Balancing Information Asymmetry in Loan Applications

Gianluca Bortoluzzi (240424)

# Introduction

The loan application process, as commonly understood, begins with a client going to his bank asking for an amount of money capable of solving a pressing need of his. The bank gathers profiling data on the client to understand the risk associated with the loan and develops one or multiple offers that the client either accepts or refuses. What is not immediately clear to the client, is that the bank has also invested in researching what aspects of the loan offer are more likely to result in a successful loan: the risk management is not the only factor considered when proposing an offer.

In the proposed study, we would like to offer a way to the clients in need of a loan a tool to balance the information asymmetry between client and banks. To do so, we would like to implement a prediction analysis on the BPI Challenge 2017 dataset to mimic the analysis financial institutions perform in order to increase profitability. Using random forest and gradient boosting -based prediction modes, our aim is providing clients a way to balance the information asymmetry in the form of access to knowledge about what aspects are evaluated before offering a loan.

In the next section, Background Literature, we will look at consumer rights in Europe regarding financial products, to understand in which way clients are protected by the legislation. We will argue that information asymmetry is present and harmful for credit market, but small steps can be taken to balance it. Next, the Research Methodology is explained detailing what data has been collected for the study and what features have been deemed relevant for the prediction task. What follows is the implementation of the two prediction models, their evaluation, and their respective feature importance analysis.

We conclude with a Discussion summarizing our journey, the gained insight and the next steps to further improve the project.

Code and analysis can be found at

# Background Literature

#### Consumer right to info

In Europe, there are several basic rights that guarantee consumer protection for financial products and services, such as moving funds, investing, and obtaining a mortgage loan (European Council, 2023). Those rights focus on ensuring that the client of a financial product has easy access to all information and are informed about the total credit costs. In fact, legislation cannot safeguard borrowers from the rightful interest of financial institutions to prioritize profit over the best interests of their clients. In a free market it’s perfectly normal to leverage one’s power to obtain as much as possible from a transaction, but it’s not necessarily the most ethical thing to do. We believe that the asymmetry of power between financial institutions and loan applicants, could be reduced by balancing the asymmetry in access to information.

#### Power of asymmetry of info

Asymmetric information has the stronger impacts on unobserved behaviors, which lead to adverse selection and moral hazard. This means that when establishing contractual conditions, if a company has access to more information than the other, it will be incentivized to use that information to increase its own profit at the expense of its customers. This can, and usually happens, in the form of:

* Adverse selection, when one contractual side takes a decision on the basis of incomplete or inaccurate information.
* Moral hazard, the tendency for someone to take on greater risk knowing that the other side will be bearing the cost.

A study based on a sample of the South African consumer loan market shows a significant amount of moral hazard, but weaker evidence of adverse selection, based on hidden information. The study finds it important to determine the existence and prevalence of any information asymmetries in case of public interventions in the credit market to avoid major market failures. Specifically, a large number of credit market interventions imply credit rationing, which arises from these asymmetric information problems (“Observing Unobservables: Identifying Information Asymmetries With a Consumer Credit Field Experiment,” 2009).

Asymmetric information is known in theory to have a significant impact on credit markets, resulting in inefficiently low credit availability (Jaffee & Russell, 1976). These effects are recognized in the literature regarding business-to-business relationships and should be noted that the research on the inefficient resource allocation for the customer point of view is not as comprehensive. While crucial in theory, information asymmetries can be challenging to recognize in real-world situations (“Observing Unobservables: Identifying Information Asymmetries With a Consumer Credit Field Experiment,” 2009), but there is evidence of worse advisory services quality in case of asymmetric access to information. In particular, it is the combination of buyers' lack of financial literacy and sellers' misuse of their informational advantage to further their own financial interests is the cause of financial institutions' subpar advisory services (Han & Jang, 2013).

In antithesis with our argument, we believe it is important to note that financial institutions offering credit to borrowers face uncertainty about their credit worthiness even if, in the process of lending, banks gather some proprietary information about borrowers’ creditworthiness (Dell’Ariccia, 2001; European Council, 2023).

#### Research question

The goal of this work is to take the role of a financial institution wanting to understand what aspects of an offer, in the context of a loan application, are more relevant to the client. Knowledge about what features can predict the behavior of the customer is leveraged to profit from financial activity, but it is only available to one of the sides.

We want to build several models using different statistical techniques, and subsequently run a feature importance analysis. In this research proposal, only the random forest and gradient boosting models will be implemented, but another technique will be briefly discussed and implemented in (hypothetical) final research study.

# Research Methodology

#### Data Description

The data used in the study originates from the Business Process Intelligence (BPI) Challenge, which serves as a pivotal platform in the field of process mining, facilitating the exchange of real-life datasets and fostering collaboration between researchers and practitioners. The datasets of the BPIC are important for the field of process mining in particular and has been chosen because it will allow to perform field specific analysis in the complete research.

The dataset contains all loan applications filed in 2016 and their handling up to February 2017, comprising a total of 1,202,267 events associated with 31,509 loan applications. Within this dataset, 42,995 offers were generated. The events are categorized into three types: Application state changes, Offer state changes, and Workflow events. We will focus on the Offer type of events.

To identify potential relationships between the features contained within the provided dataset, we explored the data using a correlation heatmap. For the chosen prediction task, only the offer attribute can have a meaningful impact. The following prior features were selected: *Monthly Cost, Number of Terms*, *Difference between Offered and Requested Amount*, *Loan Goal* and *First Withdrawal Amount*. Figure 1 shows the correlations between the analyzed features, and we can clearly see that they are not strongly pairwise correlated. This allows us to build a model knowing that all these features will contribute to the prediction.

**A screenshot of a graph

Description automatically generated****Monthly Cost**

Figure 1 - Features Correlation Heatmap

This feature represents the amount to be paid monthly by a client to settle the loan if the offer is selected. Given that this amount is based on *Offered Amount*, it was expected to find great variability in the data. The average monthly cost when analyzing the full dataset is close to €270. However, if we consider the interquartile range (IQR), the mean value results in €232. Lastly, we observed that this feature is related to *Offered Amount* and *Requested Amount*. Figure 2shows that most of the values for *Monthly Cost* are centered around this mean value. There is a right skew, suggesting that there are some, but not many, larger monthly costs within the data.

**Number of Terms**

*Number of Terms* refers to the number of payback terms that a client agrees to upon accepting a loan offer. This feature is most related to *Offered Amount* and, transitively, also with *Requested Amount*. In Figure 2, we can see that the range of “Number of terms” is quite wide. As some terms have a much higher occurrence, there may be some standard or default number of terms associated with these loans.

**First Withdrawal Amount**

This feature represents the first amount a client can withdraw after the process ends successfully.

A group of graphs showing different sizes and shapes

Description automatically generated with medium confidence

Figure 2 - Feature Distributions (part 1)

**Loan Goal (Encoded)**

*Loan Goal (Encoded)* represents the reason why a client takes the loan. This feature was constructed by applying label encoding to transform the original categorical feature into a numeric feature.

**Difference Offer Request**

*DiffOfferedReq* represents the difference between the requested loan amount by the client and the amount offered by the bank. This feature was constructed by simply taking the difference between these two original features.

The selected features offer a starting point for the model training process, by providing empirical evidence of relationships within the data which is required for effective model training. In the next section, we discuss the model training process.

A group of graphs showing different sizes and shapes

Description automatically generated with medium confidence

Figure 3 - Feature Distributions (part 2)

The dataset captures timestamped lifecycle information for each uniquely identifiable event, but multiple offers can be associated with a single application, meaning that our prediction task must differentiate different offers in the same application.

The focus of this analysis is on predicting clients' selection of offers made by the bank. We prioritize offer selection over acceptance due to the observed data patterns, wherein an accepted offer may not necessarily be selected immediately. Understanding client choice is crucial for inference and descriptive purposes, helping establishing data-driven strategies and identifying feature importance for further research.

#### Method Implementation

To implement the prediction models, we relied on two different methods to comparatively assess their performance. These models are both based on decision trees: Random Forest Classifier (RFC) and Gradient Boosting Classifier (GBC). In the next sections, we discuss their implementations and results.

##### Random Forest Classifier

The first prediction model implemented is a Random Forest Classifier. This model consists of a group of regression trees, which are independently trained on a randomly selected subset of features and predict the outcome through a ‘majority vote’. A distinguishing feature of Random Forest models is the deliberate inclusion of randomness throughout the construction of individual trees, as they bootstrap samples during the construction of trees, and select random subsets of features for each split. This deliberate randomness decorrelates the trees within the ensemble, and enhances the model's resistance to noise, which is part of our dataset, and improves model generalizability.

With the features selected, and described above the RFC was implemented in Python using the scikit-learn package. These features, namely *DiffOfferedReq,* *Loan Goal*, *First Withdrawal Amount*, *Monthly Cost*, and *Number of Terms*, are used as the independent variables of the model to predict the dependent variable *Offer* *Selected*.

For model training and evaluation, the data was divided into training, validation and test sets. This split ensures that the model is evaluated on new and unseen data. Using this setup, the RFC model obtained a test accuracy of *0.88*. This means that the model is able to correctly predict a client’s decision to accept or reject an offer in 88% of the cases. This is significantly higher than a naive classifier, where random classification would result in an accuracy of *0.5* for this binary classification task. This means that the model is effective in extracting information from the selected features to predict whether a client will accept or reject an offer.

Most importantly, now we have access to insights into the importance of each feature in the model. These importances are shown in Figure 4. These values are computed by the model by analyzing the ability of each feature to enhance the pureness of the leaves of the tree. Intuitively, this means that a higher value implies that a feature has a higher separation ability. In other words, distinguishing whether a loan will be accepted or rejected is more strongly dependent on that particular feature.

A bar graph with text on it

Description automatically generated

Figure 4 - Feature Importance for Random Forest Classifier model.

As Figure 4 shows, *Monthly Cost* appears to be the most important feature, followed by *First Withdrawal Amount*, *Number of Terms*, Difference between *Offered Amount and Requested Amount*, and finally *Loan Goal*. It is worth noting that none of these features are near zero, hence it is likely that they each play a non-trivial role in classifying whether an applicant will accept or reject a loan offer.

However, the monthly cost, first withdrawal amount and number of terms of a loan do seem to play a bigger role than the difference between offered amount and requested amount, and the loan goal. Using this information, the bank can for example customize loan offers in such a way that prioritizes the attractiveness of monthly cost, first withdrawal amount and number of terms for a client.

##### Gradient Boosting Classifier

Gradient Boosting is an optimization technique that adjusts its predictions by sequentially training a tree to approximate the negative residuals of the previous one. The algorithm then keeps performing gradient updates in the Gradient Descent fashion.

For the Gradient Boosting Classifier, the exact same training and evaluation setup as the Random Forest Classifier has been used. The resulting test accuracy is *0.68* for the GBC model, hence the model is able to correctly predict a client’s decision to accept or reject an offer in 68% of the cases. While this is still higher than a naive classifier (*0.5*) this accuracy is considerably lower than the RFC model (88%).

A bar graph with text on it

Description automatically generated

Figure 5 - Feature Importance for Gradient Boosting Classifier model.

Figure 5 depicts the feature importances associated with each feature by the Gradient Boosting Classifier. As this figure shows, unlike the Random Forest model, *First Withdrawal Amount* is the most important feature, followed by *Number of Terms*, *Monthly Cost*, Difference between *Offered Amount and Requested Amount*, and finally *Loan Goal*. This difference could be explained by the different ways in which each model extracts patterns from the data. Gradient boosting techniques iteratively optimize a given function while the random forest approach leverages randomness to avoid local optima. Since the Random Forest Classifier has a higher accuracy, the insights from that model seem to be more reliable.

# Discussion

Previous research has investigated the asymmetry of information in financial markets. This study employs computational learning techniques, namely the Random Forest Classifier (RFC) and Gradient Boosting Classifier (GBC), to assess loan offer elements that influence offer selection. Relevant features are determined via correlation analysis, with the RFC highlighting Monthly Cost as the most important, followed by First Withdrawal Amount and Number of Terms. GBC, on the other hand, shows lower evaluation results due to the algorithm susceptibility to local optima pitfalls.

Conversely, GBC's lower accuracy limits its reliability, potentially leading to suboptimal decision-making. However, its feature importance analysis still offers valuable insights, albeit less reliable than RFC.

Despite their differences, both models contribute to understanding consumer behavior and offer valuable insights for market participants. By using computational learning techniques, this research contributes to reducing information asymmetry in the lending market, aligning with consumer rights and market efficiency objectives.

In conclusion, while RFC emerges as the preferred model due to its superior accuracy and robustness, both RFC and GBC offer valuable insights into feature importance, empowering consumers to make informed interactions when dealing with financial institutions.

#### Next steps

The evaluation results of the RFC model, albeit performing better than GBC and Naïve (random) classifier, can be greatly improved. In the complete research project, we would like to adopt Predictive Process Monitoring techniques (Di Francescomarino & Ghidini, 2022; Maggi et al., 2014). This subfield of process mining is aimed at predicting the trajectory of a still ongoing, unfinished process. Predicting the outcome of a process, its duration, and subsequent activities can be highly beneficial in various scenarios, including production processes. This allows organizations to avoid undesirable outcomes, issues, and delays, but can be leveraged in the context of this study to highly increase the evaluation metrics and therefore the accuracy of insight about feature importance.

**References**

Dell’Ariccia, G. (2001). Asymmetric information and the structure of the banking industry. *European Economic Review*, *45*(10), 1957–1980. https://doi.org/10.1016/s0014-2921(00)00085-4

Di Francescomarino, C., & Ghidini, C. (2022). Predictive Process monitoring. In *Lecture notes in business information processing* (pp. 320–346). <https://doi.org/10.1007/978-3-031-08848-3_10>

European Council. (2023, October 31). *Consumer protection for financial products*. Consilium. <https://www.consilium.europa.eu/en/policies/consumer-protection/consumer-protection-for-financial-products>

Han, J., & Jang, W. (2013). Information asymmetry and the Financial Consumer Protection Policy. *Asian Journal of Political Science*, *21*(3), 213–223. https://doi.org/10.1080/02185377.2013.864511

Jaffee, D. M., & Russell, T. (1976). Imperfect information, uncertainty, and credit rationing. *The Quarterly Journal of Economics*, *90*(4), 651. https://doi.org/10.2307/1885327

Maggi, F. M., Di Francescomarino, C., Dumas, M., & Ghidini, C. (2014). Predictive monitoring of business processes. In *Lecture Notes in Computer Science* (pp. 457–472). https://doi.org/10.1007/978-3-319-07881-6\_31

Observing Unobservables: Identifying information asymmetries with a consumer credit field experiment. (2009). *Econometrica*, *77*(6), 1993–2008. https://doi.org/10.3982/ecta5781