement in the results with the model where the MI variable is replaced by the Fluctuation variable. This indicates that, as expected, it is more advantageous to work with the differences of the economic index instead of using the nominal values, because the model ends up being influenced by the scale of the MI values when it should only be influenced by the evolution of the MI in time. Furthermore, the frequentist model with the market states is the best performing model out of the three frequentist models tested. While it has a larger variance than the frequentist model with fluctuation, the $(25\%, 75\%)$ quantiles are considerably higher in the former model and it also reaches higher AUC values in general.

Considering the Bayesian models, the first major conclusion to take is that the overall performance of the models is relatively invariant to the distributions chosen for the priors. When considering all three different sets of prior distributions, the performance of those models is very similar. Additionally, contrary to the frequentist case, when considering the Fluctuation values instead of the MI values there is not a significant, but only slight, increase in the performance of the models. Although the models results' variances decrease considerably, the $(25\%, 75\%)$ quantiles remain relatively stable when compared with the previous Bayesian models with the MI variable. Furthermore, the Bayesian model with the market states has the best performance of all the other Bayesian models tested. Although it does not reach the highest levels of AUC scores, its variance is one of the lowest registered, similar to the Bayesian model with the Fluctuation variable and with a mixed configuration of prior distributions, and the $(25\%, 75\%)$ quantiles are the highest of all the models. It has the highest median at 80% AUC score, the highest 75% quantile, and the 25% is the second highest.

When comparing the frequentist and the Bayesian models we see that the Bayesian models have better performance. The frequentist model with the market states is the best performing frequentist model and even then its performance is on par with the worst performing Bayesian counterpart. This result had already been corroborated by the company managers when they mentioned that the more conservative approach of the Bayesian models more closely resemble the day-to-day situations.

We can conclude that we successfully achieved the objective of creating a market classification procedure that divided the market into Stable and Turbulent states using HMM [48] and then, the models created with that market classification and without any additional information from the market index offer better performance than the models with information from the market index. This means that the classification procedure is able to encapsulate the influence of the market while, at the same time, by not having to include additional information in the models.

Regarding the other objectives and goals set out in the beginning on this project we can also, safely affirm that, to a certain degree, they were achieved. First, considering the business side of this work, we are able to develop a model that predicts the success probability of future tenders. Additionally, this model is integrated into a business intelligence tool in Python, along side with a graphical interface that allows the company managers to, in an easy and intuitive way, obtain the success probability for a given client at a specific quarter, as well as, the evolution of the probability of success with variations of the global margin.



Success Probability Estimator

Success Probability Estimator

Client

Global Margin

Calculate Probability

Success Probability

Success Probability Estimator

Client: 5

Global Margin: 0.40

Calculate Probability

Success Probability

Success Probability Estimator

Client: 5

Global Margin: 0.40

Calculate Probability

Success Probability

Global M. 0.3 30.04%

Global M. 0.4 20.44%

Global M. 0.5 13.32%

Success Probability Estimator

Client: 7

Global Margin: 0.40

Calculate Probability

Success Probability

Global M. 0.3 58.91%

Global M. 0.4 46.18%

Global M. 0.5 33.91%

(a) Initial Interface

(b) Inserting parameters

(c) Output for New Clients

(d) Output for Regular Clients

Figure 6.2: Decision Tool Graphical Interface

In Figure 6.2 we have four different stages of the Graphical User Interface (GUI) created. These images reflect what EQS managers observe when using the decision tool. At first, the managers are presented with an empty window (Figure 6.2(a)) with two text fields on the upper right side. In the first box they can insert the ID of the client they are interested in. In the second box they can insert the global margin value they want to use for the proposal being considered. In Figure 6.2(b), we show the example where the manager is considering a proposal with a global margin value of 0.40 (40%) for the client with ID 5 in the second quarter of 2021. After clicking ”Calculate Probability” the decision tool shows what is presented in Figure 6.2(c). In the lower right side, the manager can see the estimated success probability for the K value inserted in the previous text field, as well as, the success probability values for proposals with global margins of $K + 0.10$ and $K - 0.10$. In the left side, the decision tool shows a plot of the evolution of the success probability as a function of the global margin K for the client category of client 5, which, in this case, corresponds to a New client. The black dot and dashed black line show the K value defined by the manager and its corresponding success probability for this client. The red region constitutes the probability region between $K - 0.10$ and $K + 0.10$ which, according to the managers, corresponds to the region the managers will focus on when calibrating the proposal. In Figure 6.2(d) we have the output for a different client, client 7, that is in the Regular category. We see that, we have different probability curve and different success probability values for the same levels of K. This tool allows then the company managers to better adjust their proposals in future tenders and also have more confidence in their proposals knowing that there is a data driven model backing their decisions.

This tool was presented to the managers and board members of EQS and id expected to be incorporated in the day-to-day operations. The Python code for this project can be accessed on the following Github page [49].

Turning now to the Master's project, it is also clear that the overall project can be considered successful. The academic goal set for this work is to study the influence of the market state on profit margins proposed by the partner company in tender auctions. Moreover, there is also the purpose of considering different approaches and methods, for a comparative study. Several milestones can be pointed out.

- For the construction of the MI we look into methods to develop economic index and ultimately arrive at a principal component analysis approach. Then, in order to predict future values of this index, we analyse similar works for the estimation and prediction of the Tender Price Index (TPI). This leads us to consider different methods as, Regression Analysis, Time Series, Neural Networks, Structural Equations, among others. After exploring this topic, we arrive at an integrated model that combines a regression and a time series model with an approach described in [6]. It was implemented using the improvements suggested by Yuu in [18].
- Another very important part of this work, although at first glance it does not seem to be quite related with the main focus of the study, is the client categorization process. After receiving and performing an exploratory data analysis of the data provided by the company, we realise that the information given by the ClientID variable needs some additional work in order to maximize the relevant information this variable can bring to the models. Thus, multiple clustering algorithms are tested, from k-means to hierarchical clustering but with the final algorithm for partitioning the clients being based on decision rules that relate with the business side. The categorization serves to classify the relationship between the clients and the company based on the success/failure rate of past encounters. This classification procedures proves to be essential for the good performance of future models constructed.
- Regarding the models that estimate the success probability of future auctions, several methods are considered: Logistic Regression, Neural Networks, Random Forests and XGBoost algorithm. We are able to explore multiple approaches to this problem and, in our study, the only approach that yields interesting results is the Logistic Regression one. All the other models seem to output success probabilities between 45% and 55% for all the cases, independently of the market state or the type of clients in question. Additionally, the Logistic Regression procedure allows us to consider two different approaches, a frequentist and a Bayesian one. In the Bayesian approach, we have the opportunity to provide field knowledge to the model in the form of the prior distributions of the model's parameters. Both approaches yield satisfactory results with the Bayesian model, by creating a more conservative scenario, ends up with the best performances. This is confirmed by company managers, that state that the obtained probability curves of the Bayesian models are closer to the on the ground experiences.
- The study of the market influence is done with two approaches, an explicit and an implicit one. In the explicit case models that predict the success probability of future auctions with the MI variables are constructed. While the implicit case leads us to explore the concept of Hidden Markov Models,

in which we use the MI values to estimate two market states, a Turbulent state and a Stable state. When predicting the hidden states and comparing the results with the quarterly changes in the index, we see that the predicted states characterize well the stability/turbulence of the market in those situations.

- The final step of the project consists of comparing all the models from both explicit and implicit approaches to the market influence, and where in each case the frequentist and Bayesian models are constructed. The overall takeaway of the final analysis, already mentioned in the beginning of this section, is that the Bayesian models have a more conservative and more realistic behaviour. At the same time, the implicit approach yields better results than the explicit one with the two market states created constituting a good characterization of the market.

Overall, in this project we are given a real life problem that a Portuguese company is facing where there is a lack of information on the impact of the state of the market on the performance of its proposals in tenders. In order to address this issue and to properly study the influence of the market on the performance of the company we develop a procedure with multiple steps. In each step, a different problem needs to be addressed and a different area needs to be explored. In the end, not only are we able to construct several methods to study the market influence through different approaches, but we also create models that predict the success probability of future tenders and incorporate those models into a business decision tool that can be used by the company managers to improve their performance in future cases. Thus this project's goals were achieved both at an academic and business level with satisfactory results in both ends.

Bibliography

[1] S. T. Ng, S. O. Cheung, M. Skitmore, K. C. Lam, and L. Y. Wong. Prediction of tender price index directional changes. *Construction Management and Economics*, 18(7):843–852, 2000.

[2] J. R. Hoffmeister and E. A. Dyl. Predicting outcomes of cash tender offers. *Financial Management*, pages 50–58, 1981.

[3] Eurostat: Statistical office of the european union. URL `https://ec.europa.eu/eurostat/web/main/data/database`.

[4] V. Krishnan. Constructing an area-based socioeconomic index: A principal components analysis approach. *Edmonton, Alberta: Early Child Development Mapping Project*, 2010.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] S. T. Ng, S. O. Cheung, M. Skitmore, and T. C. Wong. An integrated regression analysis and time series model for construction tender price index forecasting. *Construction Management and Economics*, 22(5):483–493, 2004.

[7] K. Chau. The implications of the difference in the growth rates of the prices of building resources and outputs in hong kong. *Engineering, Construction and Architectural Management*, 1998.

[8] J. M. Wong and S. T. Ng. Forecasting construction tender price index in hong kong using vector error correction model. *Construction management and Economics*, 28(12):1255–1268, 2010.

[9] W. K. Wong, E. Bai, and A. W. C. Chu. Adaptive time-variant models for fuzzy-time-series forecasting. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(6):1531–1542, 2010.

[10] G. B. Hua and T. H. Pin. Forecasting construction industry demand, price and productivity in singapore: the boxjenkins approach. *Construction Management and Economics*, 18(5):607–618, 2000.

[11] H. Lütkepohl. Vector autoregressive and vector error correction models. *Applied time series econometrics*, 2004.

[12] O. S. Oshodi and K. C. Lam. Using an adaptive neuro-fuzzy inference system for tender price index forecasting: A univariate approach. In *Fuzzy hybrid computing in construction engineering and management*. Emerald Publishing Limited, 2018.

[13] M. Asano, H. Tsubaki, P. K. Bhattacharyya, and M. K. Yu. 2a-2 an application of web-based hybrid approach of neural networks and the linear regression model to the water supply forecast in tokyo district 23. In *Proceedings of the annual meeting of Japanese Society of Computational Statistics 22*, pages 27–30. Japanese Society of Computational Statistics, 2008.

[14] T. P. Williams. Predicting changes in construction cost indexes using neural networks. *Journal of construction engineering and management*, 120(2):306–320, 1994.

[15] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[16] J. Miles. R-squared, adjusted r-squared. *Encyclopedia of statistics in behavioral science*, 2005.

[17] A. A. Neath and J. E. Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

[18] K. Yu. The economics of construction price inflation: Measurement. *Output and Productivity (Doctoral Dissertation, UCL (University College London)*, 2014.

[19] W. A. Fuller. *Introduction to statistical time series*, volume 428. John Wiley & Sons, 2009.

[20] H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[21] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[22] D. Kornbrot. Point biserial correlation. *Encyclopedia of statistics in behavioral science*, 2005.

[23] C. Á. Benítez. Analytical methods for logistic tenders. Master's thesis, Universitat Oberta de Catalunya (UOC), 2019.

[24] S. O. Cheung, P. S. P. Wong, A. S. Fung, and W. Coffey. Predicting project performance through neural networks. *International Journal of Project Management*, 24(3):207–215, 2006.

[25] M. Zhang, G. Johnson, and J. Wang. Predicting takeover success using machine learning techniques. *Journal of Business & Economics Research (JBER)*, 10(10):547–552, 2012.

[26] Z. Wan and D. R. Beil. Rfq auctions with supplier qualification screening. *Operations Research*, 57 (4):934–949, 2009.

[27] D. Pregibon. Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, pages 485–498, 1982.

[28] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

[29] T. T. Nguyen. Bayesian logistic regression with pymc3, 2020. URL `https://towardsdatascience.com/bayesian-logistic-regression-with-pymc3-8e17c576f31a`.

[30] S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

[31] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

[32] C. G. Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[33] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[34] A. M. Bianco and V. J. Yohai. Robust estimation in the logistic regression model. In *Robust statistics, data analysis, and computer intensive methods*, pages 17–34. Springer, 1996.

[35] P. Sur, Y. Chen, and E. J. Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability theory and related fields*, 175(1):487–558, 2019.

[36] C. Andrieu and J. Thoms. A tutorial on adaptive mcmc. *Statistics and computing*, 18(4):343–373, 2008.

[37] M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[38] M. Girolami and B. Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[39] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[40] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[41] H. Hoyt. *One hundred years of land values in Chicago: the relationship of the growth of Chicago to the rise of its land values, 1830-1933*. Beard Books, 2000.

[42] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[43] D. R. Miller, T. Leek, and R. M. Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 214–221, 1999.

[44] M. Stanke and S. Waack. Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics*, 19(2):ii215–ii225, 2003.

[45] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[46] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

[47] A. H. Fielding and J. F. Bell. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation*, pages 38–49, 1997.

[48] L. Rabiner and B. Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1): 4–16, 1986.

[49] B. Pires. Market tender performance. `https://github.com/brunomrpires/market_tender_performance`, 2021.

Appendix A

EuroStat Indexes

The EuroStat public platform gathers several economic data publically. The data are presented with multi-dimensional tables with various selection features. In order to construct and to predict future values of our market index we gathered various indexes from the platform. The indexes used for constructing the MI variable were: Producer Price in Industry Index, Producer Price in Construction Index, Turnover Index and Labour Input in Industry Index. For the Regression portion of the integrated model, the indexes gathered were: Production in Service Index, Producer Price in Service Index, Labour Input in Construction Index, Interest Rates, Production in Construction Index and GDP Index.

All of the indexes correspond to economic data from the European Union - 27 countries, as denominated in the platform, which include all the countries from the European Union as of the exit of Great Britain. Additionally, all the data obtained contained values from 2006 until the second quarter of 2020 with values being registered every quarter.

From all of the gathered variables, only the Interest Rates variables does not follow the same standardization procedure. As mentioned in previous chapters, the Interest Rates variable contains the nominal values for the interest rates of the 27 countries in the European Union. Regarding the remaining variables, all of them were extracted after a standardization transformation was applied. The EuroStat platform allows the user to extract the nominal values of the economic variable of interest, however, it is also possible to extract the information using the available indexes. In our case, we extracted the information using the 'Index, 2015=100' parameter under the unit of measure menu. With this option we can transform the nominal values of each variable into an index. The transformation, already described in previous chapters, consists of averaging the four quarterly nominal values from 2015, and consider that new value the baseline for the index which is equal to 100. Then all the quarterly values are calculated according to this new rule.

Regarding the meaning of each index, the EuroStat platform also provides a brief description of what each variable represents.

- The GDP variable corresponds to the aggregate of the Gross Domestic Product of the 27 European Union countries;

- The remaining variables can be divided by their respective fields, from Construction, to Industry and Services. These fields are defined as follows:

  - Services - refers to all activities concerning the following: transportation and storage, accommodation and food service activities, information and communication, professional scientific and technical activities and administrative and support service activities;
  - Industry - refers to mining and quarrying, manufacturing, electricity, gas, steam and air conditioning supply, water supply, sewerage, waste management and remediation activities;
  - Construction - refers solely to construction activities;

- Variables from each sector have then specific indicators from Production, Producer Price, Turnover and Labour Input. These indicators are defined as follows:

  - Production - The objective of the production index is to measure changes in the volume of output at close and regular intervals. It provides a measure of the volume trend in value added over a given reference period. The production index is a theoretical measure that must be approximated by practical measures. Value added at basic prices can be calculated from turnover (excluding VAT and other similar deductible taxes directly linked to turnover), plus capitalised production, plus other operating income plus or minus the changes in stocks, minus the purchases of goods and services, minus taxes on products which are linked to turnover but not deductible plus any subsidies on products received.
  - Turnover - The objective of the turnover index is to show the development of the market for goods and services. Turnover comprises the totals invoiced by the observation unit during the reference period, and this corresponds to market sales of goods or services supplied to third parties. Turnover also includes all other charges (transport, packaging, etc.) passed on to the customer, even if these charges are listed separately in the invoice.
  - Producer Prices - The producer prices are also known as output prices. The objective of the output price index is to measure the monthly development of transaction prices of economic activities. The domestic output price index for an economic activity measures the average price development of all goods and related services resulting from that activity and sold on the domestic market. The non-domestic price index shows the average price development (expressed in the national currency) of all goods and related services resulting from that activity and sold outside of the domestic market. When combined, these two indices show the average price development of all goods and related services resulting from an activity.
  - Labour Input - The objective of the labour input index is to show the impact of the human resources. It takes into account number of people employed, hours worked, gross wages and salaries;