

MDSAA

Master Program in
Data Science and Advanced Analytics

**A Statistical and Machine Learning Approach for Assessing the
Impact of Financial Research Reports on Clients' Trading
Behaviour**

Diogo Tomás dos Santos Peixoto

Dissertation

Presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**A STATISTICAL AND MACHINE LEARNING APPROACH FOR
ASSESSING THE IMPACT OF FINANCIAL RESEARCH REPORTS ON
CLIENTS' TRADING BEHAVIOUR**

by

Diogo Tomás dos Santos Peixoto

Dissertation as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialisation in Data Science

Advisor: Professor Mauro Castelli

Co-Advisor: Angela Pimentel

November 2023

STATEMENT OF INTEGRITY

I declare that I have conducted the present academic work with integrity. I confirm that I have not resorted to plagiarism or any other form of improper use of information or falsification of results during the process of elaborating this work. I also declare that I am aware of the Rules of Conduct and the Code of Honor of the NOVA Information Management School.

Diogo Tomás dos Santos Peixoto

Lisbon, November 2023

ABSTRACT

BNP Paribas is a market maker that buys and sells securities on its account. Market makers play a crucial role in the market by facilitating securities trading and providing liquidity. BNP has a research team that analyses different securities and makes investment suggestions, which are grouped in a document called a research report. Those are sent to BNP's clients, aiming to lead them to perform a transaction with the bank. This master dissertation seeks to determine the impact that BNP Paribas research reports containing bond trade suggestions have on clients' trade behaviour, particularly in requesting a quote (RFQ) and, subsequently, in performing a transaction. Based on a review of the literature, a problem that requires an explanation or a prediction answer leads to some differences in the model building process and the models that are advisable to be used. This research problem fits into the explanatory paradigm. The study starts with the search for variables associated with a higher likelihood of requesting an RFQ. The logistic regression model is used as an explanatory tool and assessed with standard statistical measurements like goodness-of-fit tests and residual analysis. The random forest model is then compared with the logistic regression using probability metrics rather than threshold ones aligned with the explanatory nature of the research question. Results show similar performance between the two models, with the logistic regression model ultimately chosen for its probabilistic framework. The logistic regression model answers the research question, showing that BNP clients who download a research report are 12% more likely to request an RFQ. However, the reports seem to not influence whether they choose to trade with BNP or another market maker. Another noteworthy finding is that clients who previously purchased a specific bond security before the first time it was suggested by the bank are 36% more likely to request an RFQ. This work also addresses the challenge of establishing causation, acknowledging the complexity of attributing the observed increase in RFQ requests directly to report downloads. Recommendations for future studies involve exploring causality through randomized control trials or other advanced methods. In conclusion, this research underscores the importance of both having good business knowledge for assessing the right variables and technical expertise in modelling them, with an emphasis on knowing whether the research question falls within an explanatory or predictive paradigm.

KEYWORDS

Standard Statistic, Machine Learning, Causal Inference, Explanatory Problem, Finance, Research Report, Client's Behaviour.

Sustainable Development Goals (SDG):



INDEX

1. Introduction	1
2. Literature Review	4
3. Methodology	14
3.1. Data Assembly	14
3.2. Feature Selection.....	16
3.3. Models	17
3.3.1. Standard Statistics Logistic Regression	19
3.3.2. Machine Learning	24
3.3.3. Causal Inference	27
4. Results and Discussion.....	30
4.1. Descriptive Statistics.....	30
4.2. Feature Selection.....	33
4.2.1. Independent variables included in the final model	33
4.2.2. Independent variables not included on the final model.....	37
4.3. Models Results Outcome	39
4.3.1. Standard Statistics Logistic Regression	39
4.3.2. Machine Learning	43
4.4. Business Results Outcome.....	48
5. Conclusions and Future Work	50
References.....	52
Appendix A	55
Annex B.....	65

LIST OF FIGURES

Figure 1: Bond cash flows.....	1
Figure 2: Business model studied flowchart	3
Figure 3: Probit Model	4
Figure 4: Main machine learning techniques used	5
Figure 5: Random forests performance	6
Figure 6: Relationship between bank customer information factors, and TSP and AP	7
Figure 7: Microarray variable importance	9
Figure 8: Bias-variance decomposition of mean squared error.....	11
Figure 9: Data sources and fields identification to establish the data linkage	14
Figure 10: Observational and experimental studies	16
Figure 11: In general, models that are more accurate are less explainable.....	18
Figure 12: Illustration of the logit link function's nonlinear behaviour. Linear relationships that are parallel on the logit-scale (A) are not parallel and have different slopes for a given x value on the probability scale (B).....	23
Figure 13: Popular model-agnostic interpretation techniques, classified as local or global, and as effect or importance methods.....	27
Figure 14: Example average amount sold between the treated and the control group	28
Figure 15: Histogram days difference (Trade – Opened_Only)	31
Figure 16: Histogram days difference (Trade – Downloaded_Only).....	31
Figure 17: BNP customers sector frequency	33
Figure 18: <i>T10_only</i> variable frequency and proportion bar chart against the response variable	34
Figure 19: Univariate <i>Purchased_Ticker_Before_Only</i> logistic regression model	36
Figure 20: Variable <i>report_sent_weekday</i> proportion bar chart against the dependent variable	38
Figure 21: Final logistic regression model results given by <i>Statsmodels</i> library.....	40
Figure 22: Final model logistic regression average marginal effects	41
Figure 23: Studentized Pearson residuals against fitted probabilities	42
Figure 24: Independent variables VIF values	42
Figure 25: Logistic regression model metrics results on the test dataset, trained with different class distributions.....	43
Figure 26: Random Forest feature importance results.....	45
Figure 27: PDP of 2 IV(s) using logistic regression and random forest models on the entire dataset.....	46

Figure 28: Comparison of the top 10 client numbers of RFQs matching the same trade suggestion direction against matching the opposite direction for 2022.....	47
Figure 29: Comparison of the top 10 client numbers of RFQs matching the opposite trade suggestion direction against matching it for 2022.....	48
Figure 30: <i>Tradestatus</i> variable proportion bar chart against the variable <i>downloaded_only</i>	49

LIST OF TABLES

Table 1: The two cultures to analyse data as defined by Breiman, L. (2001)	8
Table 2: Differences Between Explanatory Statistical Modelling and Predictive Analytics.....	9
Table 3: Time period reference example to perform the data source linkage.....	14
Table 4: Cramer's V Interpretation	17
Table 5: Comparison of machine learning vs. a standard statistical approach	17
Table 6: Specification of the dummy variables and the reference class	20
Table 7: Dataset class distribution	24
Table 8: Number of Observations	30
Table 9: Data variable count	30
Table 10: Final model variables meaning.....	32
Table 11: Variable <i>t10_only</i> contingency table.....	33
Table 12: Independent variables and their <i>Cramer's V</i> values	35
Table 13: Different models with respective AIC values	36
Table 14: Bond maturity date variable encode and description	37
Table 15: Variables <i>Cramer's V</i> values not included in the final model.....	37
Table 16: Train and test dataset number of observations.....	43
Table 17: Classification metrics results on the test dataset	44
Table 18: PDP values	46

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
AIC	Akaike Information Criterion
RCT	Randomized Control Trial
DV	Dependent Variable
IV(s)	Independent Variable(s)
ML	Machine Learning
MLE	Maximum Likelihood Estimation
PDP	Partial Dependence Plot
RF	Random Forest
RFQ	Request for Quotation
SS	Standard Statistic
VIF	Variation Inflation Factor

1. INTRODUCTION

BNP Paribas CIB is a *market maker*. The term market maker refers to a firm that buys and sells securities on its account. They quote two-sided markets in a given security by giving *bids* and *asks*, with *bids* being the buy price and *asks* being the sell price. Investors or clients must pay the *asking* price, provided by the *market maker*, to buy a given security. *Market makers* make money off the difference in the price *bid-ask* ($p_{bid} < p_{ask}$), also known as spread. The role of a *market maker* is not to invest in the financial markets but to provide investors with the opportunity to do so. The *market makers* provide the market with liquidity.

The ideal circumstance is when the *market maker* simultaneously finds buyers and sellers for the same financial product. However, that is not common, and hence they are compensated for the risk of retaining assets since the value of a security may drop between its purchase and sale to another buyer.

Sales teams serve as the bank's clients' point of contact. When a client wants to buy or sell a financial asset, she can request prices from *market makers* in a process called a request for quotation (RFQ).

When a customer is interested in a certain product, she can choose n *market makers* and send them an RFQ. Then, they respond with their final pricing, and the customer can either choose to purchase or sell the product with one of these n providers or do nothing. In the case of trading with BNP, the bank takes an estimated profit called client contribution (CC).

Considering the research problem this dissertation undertakes, which will be presented later, it is necessary to explain what the research team does and what bond securities are. Further information can be found in (De Franco et al., 2009).

Bonds are instruments that represent a loan from an investor to a borrower (usually a corporation or the government). Since bonds historically paid debtholders a fixed interest rate (coupon), they are referred to as fixed-income instruments. Investors can freely trade bonds between the issue date and the maturity date, after which the full amount invested must be repaid to the investor. The figure below eases understanding.

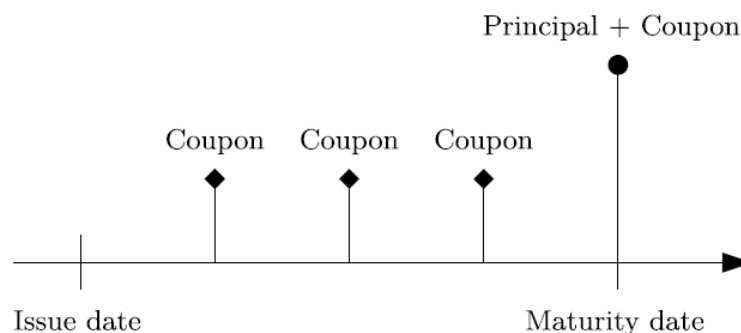


Figure 1: Bond cash flows

Note. From "Machine Learning for Financial Products Recommendation", by Barreau, B., (2020). Université Paris-Saclay (pp. 15) <https://theses.hal.science/tel-02974918/>

Bond analysts, who are part of the research team, gather and analyse data on publicly traded bond instruments and the companies that issue them. They also make investment suggestions to participants in the bond market.

Bond analysts' recommendations are likely influenced by three distinctive bond market characteristics:

- Bond investors are almost exclusively institutions, which are generally sophisticated investors, in the sense that they are more likely to have access to a variety of information sources (including their research) and have a better understanding of how to use the advice of bond analysts.
- Independent accredited rating organisations, like S&P, Moody's, or Fitch, grade bonds in exchange for payments from the companies that issue the bonds. These agencies also have priority access to information that is not generally available. Since rating agencies are alternative information intermediaries with significant reputational stakes, their disclosures can offer potentially useful information that can act as a third-party independent check on the validity and dependability of bond analysts' research.
- The value of bonds is mostly influenced by macroeconomic variables like interest rates and historical credit spreads, making their prices typically more objectively determined. Alternative bonds that have comparable cash flows and credit risk can be close alternatives to the bonds that bond analysts cover and can therefore act as pricing benchmarks.

Considering the previous points, bond analysts determine whether a company's credit fundamentals are strengthening or weakening and predict whether a company's bond instruments will perform better (or worse) than bonds with identical risks and contractual terms. That translates into an investment recommendation (i.e., buy, hold, or sell) on a certain bond security, which is defined within BNP as a *trade idea*. The *trade ideas* and their justifications are compiled in a document called a research report, such as the one shown in Figure B-1 in the annex. To ease the reading, research reports will sometimes be mentioned as just reports from now on.

The reports are attached to emails and sent to clients. The bank manages to track whether the clients open the email and download the report. Following the bank business model explained above, the reports intend to advise the clients but also to lead them to request an RFQ on the suggested security and finally to perform a trade with BNP. Please refer to the figure below. Based on that, the research problem that this dissertation tries to answer is:

What is the impact that BNP Paribas research reports with bond trade suggestions have on clients' trade behaviour?

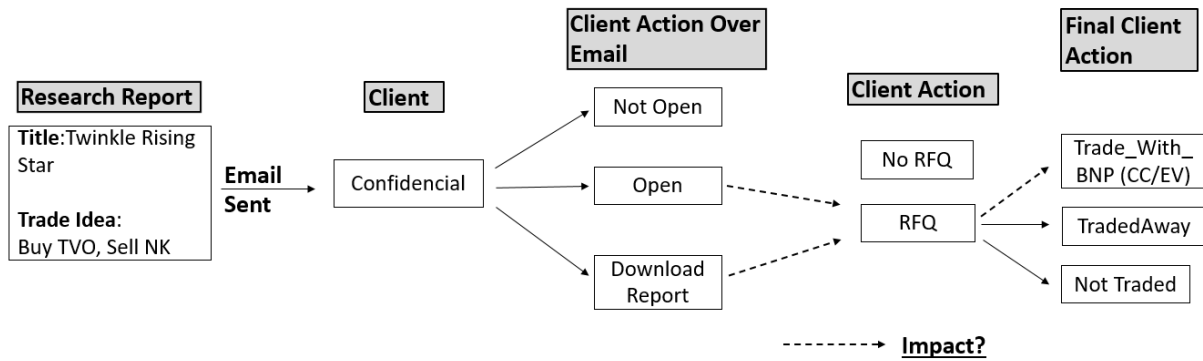


Figure 2: Business model studied flowchart

The research project has focused on studying the reports' impact on the RFQs. The impact of trading with BNP or other market makers is presented in Chapter 4.4.

The remaining document is structured as follows:

- Chapter 2 reviews the main techniques available in the academic literature to answer the research problem.
- Chapter 3 starts by presenting how and which data had to be gathered to answer the question. Then, it presents the theoretical background of the techniques used throughout the research project.
- Chapter 4 starts by describing the dataset. Then, it presents which features have been included in the model and justifies them. Afterwards, the statistical and machine learning models are introduced together with the main findings and results, including the answers to the research question.
- Chapter 5 summarises and reflects on the research process. It also makes recommendations for future work.

2. LITERATURE REVIEW

The research question presented can fit under the topic of *direct marketing*, which consists of any form of marketing that relies on reaching out to consumers directly to call for action rather than through a third party, such as mass media.

The article by Bose and Chen (2009) reviews the main models used for studying direct marketing activities. In essence, they can be grouped into statistical models and machine learning ones. The first group enlists all the regression models, including linear, logistic, and probit models. The second one entails artificial neural networks, decision trees, support vector machines, and others.

The statistical models are driven by the calculation of an approximation to a structure, which could have led to the data. In contrast, machine learning places more emphasis on the learning process that extracts rules from data.

The article (Algesheimer et al., 2010) tries to answer a similar question to this research project using a statistical model, as per the previous definition. It studies if the German eBay company's action of sending email invitations to their customers to join its community could influence participation in the community itself. That relates to the belief that communities will be effective mainly for customers who are already fans, meaning to say, customers who are already engaged and interested in the company, and therefore they would self-select themselves into the community and not by company marketing actions, such as sending an email invitation.

The study has used the Probit model. The response variable p_h is the probability of participation, which is modelled as a function of several explanatory variables, including the dummy variable $Invite_h$, which represents whether an email invitation has been sent or not to customers. The model is depicted below.

$$\begin{aligned}\rho_h = & \rho_0 + \rho_1 German_h + \rho_2 Age_h + \rho_3 Gender_h \\ & + \rho_4 Memlength_h + \rho_5 PosFB_h + \rho_6 NegFB_h \\ & + \rho_7 Invite_h.\end{aligned}$$

Figure 3: Probit Model

Note. From "The Impact of Customer Community Participation on Customer Behaviors: An Empirical Investigation", by Algesheimer, R., Borle, S., Dholakia, U. M., & Singh, S. S. (2010). *Marketing Science*, 29(4) (pp. 760) <https://doi.org/10.1287/mksc.1090.0555>

The results reveal a 22.7% higher percentage of people participating in the community if they have received the email. It also shows that both coefficients ρ_1 and ρ_3 are not significant, meaning they are German or not, and the gender features reveal no evidence of influencing the outcome with a 95% confidence level.

An important note is that the e-mail invitations were randomly distributed across the customers, allowing to directly study the impact of such action on community participation. In addition, the article states the model used could be a probit or a logit mode. That is relevant since the model used in this research project was logit, also known as logistic regression (LR).

Continuing with examples of LR applicability, the book Harrell (2015) starts the chapter on binary LR, stating that the existence or absence of a specific disease, a patient dying or not during surgery, or a customer making a purchase or not are examples of problems where LR models fit in. That is because the response variable under study is a binary one. In this research project, the request for quotation (RFQ) is also the binary response variable under study.

Zhang et al. (2021) conducted a field experiment to examine the effect of price promotions on donation behaviour. They randomly assigned participants to either a promotion condition or a no-promotion condition. Participants in the promotion condition received a discount on a product, while those in the no-promotion condition did not receive any discount. After the purchase, participants were asked to donate to a charity. To analyse the data, the researchers used binary logistic regression analysis. They treated the donation behaviour as a binary outcome variable (i.e., donate or not donate) and examined the effect of the promotion on the likelihood of donation while controlling for other factors such as age and gender. The results showed that the promotion had a significant positive effect on the likelihood of donation, indicating that participants in the promotion condition were more likely to donate than those in the no-promotion condition.

Moving to a machine learning perspective, Henrique et al. (2019) provide a literature review of machine learning techniques applied to financial market prediction. The review covers the methods used in selecting the literature, the most important articles in the field, and a classification of the reviewed articles based on the markets addressed, the type of index predicted, the variables used as inputs for the models, and the type of prediction sought. The article also discusses the challenges of predicting financial market prices using machine learning models, such as the non-linear, dynamic, and chaotic nature of financial time series data.

The article has reviewed 57 works. The most used machine learning models for predicting financial market prices are support vector machines and neural networks, as shown in Figure 4.

Table 18
Main forecasting techniques applied by each reviewed reference.

Main Method	Number of References	References
Neural Networks	42	Ang and Quek (2006), Armano et al. (2005), Ballings et al. (2015), Barak et al. (2017), Cao et al. (2005), Chang et al. (2009), Chen et al. (2003), Chiang et al. (2016), Enke and Thawornwong (2005), Fernandez-Rodriguez et al. (2000), Hájek et al. (2013), Hassan et al. (2011), Kara et al. (2011), Kamstra and Donaldson (1996), Kim and Han (2000), Kimoto et al. (1990), Krauss et al. (2017), Laboissiere et al. (2015), Leigh et al. (2002), Leung et al. (2000), Mo and Wang (2017), Oliveira et al. (2017), Patel et al. (2015), Pei et al. (2017), Rodríguez-González et al. (2011), Thawornwong et al. (2003), Tsai and Hsiao (2010), Tsai et al. (1998), Wang et al. (2012), Weng et al. (2017), Yan et al. (2017), Yoon et al. (1993), Zhong and Enke (2017), Abu-Mostafa and Atiya (1996), Donaldson and Kamstra (1999), Enke and Thawornwong (2005), Huang et al. (2005), Kim (2003), Kumar and Thenmozhi (2014), Tay and Cao (2001), Thawornwong and Enke (2004), Lahmiri (2014a), Lahmiri and Boukadoum (2015), Ballings et al. (2015), Barak et al. (2017), Bezerra and Albuquerque (2017), Chen et al. (2017), Gorenc Novak and Velušček (2016), Hájek et al. (2013), Huang and Tsai (2009), Kara et al. (2017), Oliveira et al. (2017), Pai and Lin (2005), Pan et al. (2017), Patel et al. (2015), Schumaker and Chen (2009), Weng et al. (2017), Yu et al. (2009), Huang et al. (2005), Kim (2003), Kumar and Thenmozhi (2014), Tay and Cao (2001), Lahmiri (2014b)
SVM/SVR	20	Ballings et al. (2015), Barak et al. (2017), Krauss et al. (2017), Oliveira et al. (2017), Patel et al. (2015), Weng et al. (2017), Kumar and Thenmozhi (2014)
RF/Decision Trees	7	Al Nasser et al. (2015), Hájek et al. (2013), Schumaker and Chen (2009), Weng et al. (2017), Oliveira et al. (2017)
Sentiment/Text Analysis	5	Ballings et al. (2015), Chang and Fan (2008), Chen et al. (2017), Zhang et al. (2017)
kNN	4	Bezerra and Albuquerque (2017), Donaldson and Kamstra (1999), Zhang et al. (2017), Kumar and Thenmozhi (2014)
ARIMA/GARCH	4	Chang and Fan (2008), Chen et al. (2014), Wang (2002), Wang (2003)
Fuzzy Logic	4	Leung et al. (2000), Yoon et al. (1993), Huang et al. (2005)
LDA	3	Patel et al. (2015)
NB	1	

Figure 4: Main machine learning techniques used

Note. From “Literature review: Machine learning techniques applied to financial market prediction.”, by Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). *Expert Systems With Applications*, 124, (pp. 247) <https://doi.org/10.1016/j.eswa.2019.01.012>

The same article states “Among the possible conclusions about the classification proposed here, it is to be expected that new proposed models will be compared to the benchmarks of neural and SVM networks....”. This reasoning is aligned with the continuous strive to find better algorithms’ performance, as demonstrated in the next three articles presented.

Miguéis et al. (2017) study the challenge of predicting customer response to direct marketing campaigns in the context of banking. The authors compare the performance of several response models, including logistic regression, decision trees, neural networks, and random forests. They also explore the effect of class imbalance methods on prediction performance using an oversampling method (the Synthetic Minority Oversampling Technique) and an undersampling one (the EasyEnsemble algorithm).

The Random Forests model has shown the best performance when using both balanced and imbalanced data. Its performance was better when used in combination with the EasyEnsemble method, as shown in the following figure. They highlight that these results cannot be generalized to all situations due to the specificity of each dataset.

	Random forests		
	AUC	Lift 10	Lift 20
Imbalanced	0.945	5.678	4.371
SMOTE	0.945	5.542	4.386
EasyEnsemble	0.989	7.937	4.960

Figure 5: Random forests performance

Note. From “Predicting direct marketing response in banking: comparison of class imbalance methods”, by Miguéis, V. L., Camanho, A. S., & Borges, J. (2017) *Service Business*, 11(4), (pp. 841) <https://doi.org/10.1007/s11628-016-0332-3>

This research also pinpoints the key indicators influencing responses, encompassing customers' demographic details, contact information, and socioeconomic aspects of the situation. While factors such as customers' financial status and occupational type are crucial for various banking choices, they did not significantly contribute to explaining responses within the scope of that study.

The study is relevant to direct marketing in banking because it enables banks to reduce marketing campaign costs by targeting customers who are more likely to respond positively to the offer while avoiding disturbing those who are not interested in the offer.

Feng et al. (2022) present a novel approach for bank telemarketing sales prediction called META-DES-AAP. Unlike existing machine learning-based marketing sales prediction methods that focus only on prediction accuracy, META-DES-AAP considers both accuracy and average profit maximization. The approach uses a multi-objective evolutionary algorithm to form an accuracy- and profit-based optimal base classifier pool and a dynamic-based base classifier integration method to integrate base classifiers. The predictive results include the probability of telemarketing success (TSP) and economic performance (AP), which can help marketing managers improve cost reductions. The model also supports a post hoc explanation of META-DES-AAP on TSP and AP for important factors, which helps

marketers better understand the influence and dynamics of the underlying elements that contributed to the predicted results. An example is shown in the figure below.

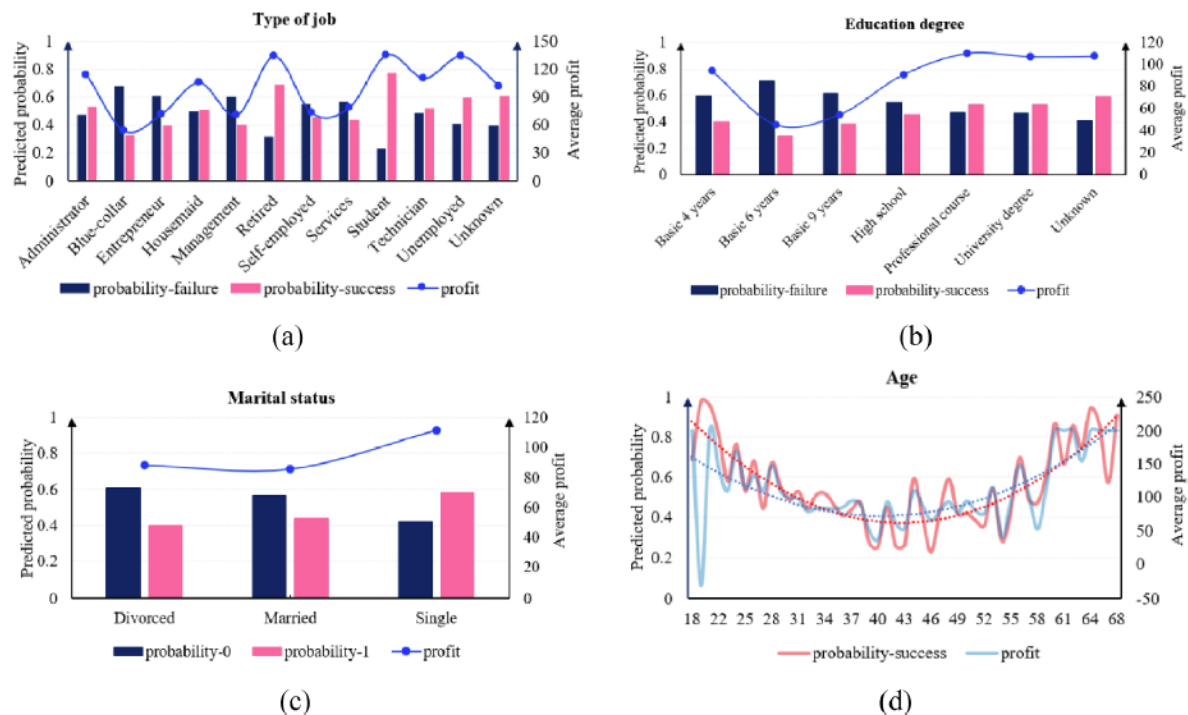


Figure 6: Relationship between bank customer information factors, and TSP and AP

Note. From “A dynamic ensemble selection method for bank telemarketing sales prediction”, Feng, Y., Yin, Y., Wang, D., & Dhamotharan, L. (2022) *Journal of Business Research*, 139, 368–382 (pp. 377) <https://doi.org/10.1016/j.jbusres.2021.09.067>

The experimental results on bank telemarketing data show that META-DES-AAP achieves the best accuracy and the largest average profit when compared across several state-of-the-art machine learning methods.

Xie et al. (2023) use three machine learning (ML) methods to find the best-performing model for predicting the success of bank telemarketing campaigns to capture consumers to subscribe to bank deposits. Telemarketing is a technique part of direct marketing that consists of contacting potential customers over the telephone. Furthermore, to understand how the chosen independent variables (IV(s)) affect the effectiveness of bank telemarketing efforts, they use partial dependence plots (PDP) to depict the marginal effects of the selected IV(s) on the customers' subscription of deposits. Further information about this technique might be found in chapters 3.3.1.3 and 3.3.2.1.

The Random Subspace-Multi Boosting algorithm is the one that achieves the best performance. By analysing the results, they have concluded that banks should give greater importance to the type of job the customer does and the month the customer was contacted.

The three previous articles and other studies, ranging from 2014 until now, have tried different algorithms on the same dataset from a retail bank in Portugal, trying to improve the prediction performance. A summary of the articles, algorithms used, and prediction results is shown in Figure B-2 in Annex B. This is an indication that machine learning practitioners have the main goal of improving prediction performance metrics and not as much to explain about the underlying data.

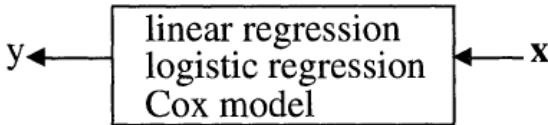
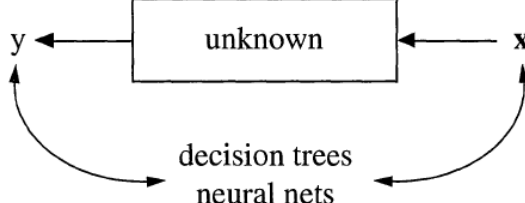
The literature review presented till now shows a separation between statistical models and machine learning ones. The reasons behind it have led to the next three articles.

Breiman, L. (2001) starts by introducing data, by definition, as being produced by a black box in which a vector of input variables (x) goes in one side and on the other side emerges the response variable y . He also distinguished the two goals in analysing data:

- Prediction. To be able to forecast the responses to forthcoming input variables.
- Information. To get some information regarding the relationship between the input and response variables.

The author pinpoints two different approaches to achieving these goals. One is called the data modelling culture, whereby one assumes that the data is generated by a given stochastic model. The other uses algorithmic models and treats the data mechanism as unknown, whose insides are complex and mysterious. A summary is presented in the table below.

Table 1: The two cultures to analyse data as defined by Breiman, L. (2001)

Cultures	Models	Model Evaluation
The data modelling culture		Goodness-of-fit tests and residual examination
The algorithm modelling culture		Measured by predictive accuracy

At that time, most of the statistics relied almost solely on the models from the data modelling culture. The author has encouraged the statisticians to start using more models from the algorithmic culture. The author has stated, “The best available solution to a data problem might be a data model; then again it might be an algorithmic model. The data and the problem guide the solution. To solve a wider range of data problems, a larger set of tools is needed.”

For instance, the study shows an example with a microarray lymphoma data set with three classes, a sample size of 81 and 4682 variables (genes). Considering the size of the dataset, an algorithm such as random forests is more suitable. From a scientific perspective, the evaluation of the significance of each of the genes was quite interesting, as displayed in the figure below.

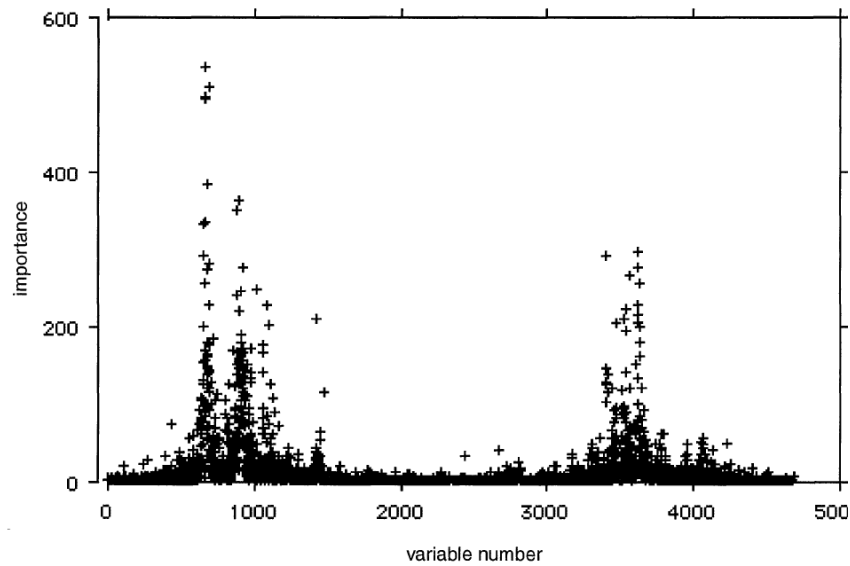


Figure 7: Microarray variable importance

Note. From "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author).", Breiman, L. (2001). *Statistical Science*, 16(3). (pp. 213) <https://doi.org/10.1214/ss/1009213726>

The author also presents that within the data modelling field, adding to the current practice of checking the data model fit using goodness-of-fit tests and residual analysis, checking the prediction accuracy on a test set is valuable to understand how good the model is and compare different models.

The article by Shmueli and Koppius (2011) starts by pointing out the differences between explanatory and predictive tasks, which are summarized in the table below.

Table 2: Differences Between Explanatory Statistical Modelling and Predictive Analytics

Step	Explanatory	Predictive
Analysis Goal	Explanatory statistical models are used for testing causal hypotheses.	Predictive models are used for predicting new observations and assessing predictability levels.
Variables of Interest	Operationalized variables are used only as instruments to study the underlying conceptual constructs and the relationships between them.	The observed, measurable variables are the focus.
Model Building Optimized Function	In explanatory modeling the focus is on minimizing model bias. Main risks are type I and II errors.	In predictive modeling the focus is on minimizing the combined bias and variance. The main risk is over-fitting.
Model Building Constraints	Empirical model must be interpretable, must support statistical testing of the hypotheses of interest, must adhere to theoretical model (e.g., in terms of form, variables, specification).	Must use variables that are available at time of model deployment.
Model Evaluation	Explanatory power is measured by strength-of-fit measures and tests (e.g., R^2 and statistical significance of coefficients).	Predictive power is measured by accuracy of out-of-sample predictions.

Note. From "Predictive Analytics in Information Systems research.", by Shmueli, M. D., & Koppius. (2011). In *Management Information Systems Quarterly*, 35(3), (pp. 557). <https://doi.org/10.2307/23042796>

The goal of finding a predictively accurate model differs from the goal of finding the true model. An optimal model for prediction purposes may be different from one obtained by estimating the "true model". Considering a linear regression model, the article states, "...although it can be used for building

an explanatory statistical model as well as a predictive model, the two resulting models will differ in many ways. The differences are not only in the statistical criteria used to assess the model, but are prevalent throughout the process of modelling: from the data used to estimate the model (e.g., variables included and excluded, form of the variables, treatment of missing data), to how performance is assessed (model validation and evaluation)”.

In addition, classical statistical education focuses on explanatory statistical modelling and statistical inference and rarely discusses prediction. On the other hand, predictive analytics are usually taught in machine learning and data mining fields.

The two previous paragraphs are insightful. The latter justifies the historical origin difference between statistical models and machine learning ones and the main purpose they have been used for. The former shows the difference relies on the task underhand, explanatory or prediction, and not on the model itself. The linear regression model, which historically falls under the realm of statistical models, can be used with a prediction intent, and hence fit under the machine learning domain.

The article acknowledges that the information systems (IS) research field, as well as other disciplines such as economics and finance, rely almost solely on explanatory statistical models, while also defending the need to integrate predictive analytics with it. Predictive analytics refers to methods that generate data predictions and assess their results. Explanatory statistical modelling relies on statistical inference techniques used to test and evaluate the explanatory power of underlying causal models.

The article points out that, even though explanation and prediction are best thought of as two separate modelling goals, they are not mutually exclusive. If the primary goal is a causal explanation, one can construct an explanatory statistical model and then, in a subsequent stage, evaluate its predictive power using predictive analytics, which can add substantial insight by:

- Comparing different explanatory statistical models.
- A good predictive power is a good reason for accepting the explanation.

In addition, predictive analytics can capture complex underlying patterns and relationships and therefore enhance current explanatory statistical models.

Kleinberg et al. (2015) also give an interesting insight into the differences between explanatory and predictive tasks.

They say that standard empirical techniques are not optimized for prediction problems because they focus on the concept of unbiasedness. For instance, ordinary least squares (OLS) is the best unbiased linear estimator. Unbiasedness implies that the mean values of the OLS estimated regression coefficients conform to the unknown population regression coefficients.

OLS reduces in-sample error. However, the goal in prediction is to perform well out of the sample, meaning a test dataset. Making sure there is no bias in the sample leads to issues in the out of sample. To see this, the mean squared error formula at a new point x from the test dataset can be decomposed in the following figure.

$$\underbrace{E_D[(\hat{f}(x) - E_D[\hat{y}_0])]^2}_{\text{Variance}} + \underbrace{(E_D[\hat{y}_0] - y)^2}_{\text{Bias}^2}.$$

Figure 8: Bias-variance decomposition of mean squared error

Note. From "Prediction policy problems.", by Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). *The American Economic Review*, 105(5), 491–495 <https://doi.org/10.1257/aer.p20151023>

Because the f varies from sample to sample, it produces variance. This must be traded off against bias. OLS assures zero bias, hence there is no bias-variance trade-off. By offering an empirical method for balancing this bias-variance trade-off, machine learning approaches were created expressly to maximise prediction performance, naturally arising from non-parametric statistics. The trade-off arises because decreasing bias often leads to an increase in variance and vice versa. Finding the right balance is crucial for building a model that generalizes well to new, unseen data.

In essence, prediction problems only require low error in \hat{Y} and do not require the coefficients to be unbiased. Machine learning techniques offer a methodological approach to accomplishing it by using the data itself to decide how to make the bias-variance trade-off and by allowing a search across a rich range of variables and functional forms. However, one must always keep in mind that because they are tuned for \hat{Y} , they do not (without many other assumptions) give very useful guarantees for the regression coefficients $\hat{\beta}$.

The following four articles present which algorithms fit better under the explanatory task paradigm and the challenges to getting reliable information from the models, which will ultimately impact business decisions.

Hoepner et al. (2020) state that while machine learning is commonly linked to neural networks, it also includes statistical techniques such as econometric regression and others.

Proponents of explainable artificial intelligence (xai) use traceable "white box" techniques, like regressions, to improve explainability to human decision makers, while proponents of deep learning, or neural networks, prefer computational efficiency over human interpretability and accept the "black box" appeal of their algorithms.

They highlight explainability as the main issue in traditional machine learning techniques like neural networks, which are not fully replicable, lack transparency and traceability, and hence make it difficult to draw explainable conclusions.

They advocate the use of "white box" algorithms, whose results are more explainable and transparent, particularly for critical industries such as healthcare and finance, where human users need to trace individual aspects of the decision process.

Hansen (2020) discusses the use of machine learning models in finance and how quants manage the complexity of these models. They use Ockham's razor as a heuristic tool, which states the simplest explanation is preferable to one that is more complex, to prevent overfitting and maintain a certain level of comprehensibility and control in the modelling process.

The study argues that explainability and comprehensibility play an important role in algorithmic trading and quantitative investment because the most complex models in finance, like multi-layer artificial neural networks, contain components that are difficult to fully understand. Comprehension involves grasping the logic and being capable of interpreting the model output. In addition to ensuring explainability, the preference for simplicity serves as a protective measure against the potential for models to learn from data noise rather than meaningful information. Less complex models with fewer parameters have a lower chance of overfitting and making inaccurate predictions.

According to one quant, he states, “I always start with the simplest model...We will for example, take a logistic regression model and use it as our benchmark model, and then do experiments based on [the following questions]: Does adding this factor help? Does using a more sophisticated model help? The more sophistication I add to the model, the more model risk I add to the strategy. [. . .] We always try to start with the simplest model we think will work and move very, very carefully to more complex models.”

Overall, the study highlights the importance of human judgement in machine-driven finance and contributes to academic discussions of the human role in the increasingly automated world of finance.

The two previous articles, Hansen (2020) and Hoepner et al. (2020) advocate the use of simpler models in the case of explanatory tasks. Aligned with it and linking to the marketing industry context, De Bruyn et al. (2020) provide an overview of the current state of artificial intelligence (AI) in marketing and its prospects.

They discuss that marketing managers must be mindful of various technological pitfalls and risks when integrating AI into their organizations. Deep-learning neural networks are AI algorithms that do more than just adjust model parameters according to human specifications. As the input data travels through the various layers of a deep neural network, the algorithm will independently reassemble the data into higher-level constructs, effectively identifying and compiling its list of predictive variables. This can lead to a few unanticipated endogeneity problems that are outside of people's control. Compared to traditional regression studies, where the analyst must identify the independent variables, this is a significant departure. These AI models, in a way, build themselves on their own. There are particular obstacles that come with this indisputable strength.

They state, “Knowing the superior ability of AI systems to leverage endogenous relationships in the data, exploit spurious correlations, replicate human biases, and make theory-free predictions, we argue that management should closely scrutinize AI-based marketing models, and that their transparency, explainability, and interpretability should be a constant and pressing concern”.

Based on that, they foresee that AI will significantly impact predicted tasks that are automatable and require small explanations. However, they consider that the issues surrounding the transfer of tacit knowledge between AI models and marketing organisations might lead to AI not living up to expectations in many marketing sectors. In the context of marketing, tacit knowledge includes skills, intuition, and creativity that are critical for decision-making but are difficult to codify and transfer to AI models. Additionally, AI models may not be able to explain their decision-making processes, which can limit their usefulness in domains where explainability is important. Therefore, marketing managers need to ensure that their AI systems are designed to transfer tacit knowledge effectively between AI

models and marketing organizations and provide explanations for their decision-making processes to maximize the benefits of AI in marketing.

Following the same topic, Rudin (2019) discusses the difficulty of applying machine learning models to applications with high stakes, like credit scoring, recidivism prediction, and medical diagnosis. Many of these models are black boxes that do not explain their predictions in a way that humans can understand. This lack of transparency and accountability can have severe consequences, as there have been cases of people incorrectly denied liberation, poor bail decisions leading to the release of dangerous criminals, etc.

The field of "explainable ML," in which a second (post hoc) model is developed to explain the first black box model, has seen a recent surge in activity. This presents an issue since explanations are frequently unreliable and may even be deceptive. Alternatively, interpretable models offer their justifications that match the model's real computations. It is important to note that the term 'explanation' here refers to an understanding of how a model works rather than an explanation of how the world works.

The author points out several key issues with explainable ML:

- Explainable ML methods can provide explanations that are not faithful to what the original models compute. Even an explanation model that makes predictions that are nearly exact matches to a black box model may employ entirely distinct features, which means it is not faithful to the black box's computation.
- Frequently, explanations do not make sense or provide insufficient information to make sense of what the black box is doing.
- Explanatory black-box models can result in an unduly complex decision-making process that is prone to human error.

Based on the above, she advocates that policymakers should reject black box models in favour of interpretable (as opposed to explainable) models, especially for high-stakes prediction applications that deeply impact human lives.

In the next article, the authors present the current state and main challenges on which machine learning and statistics fields are focused now. Hooker and Mentch (2021) state, "A newfound interest in interpretability (e.g. "explainable AI"), causality, and fairness is commonly on display at each of the most notable machine learning conferences, each of which hearkens back to fundamental ideas and philosophical considerations in traditional statistical modeling."

Considering the articles presented, the main takeaways from the literature review are:

- Explanation and prediction tasks lead to some differences in the model building process, which translate into different final models.
- Traditional statistical models, like regression ones, focus more on explanatory tasks. But they can be used for predictive tasks too.
- The predictive analytics framework, namely how the model is evaluated, can be useful in explanation tasks to measure how good the model is and compare different models.
- Machine learning algorithms, like neural networks, are more suitable for predictive tasks.

3. METHODOLOGY

3.1. DATA ASSEMBLY

Considering the business model in Figure 2 and the research problem, data from three different sources had to be linked together. These are:

- **Trades** – this data corresponds to all RFQs. It signals the interest of the client in a given product.
- **Tradeideas** - this dataset contains the BNP research team’s advice to buy or sell specific bond securities, with the advice starting at a specific time and valid for a certain period.
- **Telemetry** – this dataset shows if the clients have opened the emails and downloaded the reports attached.

The figure below shows the data source fields used to establish the necessary data linkage. Note the different word styles that identify the fields that had to match each other from the different data sources. The relationship between the data sources is many-to-many. The field’s meaning is the following:

- **MarketingClientCode** – unique client identifier
- **ISIN** – unique bond security identifier
- **ReportID** – unique report identifier
- **Direction** - identification of a trade suggestion (buy or sell)
- **BuySell** – identification of trade direction (buy or sell)

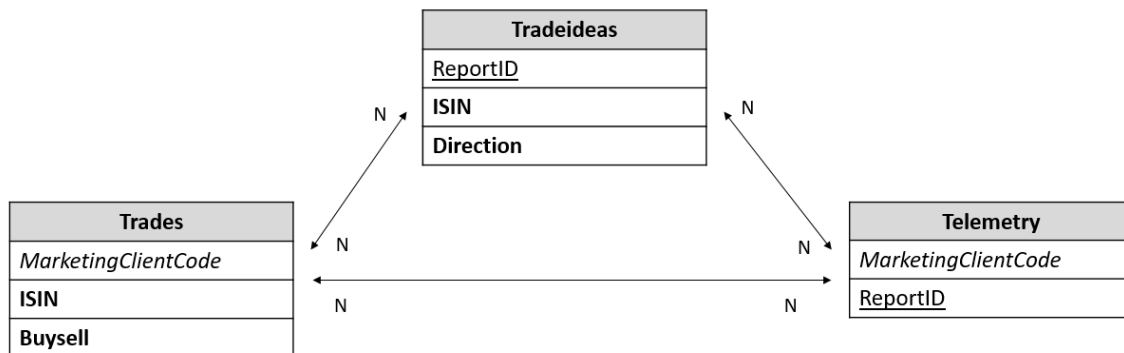


Figure 9: Data sources and fields identification to establish the data linkage

The linkage has been done with telemetry data from the three previous months against a one-month trading period. The following figure shows an example.

Table 3: Time period reference example to perform the data source linkage

Data Source	Start Period	End Period
Trades	01/01/2022	31/01/2022
Telemetry	01/10/2021	31/01/2022

The three-month period considers the nature of the bonds' security-type product. The trade ideas are opened and kept valid for extended periods, and bonds are a fixed-income product that has, in its nature, little fluctuation.

The trading time considered for this study has been the full years of 2021 and 2022.

The data output, after the linkage presented in Figure 9, can be split into three subsets:

- **Only Trades** – these are the RFQs that have not matched with telemetry events. It means clients who have not received an email or who have made an RFQ on a bond security that has not been suggested in the reports.
- **Trades-Telemetry** – trades and telemetry events have matched. It means the client has requested an RFQ on a bond security, after having received an email with a report suggestion to trade the same bond.
- **Only Telemetry** – these are telemetry events that have not matched with the trades. Those are the cases where the client has received an email, but an RFQ has not been followed.

Knowing this study aims to understand the impact of the reports in requesting or not an RFQ, the subset *only trades* dataset is discarded since there are no telemetry measurements.

Another important data pre-processing point has been data normalization. The subset *trades-telemetry* and *only telemetry* must have the same *MarketingClientCode* and *ReportID* fields. For instance, after data linkage, the aforementioned subsets had, respectively, 866 and 2161 clients, and in common 866 clients. It means that 1295 clients have received reports, but no RFQ has been followed. These clients are meaningless and constitute noise for the study since they do not allow to understand the impact of the reports in either requesting or not an RFQ.

In addition, the same trade idea can be included in different reports. That said, the subset *trades-telemetry* had for the same trade a linkage to different reports.

Table A-1 in Appendix A displays an example of the data linkage output with the subsets *only trades* and *trades-telemetry*, and with only some fields shown.

At this point, it is important to distinguish between *observational studies* and *experiments* since these concepts will be used throughout the study.

Observational studies draw inferences from a sample of the population, observing the DV and IVs without any control from the researcher. In contrast, *experiments*, such as randomized controlled trials (RCTs), where each subject is randomly assigned to a treatment or a control group. The distinction is presented in Figure 10.

This research project falls under the observational study category since past data has been gathered without any intervention.

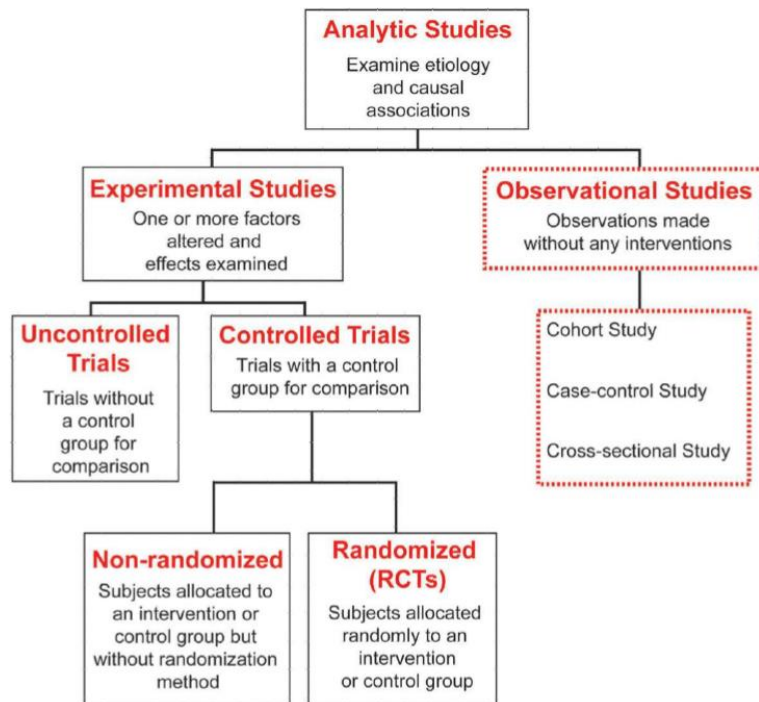


Figure 10: Observational and experimental studies

Note. From “Observational studies: Cohort and Case-Control studies”, Song, J. W., & Chung, K. C. (2010). In *Plastic and Reconstructive Surgery*, 126(6), 2234–2242 (pp. 2). <https://doi.org/10.1097/prs.0b013e3181f44abc>

3.2. FEATURE SELECTION

The variables’ selection for the model has been done as described in the book (Hosmer et al., 2013).

The model building aims to find the most parsimonious model, meaning with as few IV(s) as possible, but still explaining well the data. The analysis starts with the identification of one variable at a time, that is at least moderately associated with the response variable. For doing so and in the case of categorical variables, the *chi-squared statistic test* of independence is used, as per equation (3.1). It evaluates the independence of two variables, which are expressed in a contingency table. The latter is a kind of matrix-format table used in statistics that shows the variables’ multivariate frequency distribution.

$$H_0: \text{the variables are independent} \quad \text{vs} \quad H_1: \text{the variables are dependent} \quad (3.1)$$

The analysis has shown high *chi-squared statistic test* values, with very small *p-values*, leading invariably to the rejection of the null hypothesis (H_0). In other words, it means the IV(s) analysed were always statistically significant dependent against the DV, regardless of the threshold significance level used (usually 0.01 or 0.05 level values are considered). The *chi-squared statistic test* results are in accordance with Sullivan and Feinn (2012), which present that with a sufficiently large sample, a statistical test will almost always show a significant difference. Therefore, the *chi-squared statistic test* has demonstrated that it is not a suitable technique to measure the association of the IV(s) with the DV.

The next step was to build stacked bar charts with the DV proportions against each different IV using the contingency table frequencies. The plot’s visualization facilitates understanding the impact or

association between the IV(s) and DV variables. On top of that, a way to quantify the association was investigated. Sullivan and Feinn (2012) state that *p-values* inform if an association between variables exists, but do not quantify the effect of it. To accomplish this, *Cramer's V* association test can be used with categorical variables. The values' interpretation is based on Cohen (2013), considering two degrees of freedom, and can be checked in the table below.

Table 4: Cramer's V Interpretation

Cramer's V	Correlation Effect Interpretation
> 0.35	Large effect
0.21 < 0.35	Medium effect
<0.21	Small effect

3.3. MODELS

Regression models, such as linear and logistic regression, are one of the most important statistical tools in data analysis. They produce results that are easier to interpret and, therefore, can be categorized as interpretable models (Molnar, 2020). 'White-box' models are another name for these models found in the literature. These models have their origins in traditional statistics.

On the other hand, 'black-box' models such as neural networks, gradient boosting models, and others often provide better accuracy, but are harder to explain and interpret. These models fall into the realm of machine learning.

Table 5: Comparison of machine learning vs. a standard statistical approach

Measurements	Machine Learning	Standard Statistics (linear/logistic regressions)
Type of Data?	Multi-dimensional data that can be non-linear	Linear Data
Fit?	Best fit to learning models (generalization), that try to extract rules from data	Fit to the distribution, meaning, calculate the model parameters that could have led to the observed data distribution
Training-Test Datasets?	Yes	No
Goal?	Generally better for predictions	Generally better for inferences about the relationship between variables and their significance
Scientific question?	What will happen?	How/why it happens?

Note. Adapted from "Machine learning applications in the neuro ICU: a solution to big data mayhem?", Chaudhry, F., Hunt, R. J., Hariharan, P., Anand, S. K., Sanjay, S., Kjoller, E. E., Bartlett, C. M., Johnson, K. W., Levy, P. D., Noushmehr, H., & Lee, I. Y. (2020). In *Frontiers in Neurology*, 11 (pp. 3). <https://doi.org/10.3389/fneur.2020.554633>

Please note that both approaches overlap and are complementary (Chaudhry et al., 2020). In addition, linear and logistic regression models can also be used within the machine-learning framework.

The main dissertation's goal is to understand the reports' impact on clients' purchase behaviour. Therefore, the primary interest is not in creating a predictive model to forecast future client behaviour, but rather in understanding how one explanatory variable (reports) influences the response variable (trade behaviour). The main task is explanatory and not a predictive one.

Based on the previous paragraphs and the literature review presented in Chapter 2, the study has started with the application of the logistic regression model with a statistical approach, as presented in Table 2 and Table 5.

Afterwards, logistic regression and random forest models have been used with the predictive analytics framework, which is part of the machine learning realm. It means both models' predictive performance has been assessed on a test dataset.

The figure below depicts the general idea that more complex and accurate models tend to be less interpretable (Gulum et al., 2021). These comparisons are simplified rules and therefore may not be true for every dataset, as the size and quality of the dataset can change its performance.

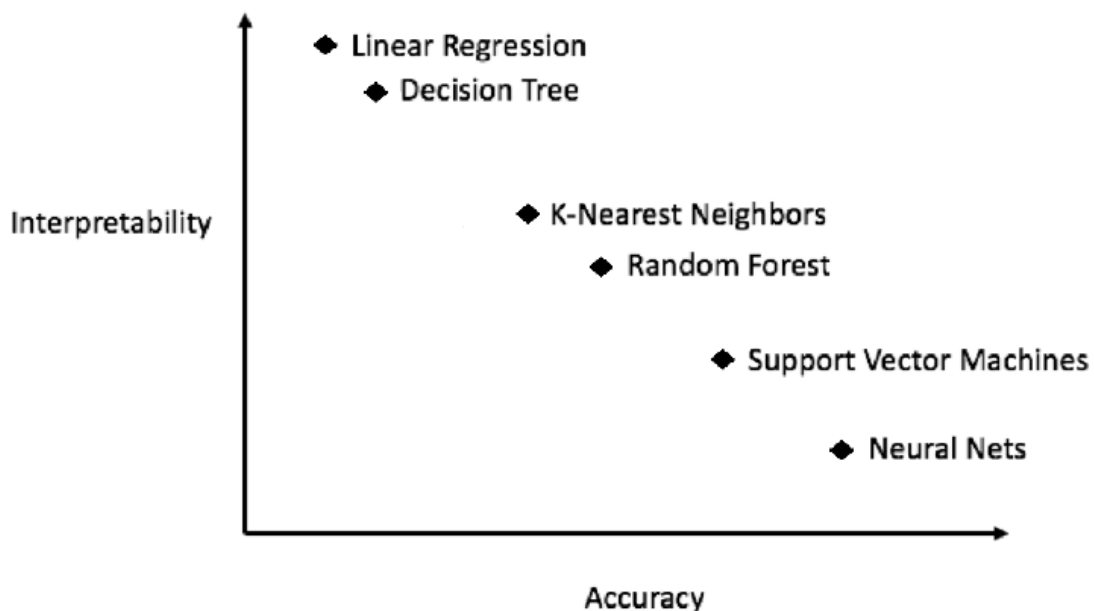


Figure 11: In general, models that are more accurate are less explainable.

Note. From "A review of Explainable deep learning cancer detection models in medical imaging", by Gulum, M. A., Trombley, C. M., & Kantardzic, M. (2021). In *Applied Sciences*, 11(10), 4573. (pp. 3). <https://doi.org/10.3390/app11104573>

However, in the last few years, a set of agnostic models has been developed to increase the interpretability of complex ML models without being tied to any ML model. These tools aim to provide insights into how models make predictions and which features are most influential. Therefore, the trade-off between interpretability and prediction, as shown in Figure 11, is reduced. That is the justification for using a more complex model, such as a random forest algorithm, knowing the dissertation's goal is to explain and not to predict. Further information about these agnostic explainable models is presented in Chapter 3.3.2.1.

3.3.1. Standard Statistics Logistic Regression

Logistic regression aims to model the relationship between one or more independent (or explanatory) variables and a discrete dependent (or response) variable. When the dependent variable can only take two values, usually zero and one, “binary logistic regression” is the model reference name.

In statistics, the logistic model (also known as the logit model) models the probability that an event will occur by making the event's log-odds a linear combination of one or more IV(s).

Formula – Logistic Regression Logit Form	Notation
$\text{logit}(\pi(X)) = \ln \left[\frac{\pi(X)}{1 - \pi(X)} \right] = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad (3.1)$	<ul style="list-style-type: none"> ▪ $\pi(X) = P(Y = 1 X) \in \{0,1\}$, meaning, it outputs a probability between 0 and 1 ▪ $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ are unknown parameters ▪ $X = (X_1, \dots, X_p)$ are the independent features ▪ $\frac{\pi(X)}{1 - \pi(X)} = Odds$

Using algebraic manipulation, the model's results can be converted from log-odds, equation (3.1), to probabilities, equation (3.2), and vice versa.

Formula – Logistic Regression Probability Form

$$\pi(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}} \quad (3.2)$$

The goal now is to estimate the β parameters. Different methods are available. The most common for LR is the maximum likelihood.

Formula – Maximum likelihood function	Notation
$l(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (3.3)$	<ul style="list-style-type: none"> ▪ $i = (1, \dots, n)$ refers to the observations' number. ▪ y_i is the dependent variable's value for observation i. ▪ x_i is a vector with the independent variables' values for observation i

To simplify calculations, the natural logarithm is applied to the likelihood function, which turns products into sums.

Formula – Log-likelihood function

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^n \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \quad (3.4)$$

To find the values $\beta_0, \beta_1, \dots, \beta_p$ that maximise the log-likelihood function, the partial derivatives with respect to the $(p+1)$ coefficients are calculated and matched to zero.

Formula – Partial Derivatives

Notation

$$\sum_{i=1}^n [y_i - \pi_i] = 0 \quad (3.5)$$

$$j = (1, 2, \dots, p)$$

$$\sum_{i=1}^n x_{ij} [y_i - \pi_i] = 0 \quad (3.6)$$

The above equations are nonlinear. They calculate the parameter β_0 and the remaining β coefficients, respectively, with numerical iterative methods.

The reader might check for further information about the LR theoretical background (Hosmer et al., 2013).

The same book also explains how the variables should be modelled to be included in the model. In the presence of polychotomous categorical variables, meaning, a categorical variable with more than two values, they shall be encoded using a technique called *one-hot encoding*. For each category within the polychotomous variable, a binary (0 or 1) dummy variable is created, except for the reference class. The reference class is the category left out to avoid multicollinearity among the dummy variables. See the example in the table below, where the reference class is the white race.

Table 6: Specification of the dummy variables and the reference class

Race (Code)	Dummy Variables		
	RACE_2	RACE_3	RACE_4
White (1)	0	0	0
Black (2)	1	0	0
Hispanic (3)	0	1	0
Other (4)	0	0	1

The continuous IV(s) modelling process is not described since it has not been used in this study.

3.3.1.1. Selection Model Criteria

After defining the relevant independent variables (IV) following Chapter 3.2, different LR models with different subsets of IV(s) can be studied. The *Akaike information criteria* (AIC) is a statistical measure commonly used to compare models with different number of parameters (Hosmer et al., 2013).

AIC handles the trade-off between the goodness of fit of the model and the simplicity of the model. The goodness of fit of a statistical model describes how well it fits a set of observations.

Formula – AIC	Notation
$AIC = -2\ln(L_m) + 2q \quad (3.7)$	<ul style="list-style-type: none">▪ L_m is the value of the likelihood model fitted.▪ q is the number of independent variables.

The best model is the one with the smallest AIC value.

To check the adequacy of the model in terms of goodness of fit, meaning global fit, the metric Pseudo-R-squared is used. The standard index used by the Python library *Statsmodels* is the McFadden's one. The formula calculation is presented in equation (3.8).

Formula – McFadden's pseudo-R-squared	Notation
$Pseudo R - squ. = 1 - \frac{\ln(L_M)}{\ln(L_0)} \quad (3.8)$	<ul style="list-style-type: none">▪ L_M is the value of the likelihood model fitted▪ L_0 is the value of the likelihood null model

In the book from the authors Hensher and Stopher (1979), McFadden states a pseudo-R-squared value between 0.20 - 0.40 represents an excellent fit.

3.3.1.2. Model Diagnostics

Once the model that best fits the data has been built, it is necessary to analyse the quality of that model. Therefore, in this section, some measurements that summarize the model's quality adjustment are presented.

Residuals

The logistic regression model's residuals must be altered to be usable, while in linear regression, the residuals are evaluated exactly as they are. The dependent variable's binary nature (0 or 1) explains this. The residuals will not be normally distributed, and their distribution is unknown because the outcome is binary (Kutner et al., 2004). The residuals that are evaluated at that point can be classified as studentized Pearson residuals, or Pearson residuals.

Plotting the studentized Pearson residuals against the predicted probability is a useful plotting technique for diagnosing logistic regression models. (Kutner et al., 2004) show that under the

assumption that the logistic regression model is correct, the expected error (difference) between the observed value and predicted value is approximately zero.

Formula	Notation
$E[y_i - \pi_i] = E[e_i] \cong 0 \quad (3.9)$	<ul style="list-style-type: none"> ▪ y_i is the dependent variable's value for observation i ▪ π_i is the calculated outcome probability for observation i ▪ e_i is the error term

They conclude that a lowess smooth line in one of the plots mentioned above would be approximately a horizontal line with zero intercept.

Multicollinearity

Multicollinearity corresponds to a situation where the data contain highly correlated independent variables. This is a problem because it lowers the predicted coefficients' precision, which lowers the logistic regression model's statistical power.

The variance inflation factor (VIF) is a good diagnostic technique that measures the degree of multicollinearity in a set of independent variables. Mathematically, it is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. A VIF value that exceeds 5 or 10 indicates a problematic amount of multicollinearity (Kutner et al., 2004).

Outliers

The impact of the outliers on the model's parameter estimates has not been assessed since the large volume of data smooths out the influence of possible extreme values.

3.3.1.3. Marginal Effects

Regression coefficients, along with standard errors and other goodness-of-fit and summary statistics, are the usual regression model outputs given by the statistical software.

The LR interpretability results, in part, from the β coefficients assessment:

- A negative or positive coefficient sign means, respectively, a decrease or increase in the response variable.
- The magnitude of each coefficient indicates the contribution to the response variable. Variables with larger coefficients are more influential on the response variable.

Also, the overall model's goodness of fit and the magnitude of change in the coefficients of the explanatory variables, by adding or removing other explanatory variables allow to understand the possible collinearity between variables and the impact on the overall model.

However, looking at equation (3.1), the LR coefficients β s are in the log-odds scale, which are not directly interpretable, apart from the sign. For instance, the coefficient β_1 is the effect of the feature X_1 on the log-odds of the outcome, not on the probability, which is often what we are looking for.

In the probability scale, equation (3.2), the response results are non-linear and conditional on the independent features. This can make the interpretation of each independent feature on the probability of interest difficult. Please see the figure below (MacKenzie et al., 2018, p. 95).

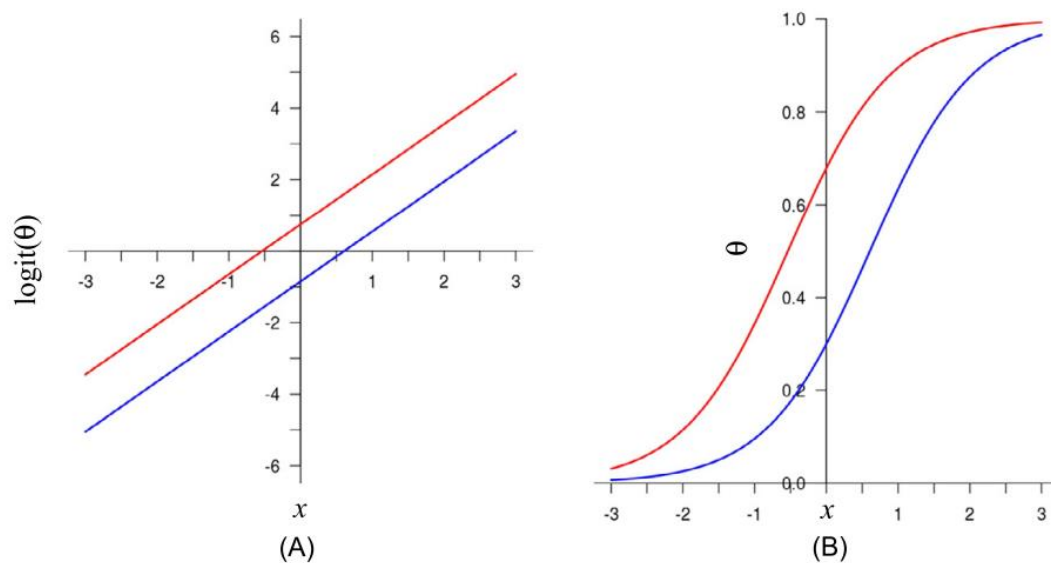


Figure 12: Illustration of the logit link function's nonlinear behaviour. Linear relationships that are parallel on the logit-scale (A) are not parallel and have different slopes for a given x value on the probability scale (B).

Note. From "Fundamental principals of statistical inference", by MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., & Hines, J. E. (2018). In Elsevier eBooks (pp. 95). <https://doi.org/10.1016/b978-0-12-407197-1.00004-1>

To interpret the model on a probability scale, we can resort to the marginal effects.

Marginal effects refer to the rate at which the outcome of the response variable changes with respect to one explanatory variable while holding all other explanatory variables constant.

In statistical models, marginal effects require the application of partial derivatives. A marginal effect is, in essence, the slope of a multi-dimensional surface with respect to one dimension of that surface. With dummy variables, the calculation is easier, requiring only the change of the explanatory variable in the study from zero to one.

The marginal effect is different for each observation, meaning a decision must be made regarding how to combine all the observations' effects to display a single marginal effect. The most common metric used is the *average marginal effect*, which calculates the marginal effect for each observation separately and then takes the mean of the marginal effects.

In practical terms, and considering the explanatory variables under study are binary, the model calculates the outcome probability with one IV's values equal to zero and one, and its mean difference is what is called the "Average Marginal Effect" for the IV under study.

The reader might check for further information about the marginal effects on (Leeper, 2013).

3.3.2. Machine Learning

In this sub-chapter, the logistic regression model has been applied under the machine learning framework. So, the question that arises is, “What is the difference between the methodology presented in sub-chapter 3.3.1?”. These are highlighted in Table 5, but a more descriptive explanation is presented:

- With regards to model selection, the problem of choosing one model from among a set of candidates is that standard statistics (SS) uses a probabilistic statistical measure like the AIC presented in Chapter 3.3.1.1, while ML uses the model that performs the best in a test dataset using the holdout method or K-fold cross-validation.
- SS uses the whole dataset to fit the model. ML fits the model on a training dataset and tests it on a test dataset.
- The SS takes advantage of the *Statsmodels* Python library, which is primarily designed for statistical analysis and hypothesis testing. ML uses the *scikit-learn* library, which is focused on machine learning.

With regard to the first two points above, Shmueli and Koppius (2011) state, “While explanatory power is evaluated using in-sample strength-of-fit measures, predictive power is evaluated using out-of-sample prediction accuracy measures”.

Based on the last statement, it is important to define some concepts. The strength-of-fit measures are, for instance, the ones presented in Chapter 3.3.1.1. The out-of-sample concept refers to data the model has not seen during training and is used to evaluate the model’s performance. In this study, the method used to evaluate the classifier was the *holdout* method. This method splits the dataset into two sets, called the training set and the test set, with the latter corresponding to the out-of-sample concept presented. Both datasets keep the same class distribution as the entire dataset, shown in the table below, using the stratification technique.

Table 7: Dataset class distribution

Dependent Variable	% Distribution
RFQ=0	92,05%
RFQ=1	7,95%

The holdout method has been used rather than the k-fold cross-validation due to its simplicity and the large dataset available.

To measure the classifier’s performance, Ferri et al. (2009) split the classification prediction evaluation metrics into three groups:

- *Threshold metrics* (e.g., *accuracy* and *F-measure*) - These measures are used when we want a model to minimise the number of errors.
- *Ranking metrics* (e.g., receiver operating characteristic curve (ROC curve) analysis and AUC) - Rank metrics are more concerned with evaluating classifiers based on how effective they are at separating classes.

- *Probability metrics* (e.g., *Brier Score* and *LogLoss*) – these metrics measure the deviation from the true probability. They are especially useful when we want an assessment of the reliability of the classifiers, not only measuring when they fail but also whether they have selected the wrong class with a high or low probability. They quantify the uncertainty in a classifier's predictions.

There are problems where the objective is to predict the correct class. In the context of the research problem in this study, it would mean predicting if the client would request an RFQ or not. For these kinds of problems, the *threshold metrics* are the most suitable to use. But that is not the research question this dissertation is trying to answer. The goal is to draw conclusions given the observed data. More specifically, the idea is to quantify the impact of the reports on the client trade behaviour with a probability. As such, the *probability metrics* are the ones most appropriate to evaluate and compare different models in this study.

Figure B-3, in Annex B, depicts a general guideline for choosing the performance metrics in accordance with the study aim.

As mentioned, *probability metrics* are used to assess the reliability and calibration of a classifier. Calibration is the concordance of predicted probabilities with the occurrence of observed cases. Kühn and Johnson (2013) state, “...we desire that the estimated class probabilities are reflective of the true underlying probability of the sample. That is, the predicted class probability (or probability-like value) needs to be well-calibrated. To be well-calibrated, the probabilities must effectively reflect the true likelihood of the event of interest.”.

In practice, imagine a dataset with 100 observations, with 80 belonging to class 1 and 20 to class 0. The model is calibrated if the mean predicted probability value for all observations belonging to class 1 is close to 0.80.

LR outputs, in general, calibrated probabilities, since they are trained using a probabilistic framework, like the maximum likelihood estimation (MLE) method, as shown in chapter 3.3.1. With that in mind, it is important to give a practical intuition of what the MLE is. Given observed data, the maximum likelihood concept relies on estimating the model parameters that maximize the likelihood of the observed data. Likelihood is the probability of getting the observed data, given a specific model. The goal is to find the model that most likely produced the observed data, which is expressed through the metric in equation 3.7.

Other ML models, such as the Random Forest (RF) algorithm, are not trained using a probabilistic framework. As such, these models' probabilities output might require calibration, which can be assessed with probability metrics, such as the *Brier Score* and *LogLoss*.

The *Brier Score* and *LogLoss* metrics formulation are as follows:

Formula – Brier Score

$$BS = \frac{1}{N} \sum_{t=1}^N (p_t - y_t)^2 \quad (3.9)$$

Notation

- p_t is the probability predicted for the observation t

Formula – LogLoss

$$\text{Log Loss} = -\frac{1}{N} \sum_{t=1}^n [y_t \log(p_t) + (1 - y_t) \log(1 - p_t)] \quad (3.10)$$

- y_t is the actual observed outcome for the observation t
- N is the total number of observations

The lower both scores are for a set of predictions, the better the predictions are calibrated.

The *ranking* metrics ROC curve and ROC AUC have been used either. The ROC curve is a graphical plot that shows the performance of a binary classifier at all different classification thresholds. The ROC AUC is the area under the ROC curve. Both the ROC curve and the AUC use probabilities for their calculations. These metrics show how effective the models are at separating classes. That is not the main goal of the research question, but still, they are good metrics to compare different models.

Since the dataset is unbalanced, as shown in Table 7, this can lead to biased models that perform poorly for the minority class. Therefore, the random undersampling technique has been applied. It consists of removing randomly selected observations from the majority class in the training dataset. The goal is for the model to learn about the minority class better, potentially improving its overall performance.

The Random Forest (RF) model was the next one assessed, aiming to improve the classification metrics. It is an ensemble learning technique that can be used in classification tasks. It works by building many decision trees during the training phase. For each observation, the class that most of the trees choose is the RF's output final classification.

It is a model that offers a good balance between interpretability and accuracy, as shown previously in Figure 11. The interpretation comes from the feature importance scores, an underlying feature of the RF, which allow for understanding the features most impactful on the model. RF captures more complex relationships in the data compared to LR since they are non-linear and linear in nature, respectively. The last statement presents one of the advantages of usual machine learning models compared to traditional statistical ones.

In Chapter 3.3.1, the LR probability output calculation has been explained. The RF calculates as follows:

1. Voting by trees: each decision tree in the RF independently classifies the observation as either class 0 or 1.
2. Counting votes and probability calculation: the RF calculates the fraction of trees that voted for class 1 out of the total number of trees. This fraction is the estimated probability that the observation belongs to class 1. The same reasoning can be applied to class 0.

3.3.2.1. Model Explanation

There are an increasing number of model-agnostic interpretation techniques for ML models. Molnar et al. (2022) distinguish between the feature effect and feature importance methods. A feature effect shows how changes in feature values affect the direction and magnitude of the expected outcome. Feature importance techniques measure a feature's impact on the performance of the model. The popular available techniques are summarized in the figure below.

		Local	Global
Feature	Effects	ICE LIME Counterfactuals Shapley Values SHAP	PDP ALE
	Importance	ICI	PI PFI SAGE

Figure 13: Popular model-agnostic interpretation techniques, classified as local or global, and as effect or importance methods.

Note. From “General Pitfalls of Model-Agnostic interpretation Methods for Machine learning Models.”, by Molnar, C., König, G., Herbringer, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., & Bischl, B. (2022). In Lecture Notes in Computer Science (pp. 39–68). https://doi.org/10.1007/978-3-031-04083-2_4

Considering this research study, the goal is to understand the global effects of the features related to the reports prepared by BNP. Based on that, the partial dependence plot (PDP) has been the technique selected.

The PDP displays the marginal effect one or two features have on the machine-learning model’s prediction outcome (Friedman, 2001). The marginal effects have been explained in Chapter 3.3.1.3. In summary, PDP shows the average prediction of a model change as a function of one or more features.

A PDP can demonstrate if a target and a feature have a linear, monotonic, or more complex relationship.

PDPs are typically generated using the entire dataset, not just the training or testing subsets. These plots help visualize the relationship between a feature and the model’s predictions by holding all other features constant. Using the full dataset provides a more accurate representation of the feature’s influence on the model’s output.

3.3.3. Causal Inference

Several research questions, including this one, intend to give an answer related to how much a phenomenon X impacts an outcome Y. It is where the causal inference framework fits in. Causal inference is concerned with the understanding cause-and-effect relationship between variables. It goes beyond describing associations (correlations) to determine whether one variable causes a change in another. It is about understanding under which circumstances correlation (association) does imply causation. This topic will be briefly introduced to understand under which circumstances causality claims could be made and if this study meets those.

Some notations will be introduced to ease the explanation:

- Treatment is a term to denote some intervention for which the effect is desired to be known. In this study, the treatment is the download of the reports, and the effect is an RFQ request or not.
- Treatment and control groups refer to units that either receive the treatment or not, respectively.

The first instinct to measure the treatment effect is to compare the outcome between the treatment and control groups. However, there might be other factors that make the outcome different, regardless of the treatment type assigned to each unit. This concept is explained visually, with an example from the book (Facure, 2023).

The following figure shows a plot with a trend, where the bigger the business size is, the more the business sells. The different colour dots distinguish the treatment and control groups, with the white and blue dots representing businesses that cut their prices and not, respectively.

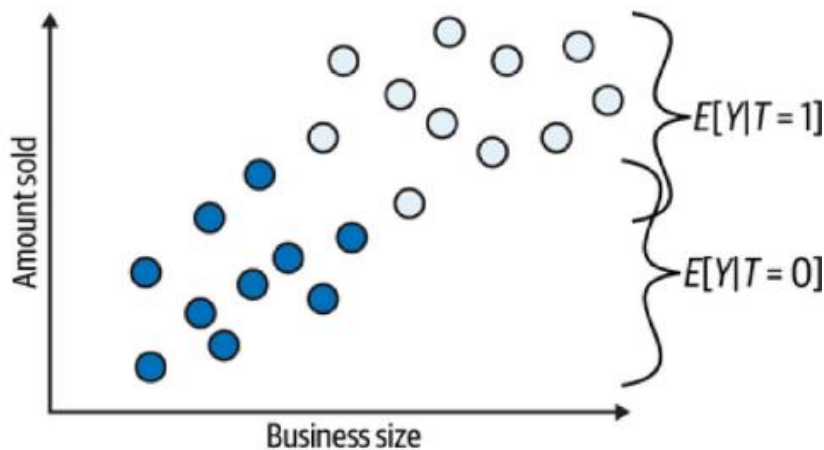


Figure 14: Example average amount sold between the treated and the control group

Note. From “Causal Inference in Python”, by Facure, M. (2023). In O'Reilly Media, Inc (pp. 38).

The difference in the amount sold between the two groups can have two causes:

- The treatment effect, caused by the price cut.
- The business size.

In this example, the causal inference challenge is to untangle both causes.

The first technique to tackle the causal inference problem is through randomization control trials, which have been shown in Figure 10. Randomly dividing units into treatment and control groups before the intervention takes place makes both groups, on average, similar to each other. Applied to this study, it would mean randomly sending reports to some clients and not to others. The ones receiving constitute the treatment group, and the others are the control group.

However, RCTs are in general difficult to perform due to ethical concerns, sample size, logistical and resource constraints, etc.

In contrast to RCTs, causal inference applied to observational data is when it is possible to check the units that got the treatment, but it is unknown how the treatment was assigned.

In the presence of observed data, the problem of causal inference can be solved with *regression models* if some conditions are met. Regression gives an interpretation for each variable, which in the economy field is named the *ceteris paribus* effect. It refers to how the DV changes with the variation of one IV while keeping the other IV(s) constant. This effect is causal if all the confounders are controlled in the model. Otherwise, the model shows an association (correlation).

A confounding variable is a variable that correlates with both the variable treatment and the outcome variable. For instance, considering the study of patients taking antidepressant pills (the treatment, or DV) and their impact on the risk of suicide (the outcome, or IV). The severity of depression is a confounder. It is correlated to both the probability of taking antidepressants and the probability of suicide.

In literature, the no inclusion of confounders is referred to with different terms, including omitted variable bias, exogeneity, conditional independence assumption not met, etc.

In observational studies, it is very difficult to make sure that all confounders are accounted for. A very good knowledge of the subject under study is necessary but might not be sufficient. Therefore, the regression results are presented with phrases like “is associated with” and “is likely to cause”, rather than statements that imply causation, such as “causes” or “results in”.

In summary, the problem of causal inference can be solved with:

- Randomized control trials. But they are not always possible, feasible, ethical, or easy to do.
- If all the confounders are observed, the regression methods reveal the causal effect.
- Other more advanced methods include difference-in-difference, instrumental variables, and regression discontinuity. These have not been presented in this chapter.

Further information about causal inference can be found in the book (Facure, 2023).

Having the causality topic been present, it is relevant to introduce the pitfalls of making causal claims based on model-agnostic interpretation methods for machine learning models, such as the ones in Figure 13, presented by (Molnar et al., 2022). The article states that conventional supervised machine learning models are designed to exploit associations rather than represent causal links. They also state, “The practitioner must carefully assess whether sufficient assumptions can be made about the underlying data-generating process, the learned model, and the interpretation technique. If these assumptions are met, a causal interpretation may be possible. The challenge of causal discovery and inference remains an open key issue in the field of ML.”. This article’s position is aligned with the difficulties expressed in this chapter in being able to make causal inference claims.

4. RESULTS AND DISCUSSION

4.1. DESCRIPTIVE STATISTICS

Following the data assembly, as described in Chapter 3.1, the dataset is explored and presented below.

The total number of observations is shown in the following table.

Table 8: Number of Observations

Dependent Variable	Number of Observations	Total Number of Observations
RFQ=0	387746	421218
RFQ=1	33472	

Table 9 displays the total distinct values for each of the main variables that comprise this study.

Table 9: Data variable count

Variable	Total Number of Distinct Values
Client	859
Report	606
ISIN ^{*1}	793
Ticker ^{*1}	320

Note ^{*1}: An International Securities Identification Number (ISIN) is a twelve-digit alphanumeric code that uniquely identifies a specific security. Bonds can be described by their issuer, also known as the ticker of the bond. Therefore, the same ticker can have different bond ISINs, which allows for differentiation of bonds' currency issuance, maturity dates, etc.

Figure 15 and Figure 16 represent the distribution frequency of the date difference in days between the day the client makes an RFQ and the day the client opens only the email or downloads the report, respectively. The results show a smaller number of RFQs as the days difference increases. In Figure 16, it is relevant to highlight the initial peak bar that shows downloading is associated with a trade immediately after.

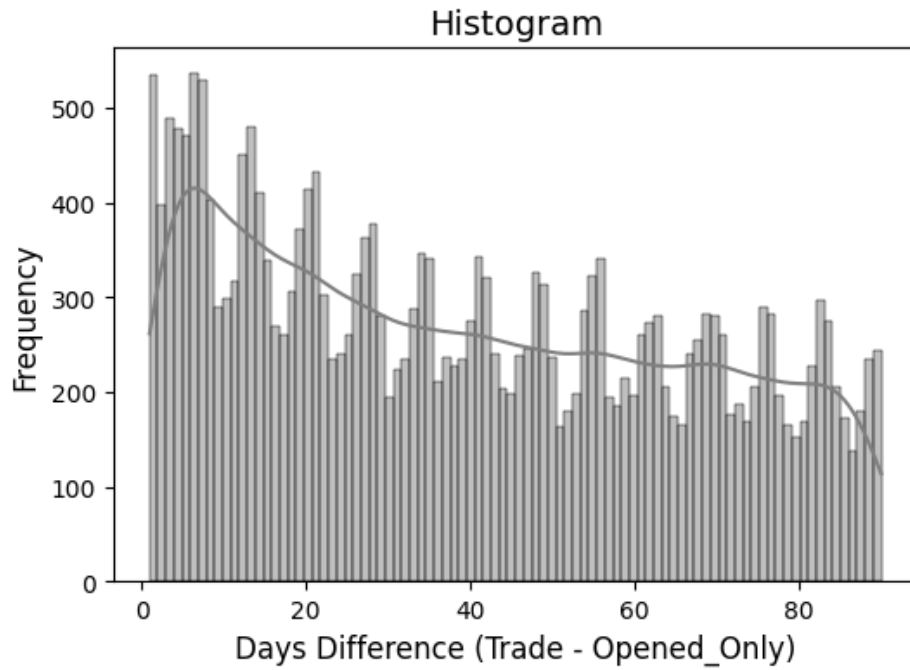


Figure 15: Histogram days difference (Trade – Opened_Only)

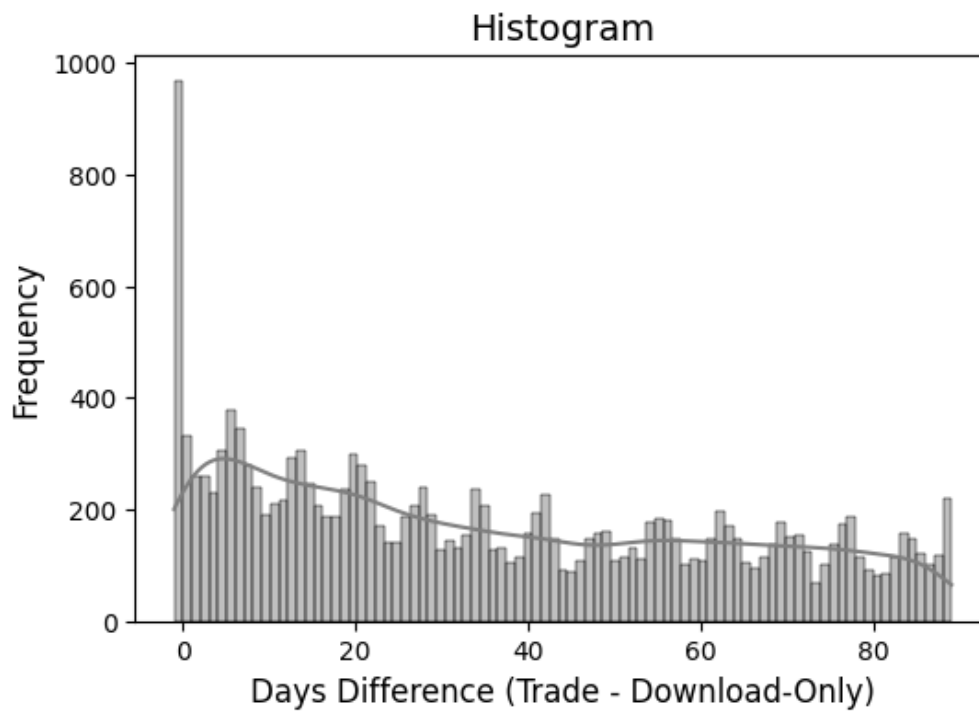


Figure 16: Histogram days difference (Trade – Downloaded_Only)

Following the methodology presented in Chapter 3.2, the study has focused on finding intuitively relevant variables that were associated with the response variable. These variables, forming the final model, are described in Table 10.

Table 10: Final model variables meaning

Category	Profile	Variables Name	Meaning	Type
Independent Variables	Client Identification	T10_only	The top ten bank clients worldwide, which are also called Titanium 10.	Binary
		C100_only	The top ninety clients worldwide, after the Titanium 10.	
		Customer_Sector	The sector in the market the client belongs to.	Categorical Nominal
	Client Telemetry	Opened_only	The client opened the email sent with a report attached but had not downloaded it.	Binary
		Downloaded_only	The client downloaded the report attached to the email.	
	Client Historical Behaviour	Purchased_ISIN_Before_Only	Clients have bought that ISIN before the first time it was ever suggested in a report.	
		Purchased_Ticker_Before_Only	Clients have bought that Ticker before the first time it was ever suggested in a report but have not bought that ISIN before, meaning the variable "Purchased_ISIN_Before_Only" would be null.	
		Downloaded_interaction	The ratio between the number of reports downloaded against the total reports received.	Categorical Ordinal
Dependent Variable	-	RFQ	Request for Quotation	Binary

The binary variables can either take a value of 0 or 1, which indicates that the attribute is absent or present, respectively.

It is important also to highlight how the client telemetry variables have been encoded.

The remaining variables, *Customer_Sector* and *Downloaded_Interaction*, are not binary. The values they can adopt are presented. Starting with the *Customer_Sector* variable, out of the 859 total different clients, their sector distribution is shown in Figure 17.

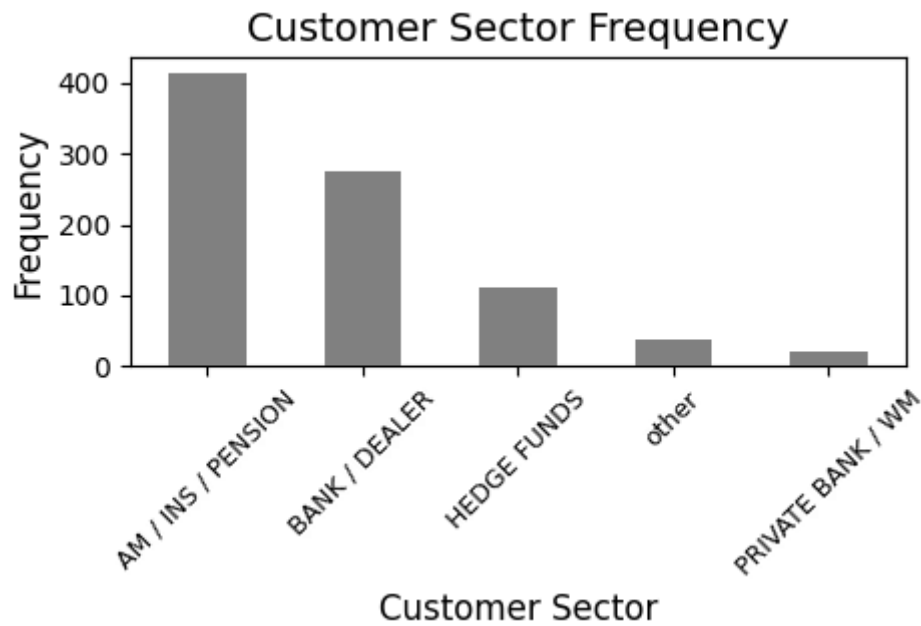


Figure 17: BNP customers sector frequency

With regards to the variable *Downloaded_Interaction*, for each client, the ratio between the total number of reports downloaded against the total reports received is calculated. After that, the clients are ranked in order based on the ratio value, and an integer number from 0 to 3 is assigned depending on the quartile they have fallen on. It means, for instance, that a value of 0 or 3 represents, respectively, the least 25% of clients and the most 25 % of clients concerning the download interaction type.

4.2. FEATURE SELECTION

4.2.1. Independent variables included in the final model

This chapter focuses on describing the feature selection process based on the methodology presented in Chapter 3.2. The variable *t10_only* is the first one to be presented. The table below shows its contingency table against the response variable.

Table 11: Variable *t10_only* contingency table

<i>t10_only</i>	0	1
<i>rfq</i>		
0	374485	13261
1	29891	3581

The frequency and proportion bar charts depicted below are based on the contingency table values. The proportion bar chart visualization allows to understand that when the client is *T10*, the proportion of RFQ =1 increases significantly. Therefore, the variable *t10_only* is moderately associated with the response variable.

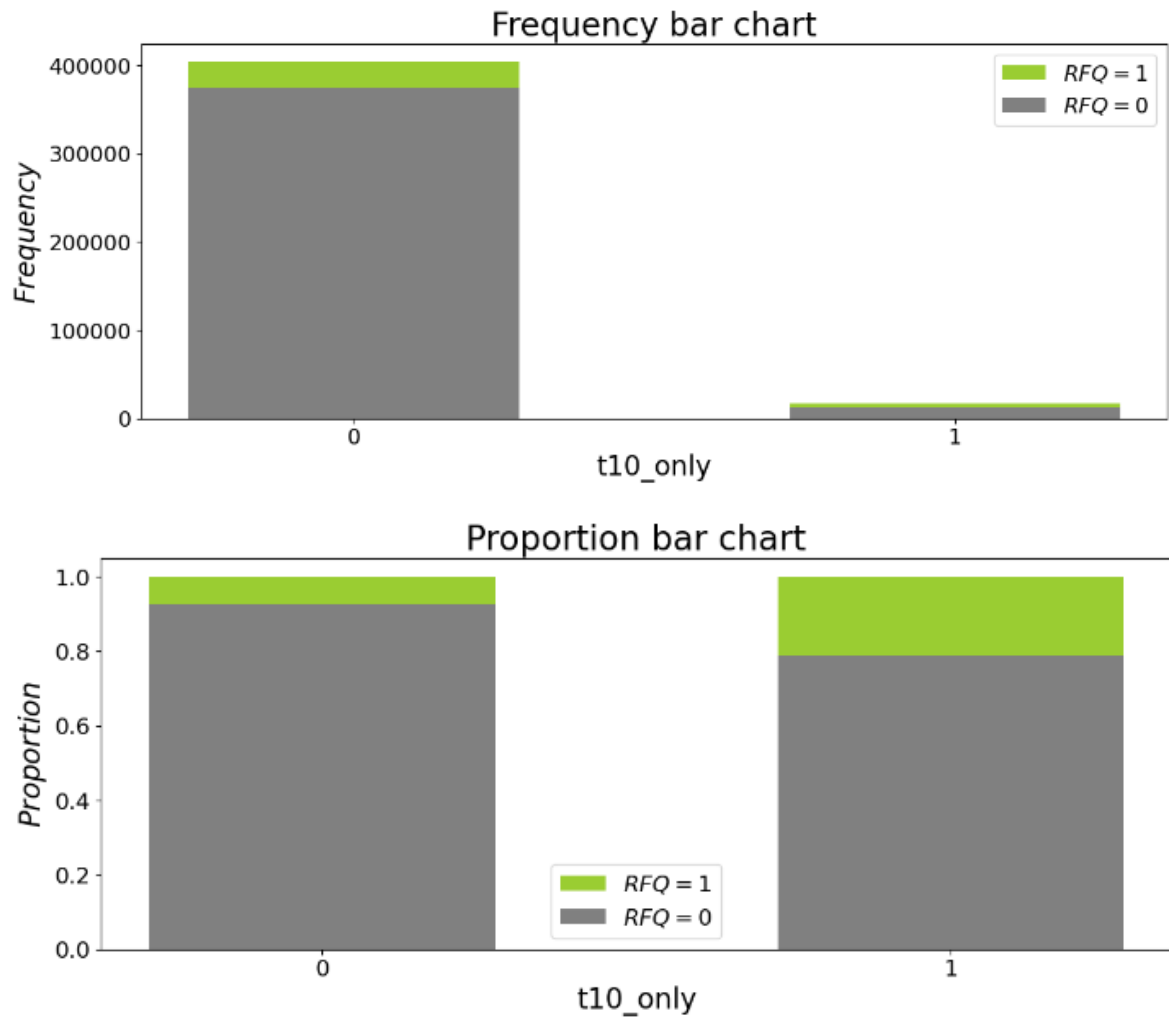


Figure 18: *T10_only* variable frequency and proportion bar chart against the response variable

The other IV(s) have been assessed following the same reasoning, meaning, by creating the contingency tables and visualizing the frequency and proportion bar charts. The visualizations are presented in Figure A-1 in Appendix A.

In addition, to quantify the association effect between DV and the different IV(s), the *Cramer's V* values are shown in the table below.

Table 12: Independent variables and their *Cramer's V* values

Independent Variable	Cramer's V with the Dependent Variable
T10_only	0,10
C100_only	0,07
Customer_Sector	0,09
Opened_only	0,02
Downloaded_only	0,21
Purchased_ISIN_Before_Only	0,41
Purchased_Ticker_Before_Only	0,06
Downloaded_interaction	0,19

Based on the results in Figure A-1 in Appendix A and Table 12, the following points are relevant to highlight:

- The variable *Purchased_ISIN_Before_Only* is the one with the highest association with the dependent variable. That can be inferred by its Cramer's V value, which is the highest, or also by the highest proportion of RFQs when the respective IV is present.
- The variable *Opened_only* is the one with the lowest association with the dependent variable. Anyway, it is kept on the model since it represents a variable that is related to the research question, and therefore its interpretation is important.
- When the variable *Purchased_Ticker_Before_Only* is present, a decrease in the proportion of positive RFQs is observed. The Cramer's value shows a positive value due to the nature of its calculation, but Figure A-3 shows a negative correlation with the dependent variable. Another way to check the association is through a univariate LR model, as shown in Figure 19, which shows a negative variable β coefficient.

Logit Regression Results						
Dep. Variable:	rfq	No. Observations:	421218			
Model:	Logit	Df Residuals:	421216			
Method:	MLE	Df Model:	1			
Date:	Fri, 22 Sep 2023	Pseudo R-squ.:	0.005746			
Time:	19:29:38	Log-Likelihood:	-1.1620e+05			
converged:	True	LL-Null:	-1.1687e+05			
Covariance Type:	nonrobust	LLR p-value:	4.901e-294			
	coef	std err	z	P> z	[0.025	0.975]
const	-2.3070	0.007	-345.203	0.000	-2.320	-2.294
purchased_ticker_before_only	-0.4572	0.013	-35.622	0.000	-0.482	-0.432

Figure 19: Univariate *Purchased_Ticker_Before_Only* logistic regression model

Till this point, the IV(s) have been studied concerning the association with the DV. Following Chapter 3.3.1.1, the subset of IV(s) can be chosen based on the metric AIC. The table below shows five different models, which differ in the number of IV(s) included.

Table 13: Different models with respective AIC values

Model	Description	AIC
A	All IV(s)	167028
B	All IV(s) - <i>opened_only</i>	167162
C	All IV(s) - <i>purchased_ticker_before_only</i>	172090
D	<i>downloaded_only</i> + <i>purchased_isin_before_only</i>	176516
E	All IV(s) - <i>downloaded_only</i> + <i>purchased_isin_before_only</i>	210428

- Notes:
- The mathematical signs (-) and (+) represent a variable excluded or included in the model, respectively.
 - All IV(s) refer to the model with all the IV(s) mentioned on Table 10

Based on the results, the following points are relevant to mention:

- The model A and C comparison allows us to understand that the variable *purchased_ticker_before_only* has a significant impact on the AIC value. It is relevant since it is the variable with the second smallest Cramer's V value in Table 12.
- The model A and B comparison allow to understand the variable *opened_only* has a small impact on the AIC value, which is aligned with the respective smallest Cramer's V value in Table 12.
- The model D, with only the two variables presented, shows a decent AIC value compared to the best model A. These variables are the ones with the highest association with the DV, as shown in Table 12.

- The model E, with the remaining IV(s) apart from the ones mentioned at the previous point, shows an AIC value far from the best model A.

The model A is the one with the smallest AIC value, and therefore, it is the one selected.

4.2.2. Independent variables not included on the final model

Other potential IV(s) have been studied but have not been included in the final model. Their frequency and proportion bar charts against the response variable are presented in Figure A-2 in Appendix A.

Starting with the bond characteristics, the variable *maturitydate*, related to the bond maturity, has been encoded as shown in the table below.

Table 14: Bond maturity date variable encode and description

Bond Maturity date		
Code	Classification	Description
0	Short-term	<= 3 years
1	Medium-term	4-10 years
2	Long-term	> 10 years

The variable *coupontype* has two main values in essence. The fixed and variable coupon rates, with the latter meaning the coupon rate alters in value during the bond's maturity lifetime.

The variable *ratinggrade* has two main values either. The investment grade (IG) and high-yield (HY) bonds. These ratings are defined by independent, accredited rating organizations, as presented in Chapter 1.

The variable *direction_indicator*, which indicates the direction of the trade suggestion with a buy or a sell, has also been studied.

These variables have the *Cramer's V* values shown in the table below. They are small, meaning they have a small association with the DV, justifying their inclusion in the final model.

Table 15: Variables *Cramer's V* values not included in the final model

Variable	Cramer's V with the Dependent Variable
maturitydate	0,01
coupontype	0,03
ratinggrade	0,02
Direction_indicator	0,03

The next variables analysed are related to datetime. These variables shall display a similar pattern across time, otherwise, they introduce noise to the models. As presented in Chapter 3.1, the data assembled is from the years 2021 and 2022. Based on that, the datetime variables have been compared between these two time periods to check their consistency.

This study acknowledges that consistency should be checked across multiple years, but there was no available data for doing so.

The variable *report_sent_weekday* identifies the weekday the report has been sent. The proportion bar charts are shown in the figure below, with the numerical encoded 0 to 6 matching the weekdays from Monday to Sunday.

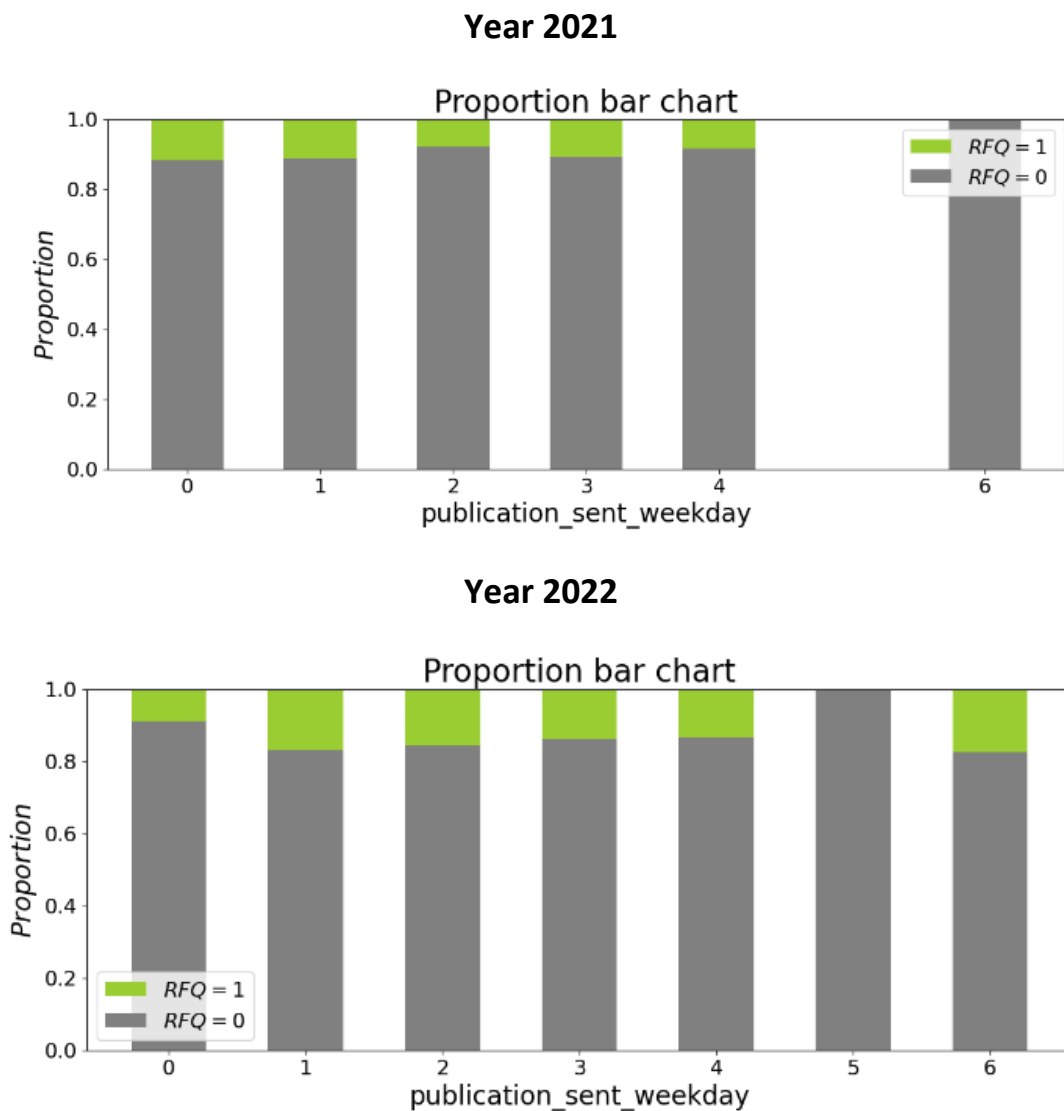


Figure 20: Variable *report_sent_weekday* proportion bar chart against the dependent variable

The results show Monday is the one with the highest and lowest proportion of RFQs in the years 2021 and 2022, respectively. It shows no consistency, and therefore, this variable has not been included in the model.

The variable *report_downloaded_weekday* identifies the weekday on which the report has been downloaded. Most of the client telemetry events have not downloaded the report, therefore most of the observations would have a null value under this feature. Hence, this variable has not been considered.

Moving to the date month type, the variable *report_sent_month* identifies the month the report has been sent. The proportions bar charts are shown in Figure A-4. The results show no consistency comparing both years and therefore, this variable has not been included in the final model.

4.3. MODELS RESULTS OUTCOME

4.3.1. Standard Statistics Logistic Regression

4.3.1.1. Results Interpretation

The LR model results, with the IV(s) defined in Chapter 4.2, are presented in Figure 21. It is the output from the Python library *StatsModels*.

The main points relevant to highlight are:

- The pseudo-r-squared value is 0.2855. It means the model fits well with the observations set as presented in Chapter 3.3.1.1.
- The variable *downloaded_interaction_medium* has a 0.664 p-value. The null hypothesis fails to be rejected for the usual significant levels of reference. Anyway, the variable is kept since there is nothing wrong with having it in the model. The dummy variable is part of the general variable *download_interaction*, and the other dummies (*downloaded_interaction_small* and *downloaded_interaction_high*) are significant.
- All the remaining variables have a p-value equal to zero. It means the null hypothesis is rejected and these variables are significant.
- The only variable with a negative β coefficient, with the meaning explained in Chapter 3.3.1.3, is *downloaded_interaction_small*.
- The variable with the highest β coefficient, with the meaning explained in Chapter 3.3.1.3, is *purchased_isin_before_only*.
- The variable *purchased_ticker_before_only* changes the β coefficient sign from the LR uniregression to the LR multiregression, shown respectively in Figure 19 and Figure 21. One of the reasons can be the addition of the confounding variable *purchased_isin_before_only* in the model. The inclusion of a confounding variable can have the effect of “adjusting” the relationship between the original independent variable and the dependent variable. This effect is shown in Figure A-5.

Logit Regression Results							
Dep. Variable:	rfq	No. Observations:	421218				
Model:	Logit	Df Residuals:	421204				
Method:	MLE	Df Model:	13				
Date:	Tue, 26 Sep 2023	Pseudo R-squ.:	0.2855				
Time:	18:14:59	Log-Likelihood:	-83500.				
converged:	True	LL-Null:	-1.1687e+05				
Covariance Type:	nonrobust	LLR p-value:	0.000				
	coef	std err	z	P> z	[0.025	0.975]	
const	-6.0811	0.072	-83.925	0.000	-6.223	-5.939	
t10_only	0.8421	0.025	33.462	0.000	0.793	0.891	
c100_only	0.3649	0.014	25.972	0.000	0.337	0.392	
opened_only	0.1759	0.015	11.663	0.000	0.146	0.205	
downloaded_only	1.5033	0.021	71.339	0.000	1.462	1.545	
purchased_isin_before_only	3.4470	0.024	146.428	0.000	3.401	3.493	
purchased_ticker_before_only	1.5492	0.024	63.882	0.000	1.502	1.597	
downloaded_interaction_small	-0.1572	0.026	-6.080	0.000	-0.208	-0.107	
downloaded_interaction_medium	0.0104	0.024	0.434	0.664	-0.037	0.057	
downloaded_interaction_high	0.1955	0.024	8.265	0.000	0.149	0.242	
customersector_AM / INS / PENSION	1.2511	0.068	18.368	0.000	1.118	1.385	
customersector_BANK / DEALER	1.4782	0.069	21.552	0.000	1.344	1.613	
customersector_HEDGE FUNDS	0.9736	0.072	13.565	0.000	0.833	1.114	
customersector_PRIVATE BANK / WM	1.7045	0.075	22.701	0.000	1.557	1.852	

Figure 21: Final logistic regression model results given by *Statsmodels* library

As explained in Chapter 3.3.1.3, the average marginal effects provide a more intuitive interpretation than the β coefficients. The results are shown in Figure 22. The column dy/dx values are interpreted as the average variation probability on the dependent variable, requesting an RFQ, when the respective independent variable is present or not, and keeping the other independent variables fixed.

Logit Marginal Effects						
=====						
Dep. Variable:	rfq					
Method:	dydx					
At:	overall					
=====						
	dy/dx	std err	z	P> z	[0.025	0.975]

t10_only	0.0583	0.002	28.130	0.000	0.054	0.062
c100_only	0.0212	0.001	25.429	0.000	0.020	0.023
opened_only	0.0099	0.001	11.507	0.000	0.008	0.012
downloaded_only	0.1200	0.002	54.779	0.000	0.116	0.124
purchased_isin_before_only	0.3618	0.003	126.715	0.000	0.356	0.367
purchased_ticker_before_only	0.0887	0.001	70.991	0.000	0.086	0.091
downloaded_interaction_small	-0.0087	0.001	-6.255	0.000	-0.011	-0.006
downloaded_interaction_medium	0.0007	0.001	0.503	0.615	-0.002	0.003
downloaded_interaction_high	0.0112	0.001	8.151	0.000	0.008	0.014
customersector_AM / INS / PENSION	0.0730	0.004	17.771	0.000	0.065	0.081
customersector_BANK / DEALER	0.0992	0.005	18.368	0.000	0.089	0.110
customersector_HEDGE FUNDS	0.0693	0.006	11.459	0.000	0.057	0.081
customersector_PRIVATE BANK / WM	0.1454	0.009	16.846	0.000	0.128	0.162
=====						

Figure 22: Final model logistic regression average marginal effects

The results interpretation shows that downloading a report is associated with an average 12.00% probability increase in requesting an RFQ, while opening an email is linked to only a 0.99% average probability increase.

Looking at all values, the variable *purchased_isin_before_only* is the one with the highest impact on the dependent variable. The results are aligned with the coefficients shown in Figure 21. If the client has purchased a specific bond security ISIN before the first time was ever suggested by the bank, it is related to an average increase of 36.18% probability in requesting an RFQ.

Other important results relevant to highlighting are:

- A T10 client is more likely to request an RFQ than a C100 one.
- The type of client more likely to request an RFQ is a private bank.

4.3.1.2. Model Diagnostics

As per the diagnostic measurements defined in Chapter 3.3.1.2, the residual plot is presented in the figure below.

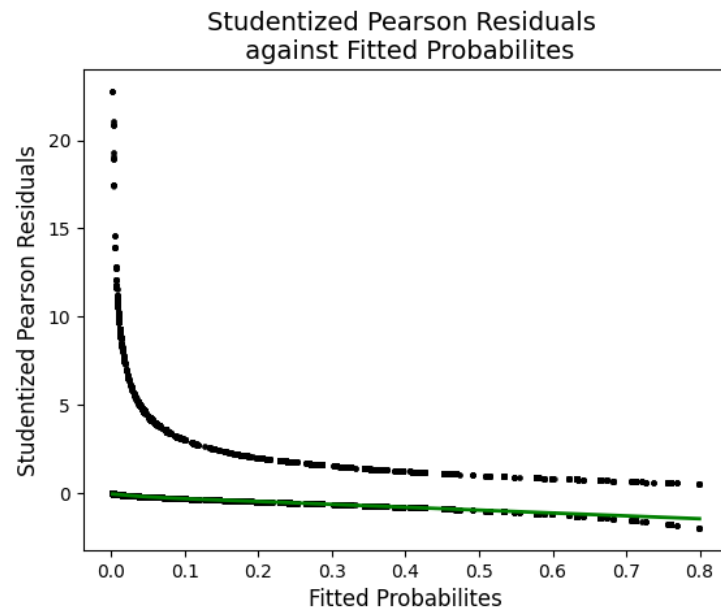


Figure 23: Studentized Pearson residuals against fitted probabilities

It appears the plots approximate a horizontal line with a 0 intercept. This suggests that there is no significant model inadequacy.

On the other hand, Figure 24 shows all independent variables with VIF values below the threshold of 5, indicating there is no multicollinearity issue.

	variables	VIF
0	rfq	1.367751
1	t10_only	1.111419
2	c100_only	1.563580
3	opened_only	1.816115
4	downloaded_only	1.299528
5	purchased_isin_before_only	1.726067
6	purchased_ticker_before_only	1.820542
7	downloaded_interaction_small	2.159972
8	downloaded_interaction_medium	2.422647
9	downloaded_interaction_high	3.228749
10	customersector_AM / INS / PENSION	3.939596
11	customersector_BANK / DEALER	2.253067
12	customersector_HEDGE FUNDS	1.604488
13	customersector_PRIVATE BANK / WM	1.134761

Figure 24: Independent variables VIF values

4.3.2. Machine Learning

The train and test dataset observations' numbers are shown in the table below. Both datasets have the same class distribution, as shown in Table 7.

Table 16: Train and test dataset number of observations.

Dataset	Number of Observations	% Of Full Dataset
Train	336974	80%
Test	84244	20%

The LR model trained on the training dataset has been the first one studied. Its results are presented in Table 17.

Afterwards, the random undersampling technique covered in Chapter 3.3.2 has been applied to the LR model as well. The results are displayed in Figure 25. Different undersampling proportions have been calculated, which are marked with a dot in the plot. The undersampling proportion values can vary from 0 to 1, which correspond, respectively, to the entire dataset class distribution or one with the same number of observations on both binary DV classes.

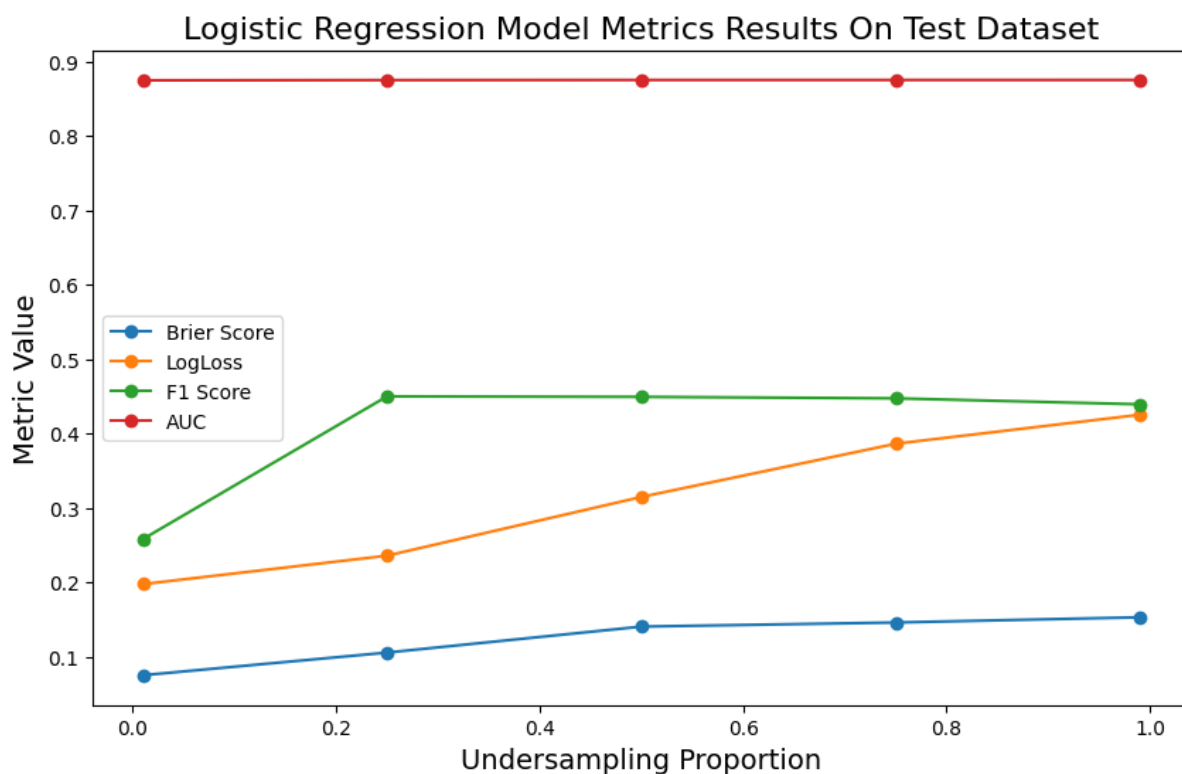


Figure 25: Logistic regression model metrics results on the test dataset, trained with different class distributions

The main points relevant to highlight are:

- As we reduce the majority class number of observations on the training dataset, the probability metrics *LogLoss* and *Brier Score* get worse. It means the predicted probabilities are further deviating from the true underlying probability of the sample, which is the test dataset.
- The F1 score metric increases initially and then stabilizes. The increase is justified by the model better predicting the minority class, as we reduce the number of observations from the majority class on the training dataset. The F1 score has been calculated with the probability threshold to define the predicted class equal to 0.50.
- The AUC score is approximately constant.

These results show the importance of selecting the correct metric in light of the research project's problem. If the goal was to predict the observations' class, the F1 score metric should guide the model choice. However, considering the research problem in the study and the reasoning presented in Chapter 3.3.2, the *probability metrics* are the ones relevant to this research project. Based on that, the random undersampling technique does not improve the results.

Finally, the RF with default parameters has been applied.

Both the RF and LR models' classification metrics results on the test dataset are presented in Table 17.

Table 17: Classification metrics results on the test dataset

Model	Average Predicted (RFQ=1) ^{*1}	Metrics			
		Threshold	Ranking	Probability	
		F1-score	AUC	Brier Score	Log Loss
Logistic Regression	7,92%	0,26	0,87	0,057	0,197
Random Forest	7,91%	0,25	0,88	0,056	0,195

*1 – it is the mean of the predicted probabilities for each observation belonging to class 1.

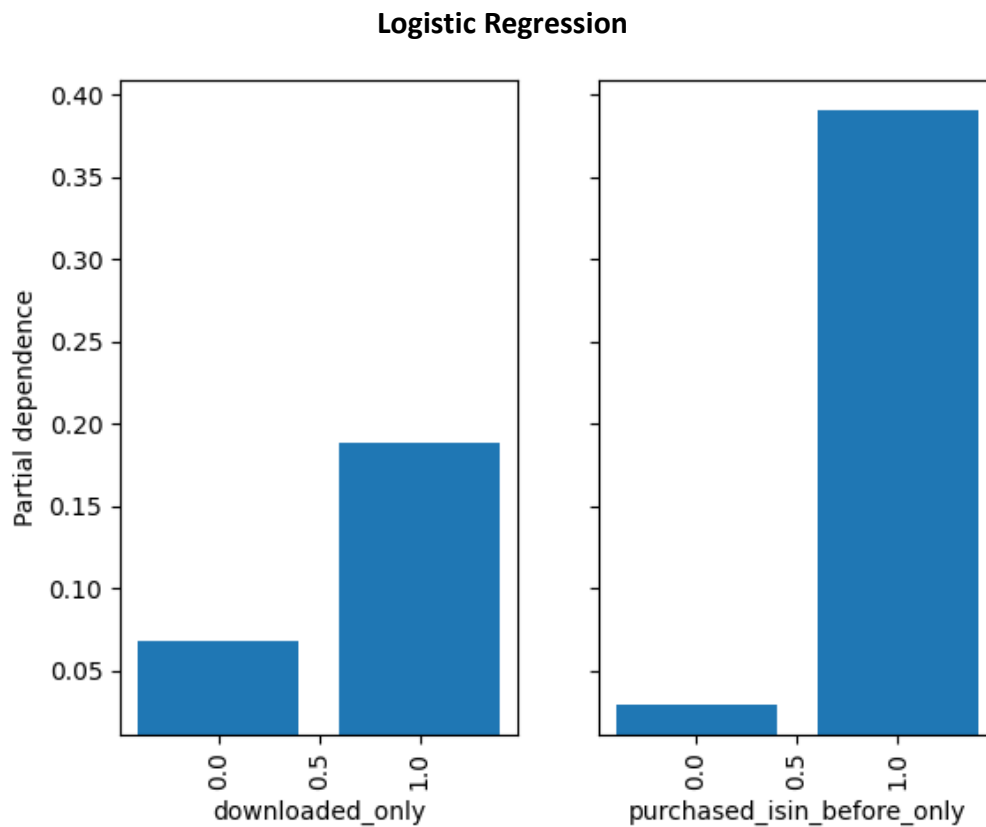
The results show a marginal improvement using the RF model compared to the LR one. A possible reason for the result's similarity is the use of structured data with meaningful features. Rudin (2019) states “.... problems that have structured data with meaningful features, there is often no significant difference in performance between more complex classifiers (deep neural networks, boosted decision trees, random forests) and much simpler classifiers (logistic regression, decision lists) after preprocessing.”. Based on that, no other ML models have been tried.

The RF feature importance results are presented in the figure below. The two most important IV(s) are aligned with the LR coefficients magnitude shown in Figure 21. The remaining IV(s) are not aligned.

<code>purchased_isin_before_only</code>	0.630442
<code>downloaded_only</code>	0.144939
<code>downloaded_interaction_high</code>	0.038421
<code>t10_only</code>	0.032058
<code>purchased_ticker_before_only</code>	0.031507
<code>c100_only</code>	0.026275
<code>opened_only</code>	0.018043
<code>customersector_AM / INS / PENSION</code>	0.017671
<code>customersector_BANK / DEALER</code>	0.016006
<code>customersector_HEDGE FUNDS</code>	0.013998
<code>downloaded_interaction_small</code>	0.012844
<code>downloaded_interaction_medium</code>	0.010072
<code>customersector_PRIVATE BANK / WM</code>	0.007722

Figure 26: Random Forest feature importance results

The ML models results are interpreted through PDPs, as presented in Chapter 3.3.2.1. The PDFs are applied to the full dataset. The two most impactful variables' results, *downloaded_only* and *purchased_isin_before_only*, are the only ones presented.



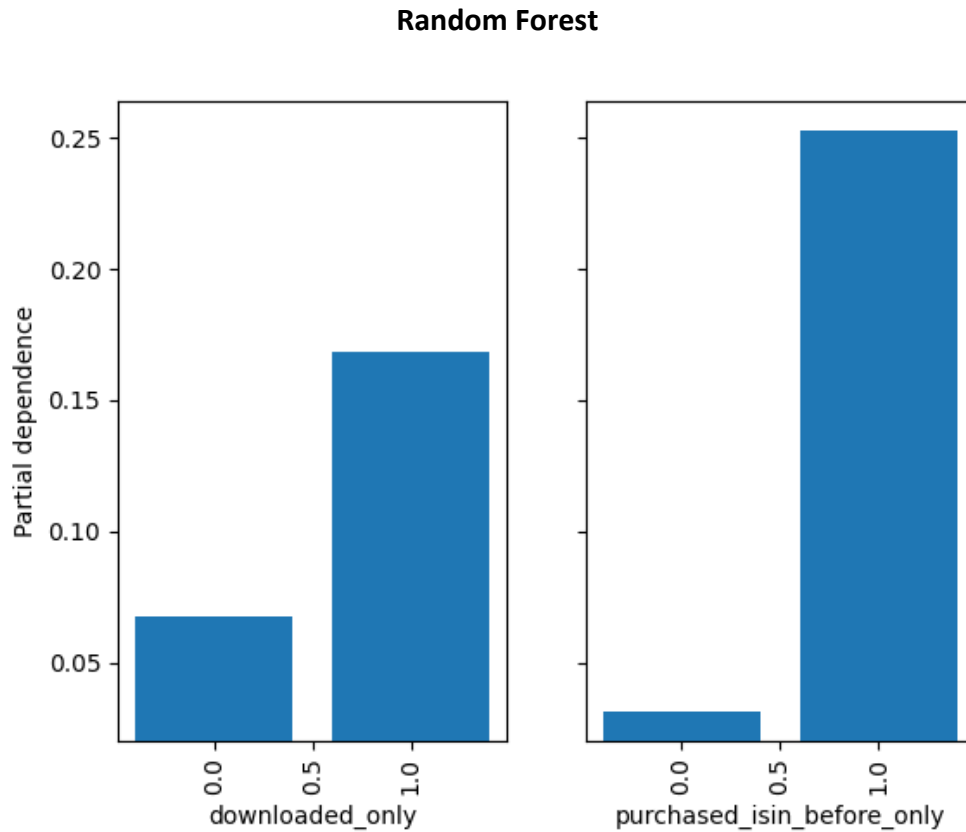


Figure 27: PDP of 2 IV(s) using logistic regression and random forest models on the entire dataset

Table 18: PDP values

Model	Dependent Variable					
	downloaded_only			purchased_isin_before_only		
	0	1	Δ^{*1}	0	1	Δ^{*1}
Logistic Regression	0,0682	0,1881	0,1199	0,0294	0,3910	0,3616
Random Forest	0,0678	0,1687	0,1008	0,0316	0,2528	0,2212

Note ^{*1}: It is the difference between PDP values with the respective feature value equal to 1 and 0

The following points are relevant to highlight:

- The PDP logistic regression results (Δ) are the same as the ones presented in Figure 22. It is expected since both are calculated using the marginal effects technique.
- The results between the two models are similar for the variable *downloaded_only* but differ for the variable *purchased_isin_before_only*. The latter variable has a likely impact on the DV of 36,16% and 22,12%, respectively, using the LR and RF models. This difference stems from each different model assumption and the way the probabilities are calculated, which have been explained in Chapter 3.3.

It is interesting to notice that despite LR and RF models outputting the same classification metric results on the test dataset, as shown in Table 17, the PDP values for the variables assessed are different, especially for the variable *purchased_isin_before_only*.

With reference to the research question, the answer lies in the interpretation of the variable *downloaded_only*. For this variable, the LR and RF models provide slightly different results, respectively, with 11,99% and 10,08% likely impacts on the DV. The problem now lies in knowing which one is more correct or should be considered. The difference in results is a topic that could be further investigated in a future study. At this time, the LR results are the ones that will be considered for the following reasons:

- The Random Forest algorithm is not deterministic since the ensemble method introduces randomness due to bootstrapping and feature selection.
- LR are trained using a probabilistic framework, like the maximum likelihood estimation method.

One question that could be raised is if a causality claim like “reports cause on average a 12% increase in requesting an RFQ” is valid. Considering the causal inference methodology introduced in Chapter 3.3.3, all the confounders would have to be accounted for. It is a difficult scenario to achieve. For instance, the following confounders are difficult to measure and include:

- Client satisfaction and trust in BNP Paribas
- Other market makers’ reports with their tradeideas

With regard to the first example, it is interesting to further investigate. The most direct way to measure client satisfaction would be through surveys. An indirect way is to measure the number of RFQs per client, whether the direction is aligned with or against the tradeidea direction. Figure 28 and Figure 29 display, respectively, the top 10 clients whose number of RFQs’ direction is aligned with and against the tradeidea direction. The names of the clients on the X-axis are not displayed due to confidentiality reasons.

Comparison Top 10 Client Number of RFQs Matching the Same Trade Suggestion Direction Against Matching the Opposite Direction for Year 2022

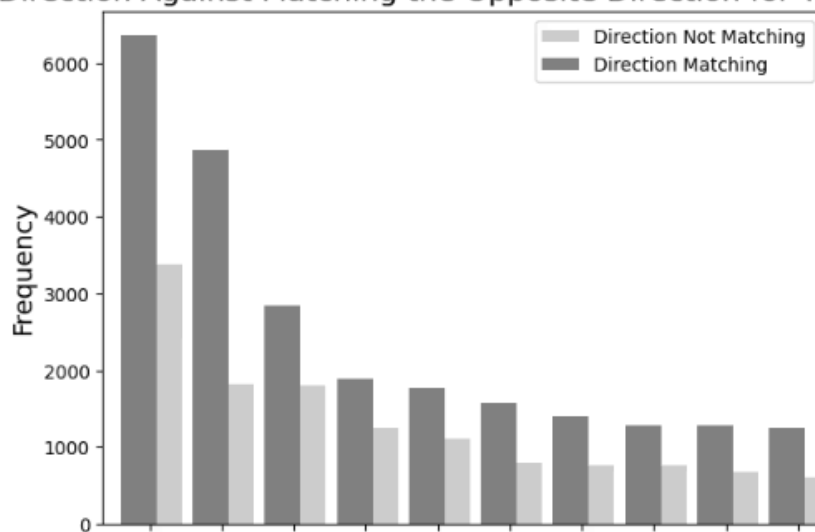


Figure 28: Comparison of the top 10 client numbers of RFQs matching the same trade suggestion direction against matching the opposite direction for 2022

Comparison Top 10 Client Number of RFQs Matching the Opposite Trade Suggestion Direction Against Matching it for Year 2022

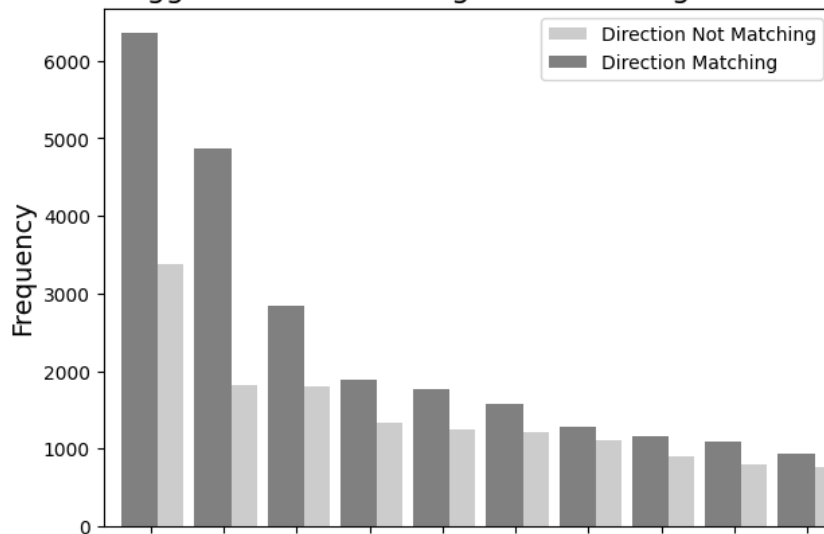


Figure 29: Comparison of the top 10 client numbers of RFQs matching the opposite trade suggestion direction against matching it for 2022

The results show in both cases that each client has always had a higher number of RFQs matching the trade idea direction. Overall, it is a sign the clients trust BNP trade ideas.

Another question is to what degree the variable `opened_only` could also measure the reports' impact. Reports are comprised of trade ideas. BNP bond analysts have confirmed that some emails might have trade ideas in the email body itself, but others do not. Based on that and knowing that an email could also be opened to delete immediately afterwards, this variable is not considered relevant to measure the reports' impact.

4.4. BUSINESS RESULTS OUTCOME

One interesting exercise is to calculate the expected decrease in the total number of RFQs if the clients have not downloaded any reports. It is done by observing the mean predicted probability difference of all observations belonging to class 1 between the original dataset and the modified original dataset with the IV `downloaded_only` equal to zero. The LR model, presented in Chapter 4.3.1, has been the one used. A total of 4742 RFQs would be expected not to be requested out of the total of 33472 original RFQs.

Another inquiry is concerning the reports' impact on clients' final action. After a client's RFQ, he can trade with BNP, with another institution, or not do any trade. The representative schema is shown initially in Figure 2.

Following a similar reasoning as the one presented in Chapter 4.2, the proportion bar chart of the variable `tradestatus`, which defines the final client trade type, against the variable `downloaded_only` is shown in Figure 30.

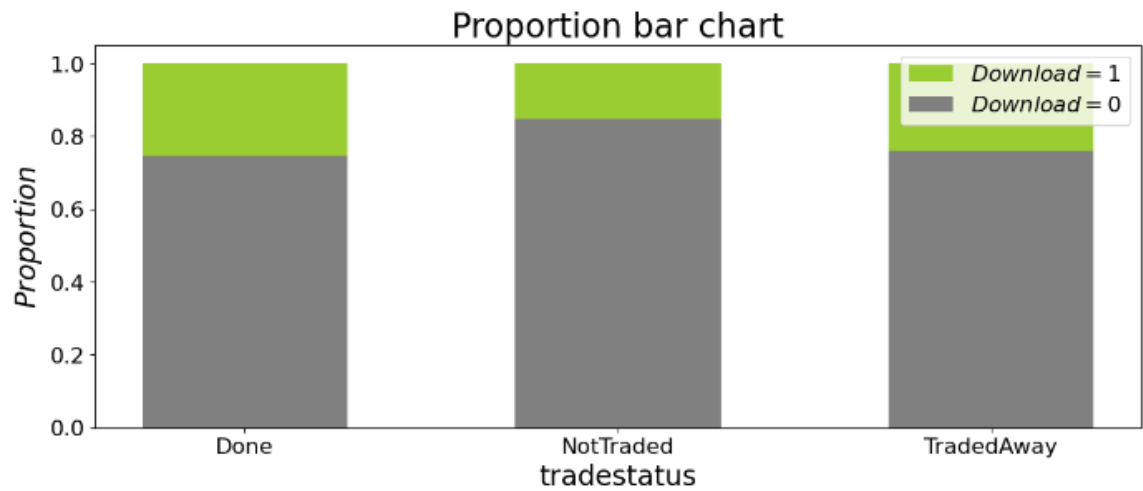


Figure 30: *Tradestatus* variable proportion bar chart against the variable *downloaded_only*

- Notes:
- Done = client has done a trade with BNP
 - NotTraded = client has not done a trade
 - TradedAway = client has done a trade with another market maker institution

The results show that downloading or not downloading a report has no impact on the client making a trade with BNP or another institution. On the other hand, it shows the client that downloads a report has a higher chances of trading with BNP or another institution than not doing it.

Other variables will influence the client's trading with BNP or another institution, but their study is outside of this research project's scope.

5. CONCLUSIONS AND FUTURE WORK

This dissertation aimed to understand the impact that the BNP Paribas research reports with bond trade suggestions had on clients' trade behaviour. The analysis concludes that clients who download the attached report to the email sent show, on average, a 12% higher likelihood of requesting an RFQ. However, the results show the reports do not influence whether the client conducts a trade with BNP or with another market maker's competitor.

The literature review effectively distinguishes between explanatory and predictive tasks. The research question fits the explanatory paradigm. A predictive task would be, for instance, trying to predict if the client would request an RFQ. In addition, the literature review has also shown that historically, classical statistical education focuses on explanatory statistical modelling and statistical inference, while machine learning focuses on predictive tasks.

Given the previous distinctions and the dissertation's nature, the initial use of the logistic regression model is justified. It is important to highlight that the underlying mathematical concept of LR remains the same in the case of a predictive task. The use of the LR model as an explanatory instrument had an impact on the way feature selection has been undertaken and on the metrics used to assess the best LR model.

Subsequently, the Random Forest (RF) model is introduced and compared against the LR model. The application of the holdout method, which consists of checking the prediction performance on the test dataset, aligns with machine learning practices, though, given the explanatory task, probability metrics are used rather than threshold metrics within the classification metric structure. Both models present similar performance on the test dataset, as evaluated by the Brier score and Log Loss probability metrics, leading to the exclusion of more complex machine learning models evaluation.

The study's impact of a single independent feature on the dependent feature RFQ has been done with the partial dependence plot (PDP) technique. Results differ between the RF and LR models, with the LR results taking precedence due to the model's probabilistic framework, namely, the maximum likelihood estimation (MLE) method.

In cases where the main task had a predictive intent, the random undersampling technique has emerged as the most effective, as indicated by the F1-score threshold metric.

Additional insights highlight that clients who purchased a specific bond security before the first time it was ever suggested by the bank have, on average, 36% higher chances of requesting an RFQ. However, variables related to bond characteristics and the timing of report delivery reveal very little association with a higher probability of RFQ requests.

The research project has also raised the question of how the research question would be answered with a causality claim like "reports cause on average a 12% increase in requesting an RFQ". This point is linked to the famous quote "Association is not causation". Considering the causal inference methodology introduced in Chapter 3.3.3 and knowing the study was based on observed data, a causality claim would be valid if all the confounders had been accounted for. In general, it is a difficult scenario to achieve. Based on that, the results linking the independent variables' impact on the dependent variable should not be read from a causal perspective.

Concerning future studies, in case the researcher aims to make a causal claim about the reports' impact, the following options could be considered. Firstly, randomised control trials, whereby the reports are randomly sent to some clients and not to others. Other options entail the use of advanced causality methods such as difference-in-difference, instrumental variables, and regression discontinuity.

Subsequent research could also further investigate and explain the different results outputted by the PDP technique when used with the LR and RF models. In addition, work on the discovery of other independent variables with an impact on the dependent variable RFQ or trying a different timeline difference to link trades and telemetry data, as shown in Table 3, could be valuable.

In summary, this dissertation confirmed that good domain knowledge about the business for assessing the right variables is as important as technical expertise about the necessary models to sort the problem out. In addition, a clear research problem definition is needed to clearly distinguish its explanatory or predictive nature, therefore contributing to the overall robustness of the study. Ultimately, the research project has managed to demonstrate the impact BNP research reports have on clients' trade behaviour.

REFERENCES

- Algesheimer, R., Borle, S., Dholakia, U. M., & Singh, S. S. (2010). The Impact of Customer Community Participation on Customer Behaviors: An Empirical Investigation. *Marketing Science*, 29(4), 756–769. <https://doi.org/10.1287/mksc.1090.0555>
- Barreau, B. (2020). Machine Learning for Financial Products Recommendation. Université Paris-Saclay <https://theses.hal.science/tel-02974918/>
- Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3). <https://doi.org/10.1214/ss/1009213726>
- Bose, I., & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, 195(1), 1–16. <https://doi.org/10.1016/j.ejor.2008.04.006>
- Brownlee, J. (2021). Tour of Evaluation Metrics for Imbalanced Classification. MachineLearningMastery.com. <https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/>
- Chaudhry, F., Hunt, R. J., Hariharan, P., Anand, S. K., Sanjay, S., Kjoller, E., Bartlett, C. M., Johnson, K. W., Levy, P. D., Noushmehr, H., & Lee, I. Y. (2020). Machine learning applications in the neuro ICU: a solution to big data mayhem? *Frontiers in Neurology*, 11. <https://doi.org/10.3389/fneur.2020.554633>
- Cohen, J. (2013). Statistical Power Analysis for the Behavioral Sciences. In *Routledge eBooks*. <https://doi.org/10.4324/9780203771587>
- De Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, J. K. U., & Von Wangenheim, F. (2020). Artificial intelligence and Marketing: Pitfalls and opportunities. *Journal of Interactive Marketing*, 51, 91–105. <https://doi.org/10.1016/j.intmar.2020.04.007>
- De Franco, G., Vasvari, F. P., & Wittenberg-Moerman, R. (2009). The informational role of bond analysts. *Journal of Accounting Research*, 47(5), 1201–1248. <https://doi.org/10.1111/j.1475-679x.2009.00348.x>
- Facure, M. (2023). *Causal Inference in Python*. O'Reilly Media, Inc.
- Feng, Y., Yin, Y., Wang, D., & Dhamotharan, L. (2022). A dynamic ensemble selection method for bank telemarketing sales prediction. *Journal of Business Research*, 139, 368–382. <https://doi.org/10.1016/j.jbusres.2021.09.067>
- Ferri, C., Hernández-Orallo, J., & Modroi, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1), 27–38. <https://doi.org/10.1016/j.patrec.2008.08.010>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5). <https://doi.org/10.1214/aos/1013203451>

- Gulum, M. A., Trombley, C. M., & Kantardzic, M. (2021). A review of Explainable deep learning cancer detection models in medical imaging. *Applied Sciences*, 11(10), 4573.
<https://doi.org/10.3390/app11104573>
- Hansen, K. B. (2020). The virtue of simplicity: On machine learning models in algorithmic trading. *Big Data & Society*, 7(1), 205395172092655. <https://doi.org/10.1177/2053951720926558>
- Harrell, F. E. (2015). Binary logistic regression. In Springer series in statistics (pp. 219–274).
https://doi.org/10.1007/978-3-319-19425-7_10
- Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems With Applications*, 124, 226–251.
<https://doi.org/10.1016/j.eswa.2019.01.012>
- Hensher, D. A., & Stopher. (1979). Behavioural Travel Modelling. Croom Helm, London.
<https://doi.org/10.4324/9781003156055>
- Hoepner, A. G. F., McMillan, D. G., Vivian, A., & Simen, C. W. (2020). Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective. *European Journal of Finance*, 27(1–2), 1–7.
<https://doi.org/10.1080/1351847x.2020.1847725>
- Hooker, G., & Mentch, L. (2021). Bridging Breiman’s Brook: From algorithmic modeling to Statistical learning. *Observational Studies*, 7(1), 107–125. <https://doi.org/10.1353/obs.2021.0027>
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. In Wiley series in probability and statistics. <https://doi.org/10.1002/9781118548387>
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *The American Economic Review*, 105(5), 491–495. <https://doi.org/10.1257/aer.p20151023>
- Kühn, M., & Johnson, K. (2013). Applied Predictive Modeling. In Springer eBooks.
<https://doi.org/10.1007/978-1-4614-6849-3>
- Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004). *Applied linear statistical models* (5th ed.). New York, NY: McGraw-Hill Irwin.
- Leeper, T. J. (2017) Interpreting regression results using average marginal effects with R’s margins.
<https://www.semanticscholar.org/paper/Interpreting-Regression-Results-using-Average-with-Leeper/9615c76bd5d81f7ebbbdac9714619863dc3a2337>
- MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., & Hines, J. E. (2018). Fundamental principals of statistical inference. In Elsevier eBooks (pp. 71–111).
<https://doi.org/10.1016/b978-0-12-407197-1.00004-1>
- Miguéis, V. L., Camanho, A. S., & Borges, J. (2017). Predicting direct marketing response in banking: comparison of class imbalance methods. *Service Business*, 11(4), 831–849. <https://doi.org/10.1007/s11628-016-0332-3>
- Molnar, C. (2020). *Interpretable Machine learning*. Lulu.com.

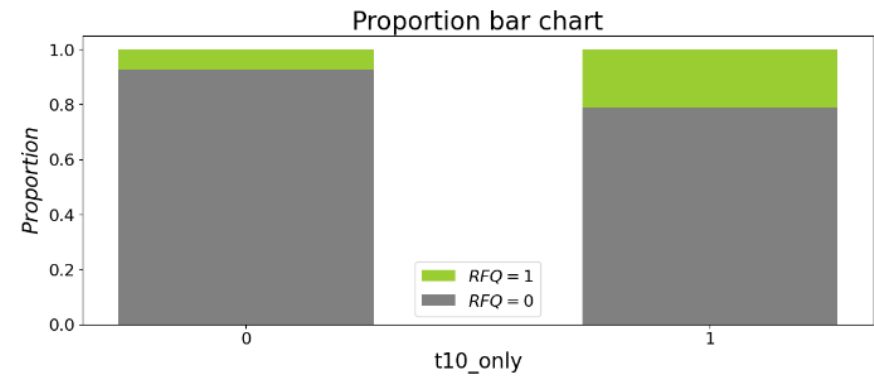
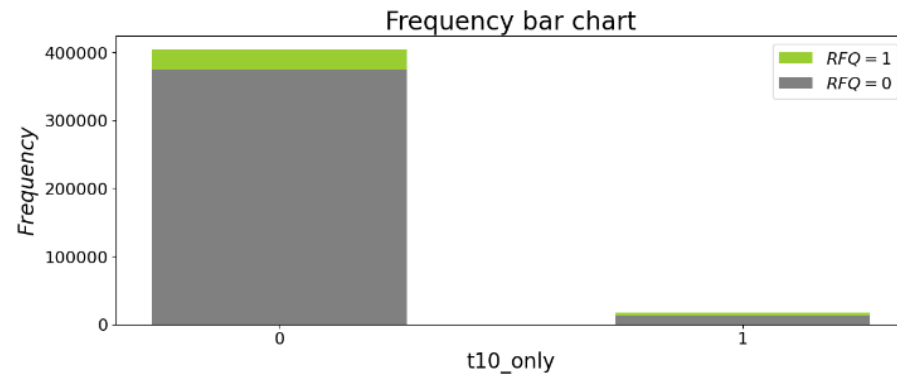
- Molnar, C., König, G., Herbringer, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., & Bischl, B. (2022). General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. In *Lecture Notes in Computer Science* (pp. 39–68).
https://doi.org/10.1007/978-3-031-04083-2_4
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
<https://doi.org/10.1038/s42256-019-0048-x>
- Shmueli, M. D., & Koppius. (2011). Predictive Analytics in Information Systems research. *Management Information Systems Quarterly*, 35(3), 553. <https://doi.org/10.2307/23042796>
- Song, J. W., & Chung, K. C. (2010). Observational studies: Cohort and Case-Control studies. *Plastic and Reconstructive Surgery*, 126(6), 2234–2242.
<https://doi.org/10.1097/prs.0b013e3181f44abc>
- Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, 4(3), 279–282. <https://doi.org/10.4300/jgme-d-12-00156.1>
- Xie, C., Zhang, J., You, Z., Xiong, B., & Wang, G. (2023). How to improve the success of bank telemarketing? Prediction and interpretability analysis based on machine learning. *Computers & Industrial Engineering*, 175, 108874. <https://doi.org/10.1016/j.cie.2022.108874>
- Zhang, K., Cai, F., & Zhengyu, S. (2021). Do promotions make consumers more generous? The impact of price promotions on consumers' donation behavior. *Journal of Marketing*, 85(3), 240–255.
<https://doi.org/10.1177/0022242920988253>

APPENDIX A

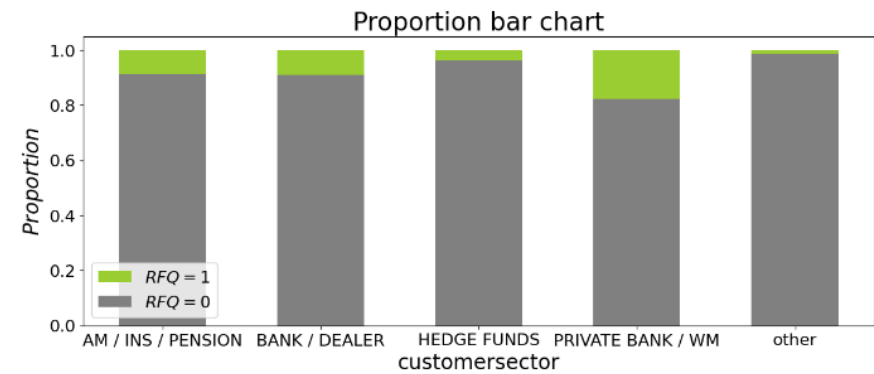
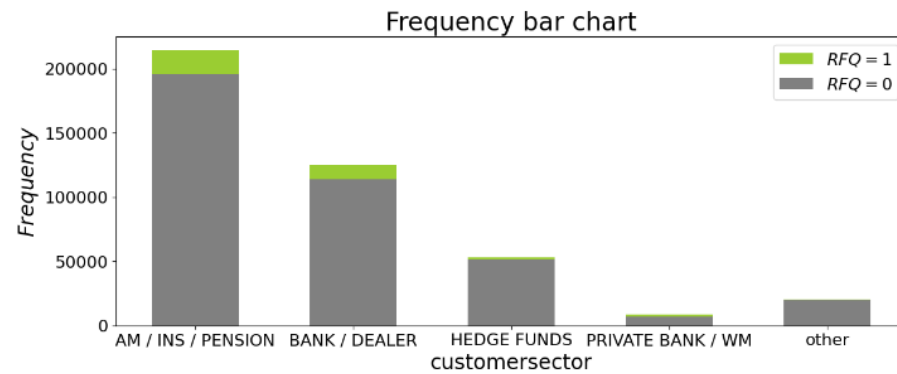
Table A-1: Example data linkage output

TRADE_ID	Client Name	Trade Date	Bond Security Name	TELEMETRY_ID	Report Title	Send	Opened	Downloaded	RFQ
1004	Confidential	2022-02-28	SPMIM 3 3/4 09/08/23	2500	Saipem - Where there's a well there's a way	1	1	0	YES
3050	Confidential	2022-02-28	ROLLS 4 5/8 02/16/26	3156	Credit Strategy RV - Closing Long Cash/CDS basis on € ROLLS 4.625% 11/25- 2/26	1	1	0	YES
	Confidential			719560	Colombia rates: 1y1y IBR payer	1	0	0	NO
	Confidential			721452	Chinese property pre- sales update 12 Nov 2021	1	1	0	NO

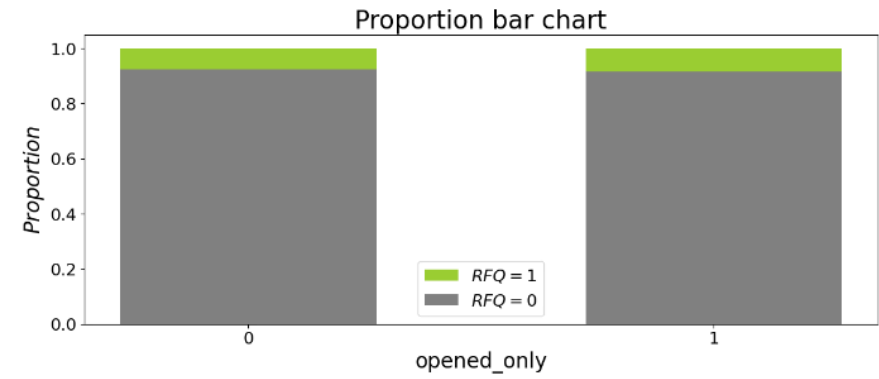
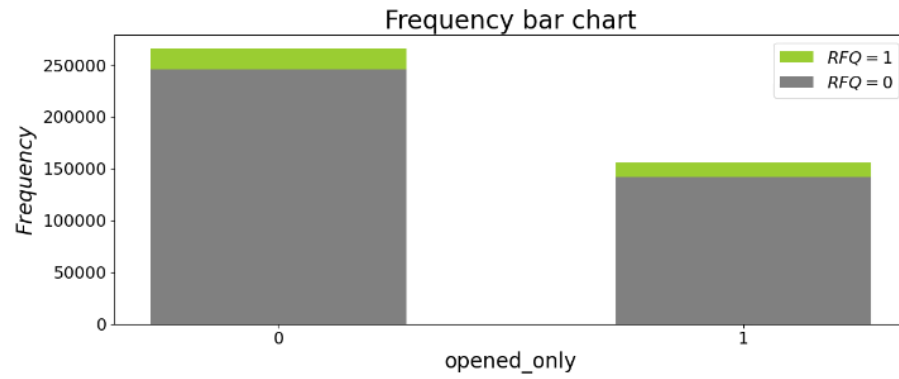
Variable - *C100_only*



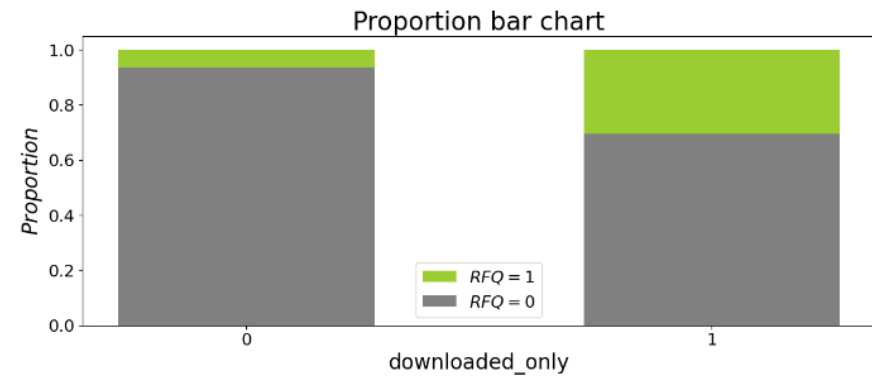
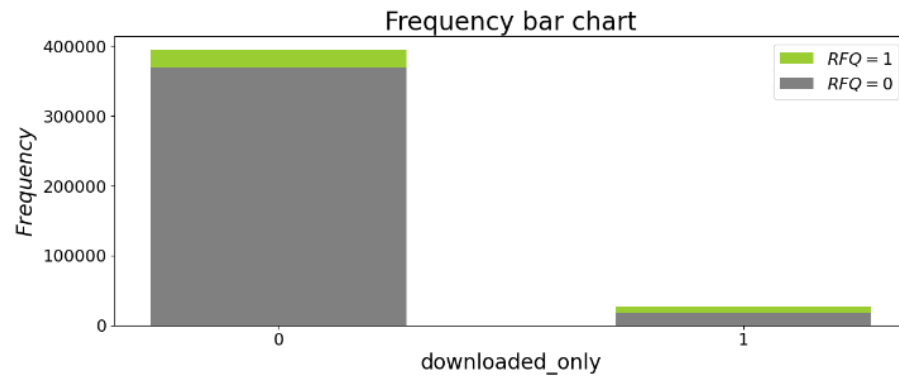
Variable - *CustomerSector*



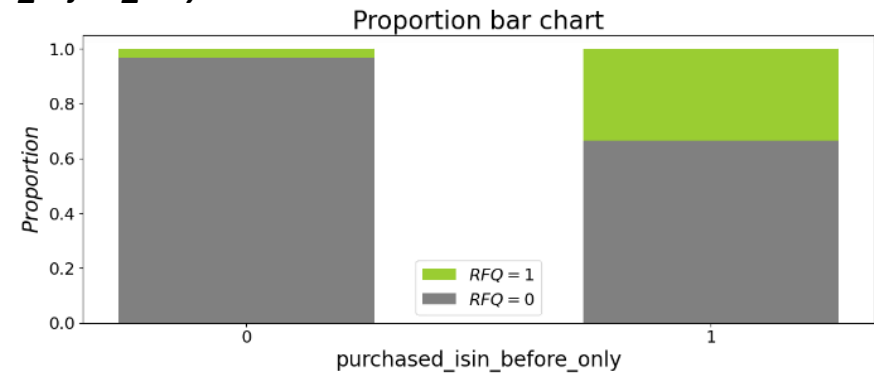
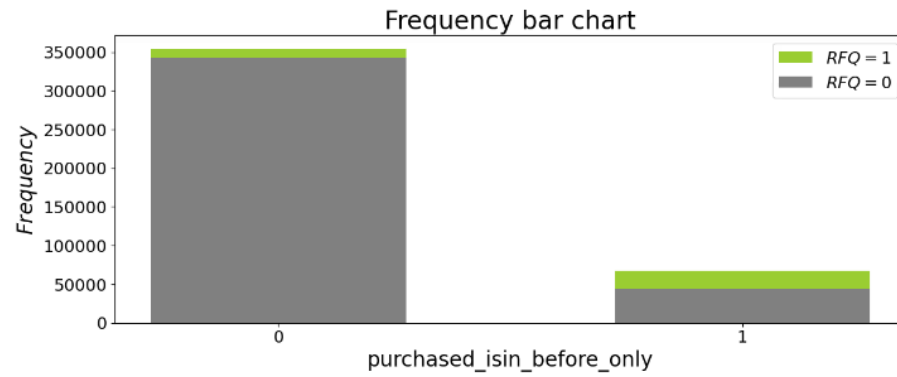
Variable - *Opened_only*



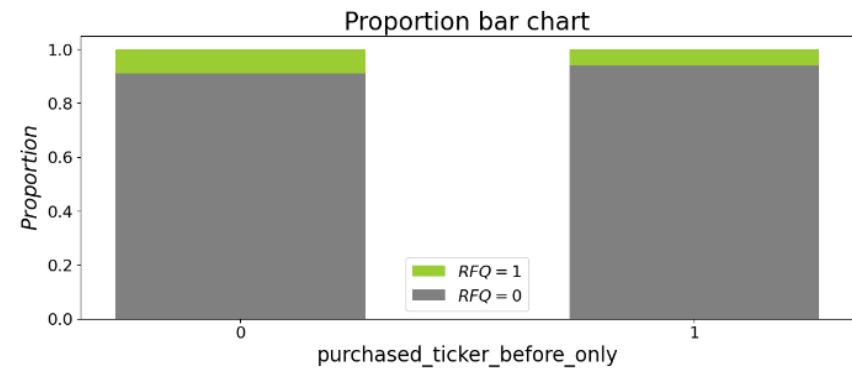
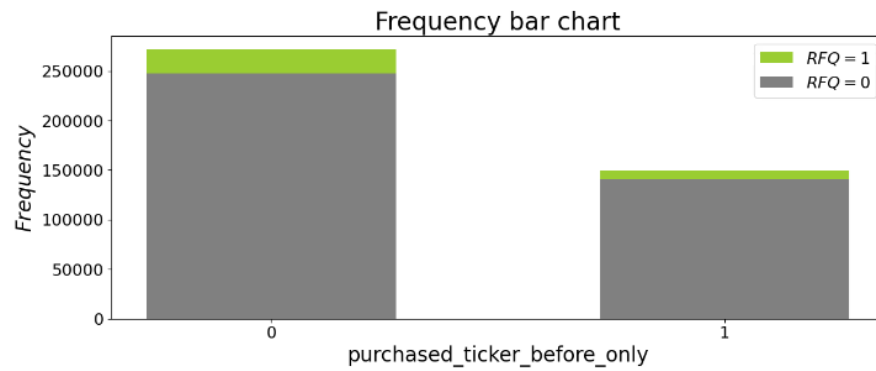
Variable - *Downloaded_only*



Variable - *Purchased_ISIN_Before_Only*



Variable - *Purchased_Ticker_Before_Only*



Variable - *Downloaded_Interaction*

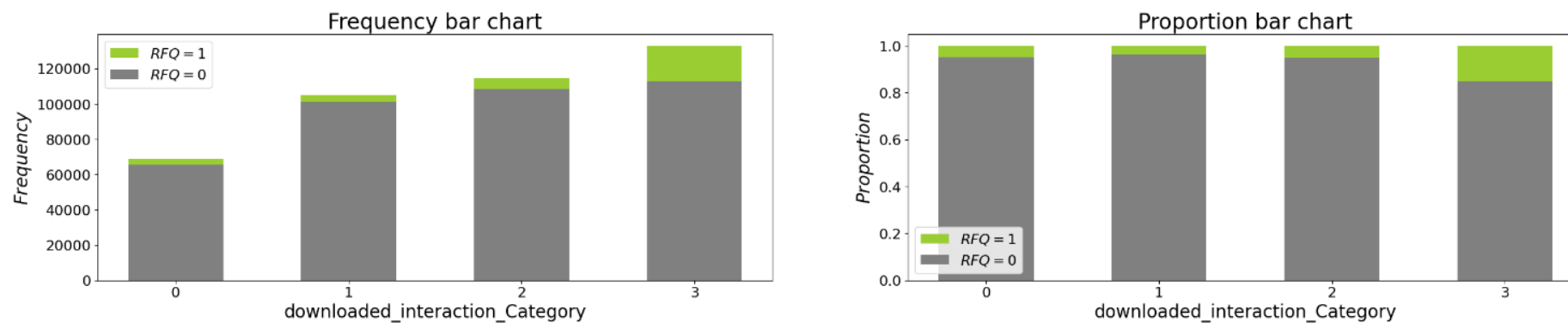
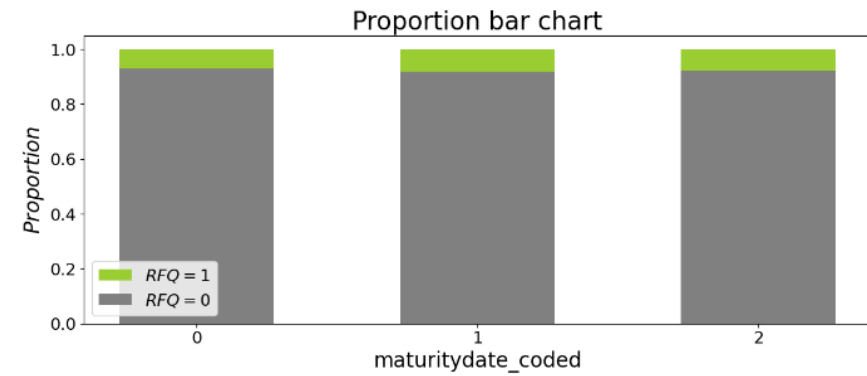
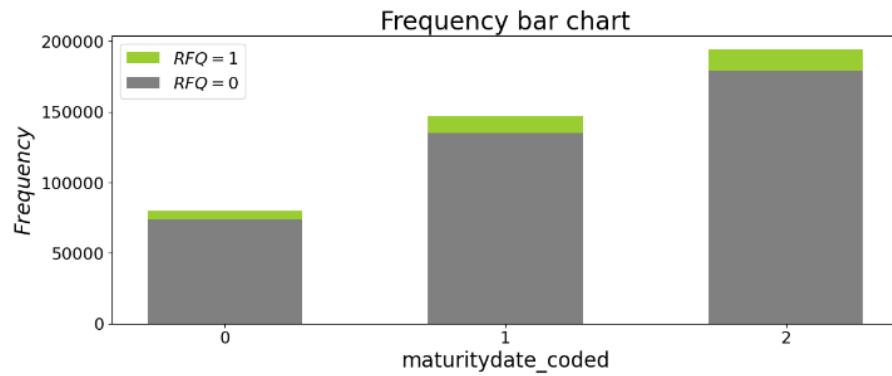
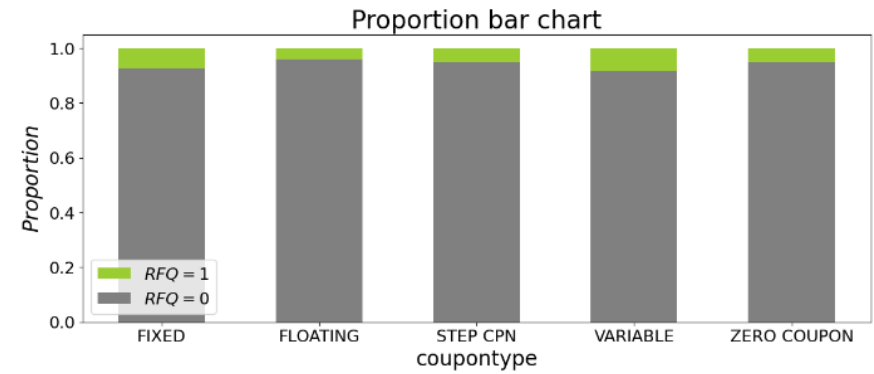
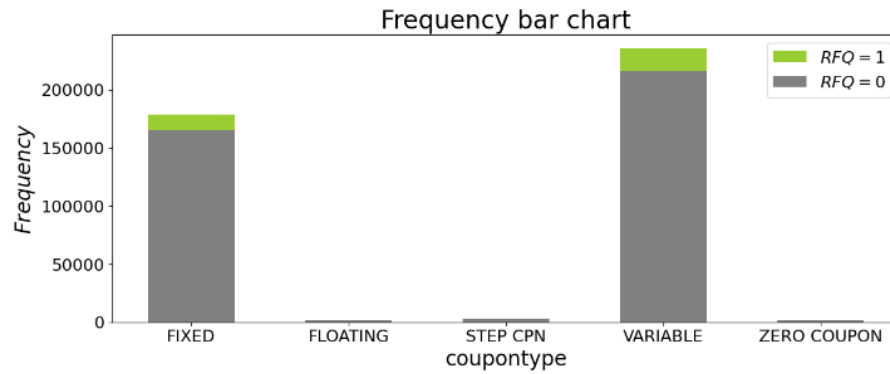


Figure A-1: Independent variables frequency and proportion bar charts against the dependent variable

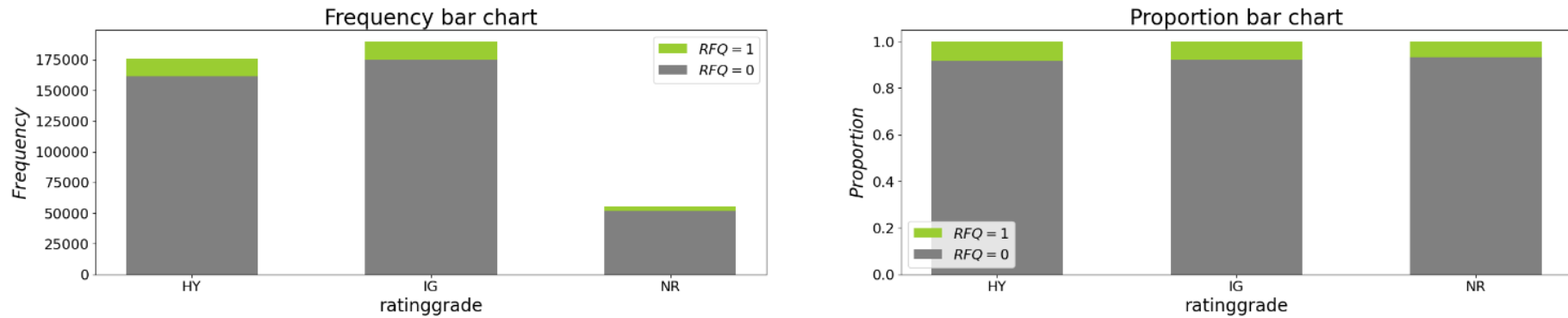
Variable - *MaturityDate*



Variable - *CouponType*



Variable - *RatingGrade*



Variable - *Direction_Indicator*

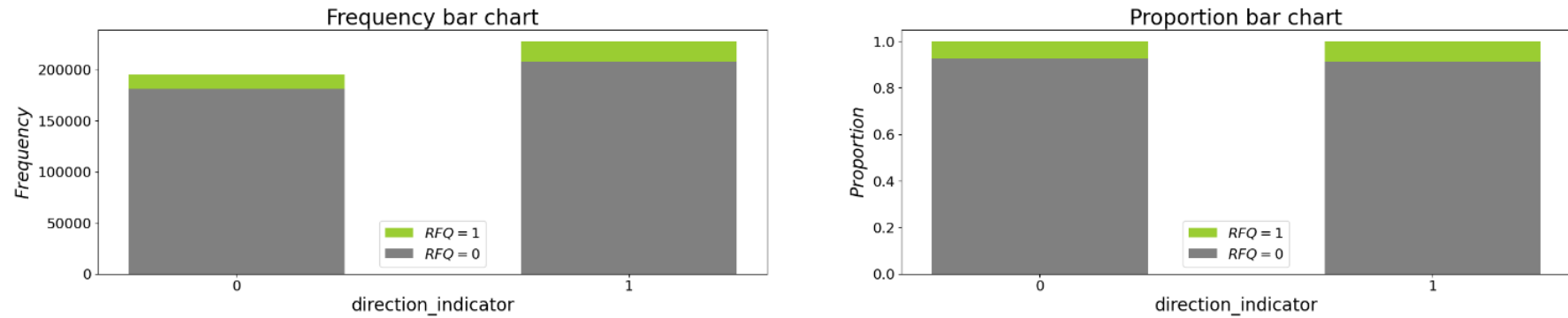


Figure A-2: Not included bond characteristics variables frequency and proportion bar charts against the dependent variable

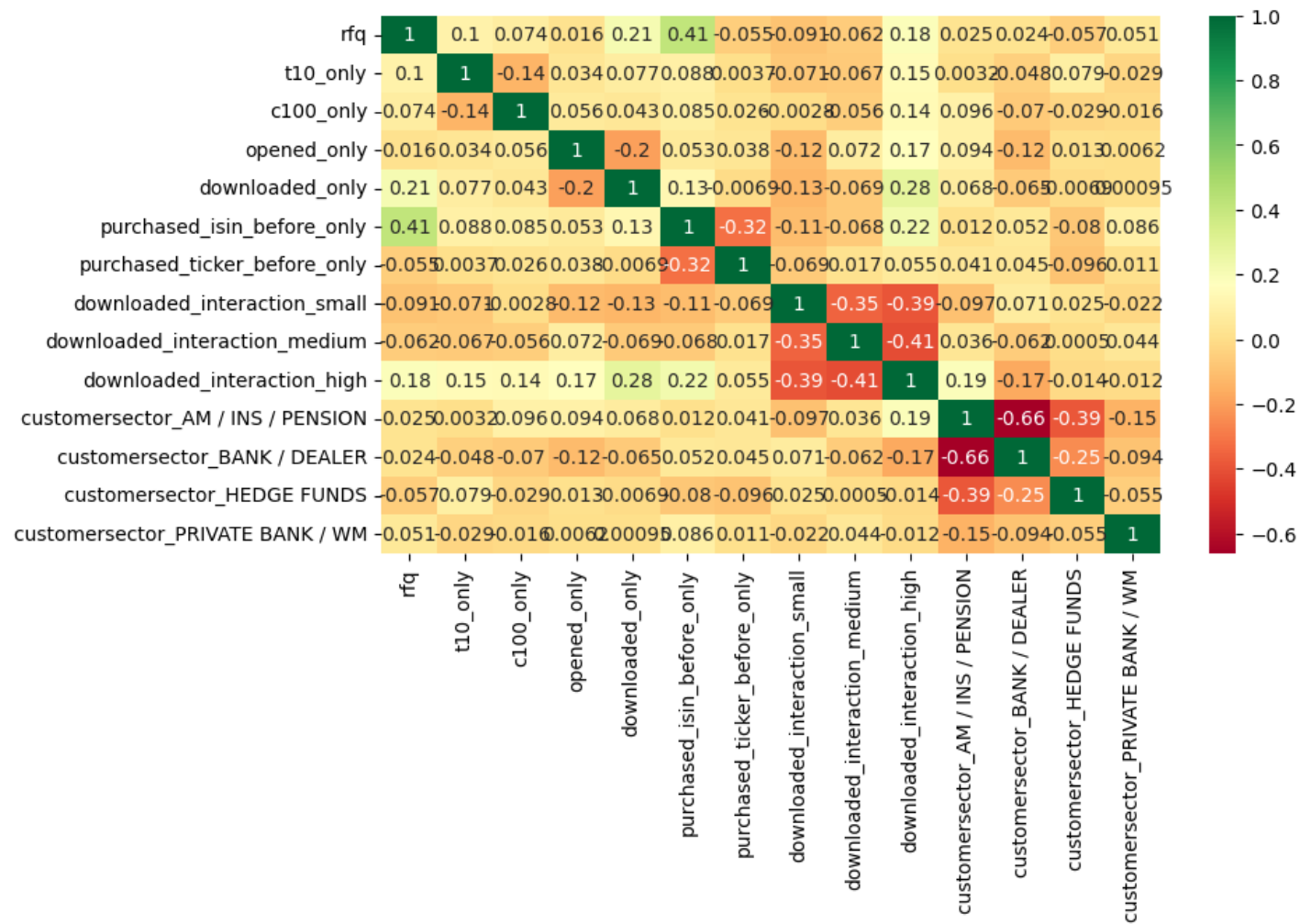
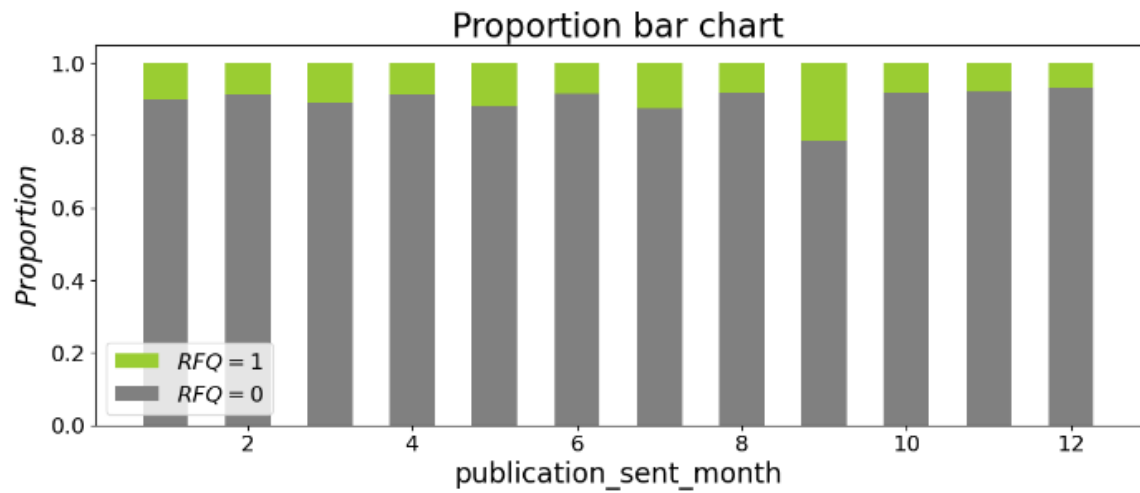


Figure A-3: Correlation Matrix

Year 2021



Year 2022

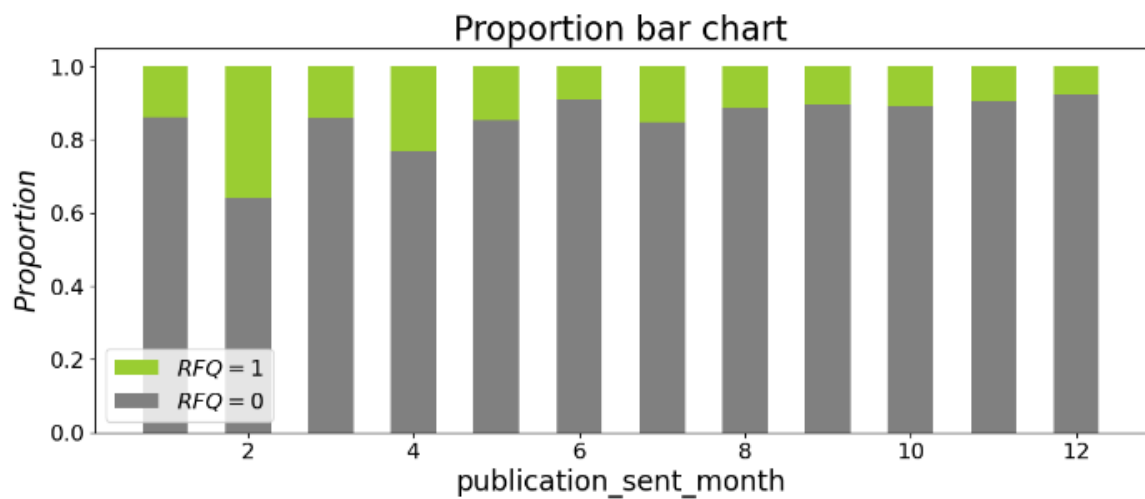


Figure A-4: Not included datetime variables proportion bar charts against the dependent variable

Dep. Variable:	rfq	No. Observations:	421218
Model:	Logit	Df Residuals:	421215
Method:	MLE	Df Model:	2
Date:	Thu, 02 Nov 2023	Pseudo R-squ.:	0.2419
Time:	15:46:17	Log-Likelihood:	-88599.
converged:	True	LL-Null:	-1.1687e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P> z 	[0.025	0.975]
const	-4.5033	0.021	-212.200	0.000	-4.545	-4.462
purchased_ticker_before_only	1.7391	0.024	72.814	0.000	1.692	1.786
purchased_isin_before_only	3.8151	0.023	167.703	0.000	3.770	3.860

Figure A-5: Logistic regression model with two IV

ANNEX B

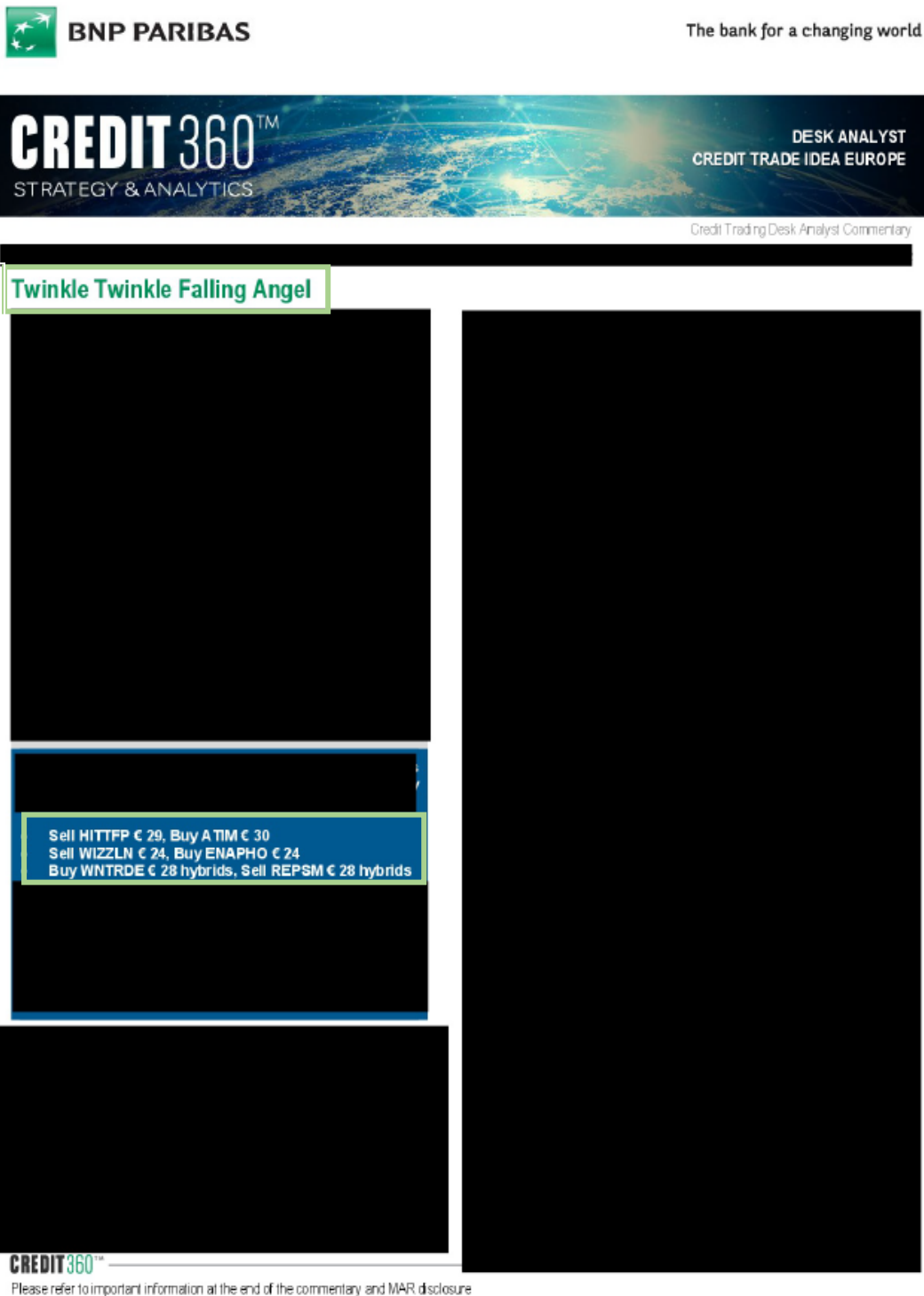


Figure B-1: Report example

Method	Average Accuracy	FNR	FPR	F-Measure	AUC
(A) Our results with V_s					
RS ^a	93.3 %	8.3 %	5.1 %	93.2 %	0.980
MB	93.9 %	8.4 %	3.8 %	93.8 %	0.983
RS-MB ^b	94.8 %	<u>7.1 %</u>	3.3 %	<u>94.7 %</u>	<u>0.990</u>
(B) Our results with V_o					
RS-MB	94.7 %	7.3 %	3.2 %	94.6 %	0.988
(C) Other results					
1) Neural network (Moro, Cortez, & Rita, 2014)	–	–	–	–	0.929
2) Artificial immune networks (Lu et al., 2016)	95.9 %	36.7 % ^c	<u>0.8 %^c</u>	73.0 % ^c	–
3) Naïve associative classifier (Villuendas-Rey et al., 2017)	78 %	–	–	–	0.76
4) Random forest with easy ensemble (Migueis, Camanho, & Borges, 2017)	–	–	–	–	0.989
5) Meta-cost-multilayer perceptron (Ghatasheh et al., 2020)	77.48 %	19.2 %	22.9 %	–	–
6) Cost sensitive classifier-multilayer perceptron (Ghatasheh et al., 2020)	84.18 %	38.6 %	12.8 %	–	–
7) Subset inference of general nonparametric Bayesian methods (Ni et al., 2020)	–	–	–	–	0.825
8) Neural network-based multiple criteria decision aiding (Guo, Zhang, Liao, Chen, & Zeng, 2021)	–	–	–	–	0.889
9) A dynamic ensemble selection method that considers the accuracy and average profit with meta-training (Feng, Yin, Wang, & Dhamotharan, 2022)	89.39 %	–	–	–	0.894
10) class membership-based classifier (Tékouabou et al., 2022)	<u>97.3 %</u>			93.2 %	0.959

Figure B-2: Prediction results comparison

Note. From “How to improve the success of bank telemarketing? Prediction and interpretability analysis based on machine learning”, by Xie, C., Zhang, J., You, Z., Xiong, B., & Wang, G. (2023). Computers & Industrial Engineering, 175, 108874 (pp. 7) <https://doi.org/10.1016/j.cie.2022.108874>

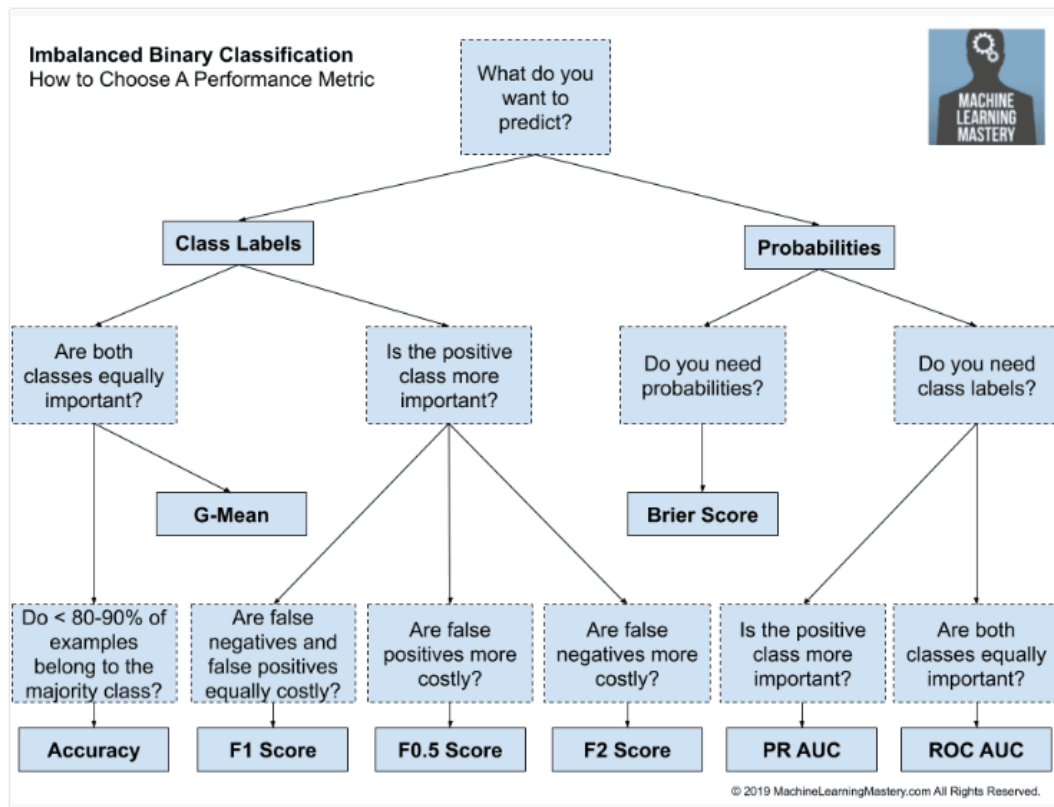


Figure B-3: Classification performance metric choice diagram

Note. From "Tour of Evaluation Metrics for Imbalanced Classification. MachineLearningMastery.com.", by Brownlee, J. (2021). <https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/>