PROJECT PROPOSAL

Credit Card Debt: Using analytics to drive innovation in the banking industry

Teresa Ventaja

Higher Diploma in Science in Data Analytics

School of Computing, National College of Ireland

Email: [x@x.ie](mailto:x@x.ie)

-

*Abstract*— This project aims to apply various analytics techniques to data generally used in the financial sector. Banks and credit institutions play an essential role in society by lending money to individuals, increasing cash flow and contributing to economic growth, but they need an adequate risk management strategy to minimize loses.

Keywords — Machine Learning, Default Prediction, Classification Algorithms, Random Forest, Decision Trees

# Introduction

I have personal reasons to engage in this project, since I am pursuing a career as a Data Analyst in the banking industry. Applying techniques learned in the course will help me to understand the technical challenges the industry faces, constituting an excellent complement for my professional background in e-banking. Therefore, I decided to center this project on the “Default Payments of Credit Card Clients” [1] dataset.

The fictional assumption that this is a consulting project from a client (a bank) that wants to leverage the application of data analytics to drive innovative changes will help to simulate a real business scenario. The goals of the projects are:

1. To renew the global strategy of credit card debt by uncovering risk patterns.

2. Based on the new strategy, implementing business process improvement strategies across multiple departments: HR, credit service, risk operations and community management.

# Problem Definition

To reach an agreement on the services needed, a series of elicitation methods [2] have been conducted. The requirements will translate into manageable objectives, which will guide the execution of the project.

A. **Brainstorming**: semi-formal meetings with a flexible agenda where participants can speak spontaneously but with a moderator, the Project Manager.

B. **Data Mining**: evaluate different technical approaches which could result into the desired outcome and present a proposal to the Bank’s BI Mr. Incorporate estimation of accuracy, timeframes, efficiency and costs.

C. **Estimation**: formal meeting where a preliminary invoice is presented to the bank, with a list of operational and technical resources needed. Negotiation of budgeting and deadlines.

D. **Risk Management**: 2 meetings to identify possible constraints and threats which could potentially impede the successful implementation of the project, documenting them. This information is passed to an outsourced company which creates the RM plan.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Scope | Technique | Stakeholders | Materials | Timeframe |
| Understanding the business needs | Brainstorming | Bank Credit Director, Bank BI Manager, Bank Operations Mr., Project Manager | Boardroom, recording equipment | 2 days, 4h/day |
| Diagnose which data mining methods are more suitable | Data Mining | Project Manager, Bank BI Manager | Project Proposal, Rstudio, Tableau, Microsoft Excel, SPSS | 2 days |
| Calculation of resources and budgeting | Estimation | Bank BI Mr., Bank Investments Mr., Sales Representative, Project Mr. | PowerPoint presentation on services, preliminary invoice | 2h meeting |
| Controlling constraints and creating a contingency plan | Risk Management | Outsourced company, Bank Credit Director, Bank BI Mr., Project Manager | Services risk analysis plan, quality audit plan, service delivery contract | 2 days, 2h meetings |

Figure 1: Elicitation Activity Plan

After these activities have been performed, the following requirements have been identified:

|  |
| --- |
| Uncovering credit risk patterns using 2 different machine learning techniques and compare their performance: decision trees and random forests |
| Implementing a dashboard where Credit Services Specialists in the bank have a view on the demographics of credit holders and other consumer trends |
| Building a model to predict credit limit based on the age and the repayment status |
| Identifying if credit allowance bias by genre exists, to improve corporate social responsibility rates |
| Create a Community Management plan in Twitter, starting by understanding how the topic of credit card debt is being discussed in the platform |
| Budget of 15.000 EUR for 240h of consultancy work |
| Consultant delivers quality audit plan and service delivery contract |
| Bank providing laptop |
| Consultant add a Data Security and GDPR compliance report on the technical design, including a penetration testing report and bug/incident report tracking system |
| Bank manages risk management plan other regulatory requirements by its own means |

# Current Status of Art

* **Statistical analysis**: descriptive statistics will help choosing the appropriate design and data mining techniques. Hypothesis testing is usually underrated, but in this case, a t-test suffices to understand if genre bias exists [3]. Since the explanatory power of this analysis is limited, demographics will be explored with graphical methods, such as visualizations with Tableau.
* **Regression model**: multiple linear regression to predict an outcome from two independent variables. For that, we need variables that are uncorrelated to each other but are both related to the predicted variable [4].
* **Data Mining & Classification Algorithms**: Decision Trees and Random Forests (CART analysis) involve segmenting associations into a set of splitting rules, making their interpretation simple and useful [5]. Therefore, they will be used in this project as the main resource for discovering credit risk patterns.
* **Text Mining**: it consists in analyzing text to extract information that is useful for specific purposes. That information needs to be actionable, that is, capable of providing a basis for actions to be taken automatically, and comprehensive (the information helps to explain the data) [6]. Specifically, web scrapping and topic modelling will be used.

# Engineering Approach

This project is based on the CRISP-DM methodology due to its strong focus in the business and data understanding, and the flexibility that its cyclical feedback system involves [7]. It consists on the following phases:

Diagram

Description automatically generated

Figure 2: CRISP-DM Phases [8]

In terms of deliverables, 2 proposals can be of interest. The first one is a cost-effective solution covering the main business objectives, and the second amplifies the scope, but it is more costly.

**Solution 1**: hypothesis testing to check if there is genre bias. Creating a dashboard with demographic visualizations. Creating a regression model to predict credit limit based on age and repayment status. Performing Decision Trees and Random Forests models and describing credit risk patterns based on the best performing model.

**Solution 2**: Same as Solution 1, adding Twitter text scrapping plus topic modelling to get a sense of the discussions on the topics “Credit card debt” and/or “Consumer debt”.

I will put in practice data visualization principles and statistical techniques, such as descriptive, hypothesis testing and multiple linear regression. I selected to use CART and topic modeling to challenge myself to do something that I have never done before, because in previous projects I used other techniques.

# Tasks & Deliverables

|  |  |  |
| --- | --- | --- |
| Section | Tasks | Deliverables |
| Credit Card Debt - Using analytics to drive innovation in the banking industry | Problem Statement | Proposal - service offer |
|  | Requirements Elicitation | Auditing report |
| Objectives | Aim of the Research | Service delivery contract |
|  | Analysis Topics |  |
| Related work | Background Information on Consumer Debt and Credit Card Debt | Report |
|  | Literature review on Machine Learning applied to financial analysis | Prototype and proposal, including costs & resources |
| Methodology | Research Design |  |
|  | Data description & preparation |  |
|  | Statistical Analysis | Report on main findings |
| Trend Analysis | Dashboard with demographics & trends | Dashboard |
|  | Credit Allowance - genre bias | Dashboard |
| Modelling | Multiple linear regresion - prediction of credit limit | Report |
|  | Decision trees & random forests, testing and evaluation, confusion matrix | Manual with schematics |
|  | Twitter text scrapping, topic modeling and evaluation | Report |
| Conclusions | Conclusions |  |
|  | Future work | Prototype - proposal of further analysis |

Figure 3: Tasks & Deliverables

# Project Management

The execution of the project will follow this plan:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Activity** | **Description** | **Ownership** | **Start Date** | **End Date** | **Budget** | **Resources** |
| Problem definition | Brainstorming | Bank BI Manager | 04/10/20 | 06/10/20 | N/A | Zoom |
| Sourcing data | Getting datasets and materials | Project Manager, Bank Credit Director | 07/10/20 | 18/10/20 | 2,000 € | Bank provides laptop i7, 16GB RAM memory |
| Proposal report | Service delivery contract for signature | Project Manager | 04/10/20 | 18/10/20 | 15,000 € | Microsoft Suite |
| Risk Management Plan | Drafting the plan | Outsourced | 19/10/20 | 25/10/20 | 1,000 € | N/A |
| Quality Audit Plan | Drafting the plan | Outsourced | 26/10/20 | 30/10/20 | 1,000 € | N/A |
| Consulting banking expert | Virtual meeting | Project Manager | 07/10/20 | 07/10/20 | N/A | Zoom |
| Exploratory analysis of dataset | Statistical analysis: descriptive, normality and t-test (genre bias) | Project Manager | 19/10/20 | 22/10/20 | N/A | Excel, SPSS |
| Data Modelling | Decision trees and random forest | Project Manager | 23/10/20 | 29/10/20 | N/A | Rstudio |
| Twitter Scrapping | Collecting Tweets from the social networking platform | Project Manager | 30/10/20 | 01/11/20 | N/A | Rstudio |
| Preparing project report | Based on the structure designed on "Tasks and Deliverables" section | Project Manager | 02/11/20 | 06/12/20 | N/A | Microsoft Suite |
| Dashboard | Analyse trends and present them in a structured way | Project Manager | 03/11/20 | 09/11/20 | N/A | Tableau |
| Regression model | Prediction of credit limit | Project Manager | 10/11/20 | 15/11/20 | N/A | SPSS |
| Analyse classification model | Add to report | Project Manager | 16/11/20 | 22/11/20 | N/A | Rstudio |
| Topic modeling analysis | Add to report | Project Manager | 23/11/20 | 29/11/20 | N/A | Rstudio |
| Present Project Report | Deliver to client | Project Manager | 10/12/20 | 10/12/20 | N/A | N/A |
| Bank implementation | Community Management plan, Data Security and GDPR compliance, designing new innovation strategy | Bank's Board of Directors | 12/12/20 | 31/12/20 | On bank's discretion | On bank's discretion |

Figure 4: Project Management Plan

Chart, bar chart

Description automatically generated

Figure 5: Gantt Chart

Every aspect of the project design has been thought based on practicing techniques that I have never worked with before. Over 80% of the project objectives are challenging for me and will be useful for my career. In that sense, I declare that I will write the project on my own, not copying from other sources.

# Conclusions

The combination of multiple disciplines, such as statistics, computer programming, data visualization and business analytics is what defines the robustness of this work, while a single technique on its own would limit its explanatory power. This project proves that realistically, in roughly three months, with a relatively low budget (20000€-25000€) it is possible to leverage analytics for business process improvement, optimizing efficiency. Ultimately, this will result in a significant reduction of monetary loses.

##### References

1. Yeh, I-C. & Lien, C. H., “Default of credit card clients Data Set”, *UCI Machine Learning Repository”* 2009. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[Accessed on: Oct. 17, 2020].

1. IIBA, “Chapter 4: Elicitation and Collaboration”, *BABOK 3 ONLINE - A Guide to the Business Analysis Body of Knowledge*, p. 56, 2015 [Online] Available: <http://babokonline.org/> [Accessed on: Oct. 17, 2020].
2. Salkind, N. J., “t(ea) for Two (Again) Tests Between the Means of Related Groups”, *Statistics for People Who (Think They) Hate Statistics*, p. 288, 2017, 6th Edition.
3. Salkind, N. J., “The More Predictors the Better? Maybe”, *Statistics for People Who (Think They) Hate Statistics*, p. 288, 2017, 6th Edition.
4. James, G, Witten, D., Hastie, T. & Tibshirani, R., “Lab: Decision Trees”, *An Introduction to Statistical Learning with Applications in R*, p. 323, Springer: 2017
5. Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J., “Text Mining”, *Practical Machine Learning Tools and Techniques*, p. 515, MK: 2017, 4th Edition
6. Azevedo, A. & Santos, M. F., “KDD, semma and CRISP-DM: A parallel overview”, *IADIS European Conference on Data Mining 2008*, p. 182, July 24-26, 2008, Available: https://www.researchgate.net/publication/220969845\_KDD\_semma\_and\_CRISP-DM\_A\_parallel\_overview [Accessed on: Oct. 17, 2020].
7. Vorhies, W., “CRISP-DM – a Standard Methodology to Ensure a Good Outcome”, Data Science Central, July 26, 2016 [Online] Available:  [https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome/](http://babokonline.org/) [Accessed on: Oct. 17, 2020].
8. Yeh, I. C., & Lien, C. H., “*The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications*”, Department of Information Management, Chung Hua University, Taiwan. Department of Civil Engineering, Tamkang University, Taiwan. 36(2), 2009, 2473-2480.
9. Payments Services Statistics, *Central Bank of Ireland*, 15 October 2019 [Online]. Available: <https://www.centralbank.ie/statistics/data-and-analysis/payments-services-statistics> [Accessed on: Dec. 1, 2020]
10. Y. Zhao, Y. Zhao & I. Song, “Predicting New Customers' Risk Type in the Credit Card Market”, *Journal of Marketing Research*, vol. 46, pages 506-517, 2009 [Online]. Available: <http://www.bm.ust.hk/mark/files/staff/Ying/Ying-JMR-2009.pdf> [Accessed on: Dec. 1, 2020]
11. T. M. Rodica, “The credit risk - Component of the baking risks”, *University of Oradea, Faculty of Economics*, vol. 1(2), pages 430-437, December 2011 [Online]. Available: Annals of Faculty of Economics, <https://ideas.repec.org/s/ora/journl.html> [Accessed on: Dec. 1, 2020]
12. I-C. Yeh & C. H. Lien, “Default of credit card clients Data Set”, *UCI Machine Learning Repository”* 2009. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[Accessed on: Dec. 01, 2020].

1. S. Khemakhem & Y. Boujelbene, “Predicting credit risk on the basis of financial and non-financial variables and data mining”, *Review of Accounting and Finance*, vol. 17, August 2018. [Online]. Available: <https://www.researchgate.net/publication/327119825_Predicting_credit_risk_on_the_basis_of_financial_and_non-financial_variables_and_data_mining> [Accessed on: Dec. 01, 2020].
2. F. Butaru, Q. Chen, B. Clark, S. Das, A. W. Lo & A. Siddique, “Risk and risk management in the credit card industry”, *Journal of Banking & Finance*, vol 72, pages 218-239, 2016 [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378426616301340> [Accessed on: Dec. 01, 2020].
3. C. Zhang, C. Liu, X. Zhang & G. Almpanidis, “An up-to-date comparison of state-of-the-art classification algorithms”, *Expert Syst.* Appl. 82 pages 128–150, April 2017 [Online]. Available: <http://dx.doi.org/10.1016/j.eswa.2017.04.003> [Accessed on: Dec. 01, 2020].
4. J. Zhou, W. Li, J. Wang, S. Ding & C. Xia, “Default prediction in P2P lending from high-dimensional data based on machine learning”, Physica A: Statistical Mechanics and its Applications, vol. 534, 2019 [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378437119313652> [Accessed on: Dec. 01, 2020].
5. L. W. Woosley & M. B. Sherman, “Big data, machine learning and bias”, *The RMA Journal,* vol. 101, Iss. 10, pages 26 – 31. Philadelphia, Jul/Aug 2019 [Online]. Available: <https://rmajournal.org/rmajournal/july_august_2019/MobilePagedArticle.action?articleId=1504204> [Accessed on: Dec. 01, 2020].
6. N. J. Salkind, “t(ea) for Two (Again) Tests Between the Means of Related Groups”, *Statistics for People Who (Think They) Hate Statistics*, page 288, 2017, 6th Edition.
7. N. J. Salkind, “The More Predictors the Better? Maybe”, *Statistics for People Who (Think They) Hate Statistics*, page 376, 2017, 6th Edition.
8. G. James, D. Witten, T. Hastie & R. Tibshirani, “Lab: Decision Trees”, *An Introduction to Statistical Learning with Applications in R*, Springer page 323, 2017
9. M. Schonlau & R. Y. Zou, “The random forest algorithm for statistical learning”, *The Stata Journal*, vol, 11, Iss. 1, pages 3 – 29, March 2020 [Online]. Available: <https://doi.org/10.1177/1536867X20909688> [Accessed on: Dec. 01, 2020].
10. C. Dhaoui, C. Webster & L. Tan, “Social media sentiment analysis: lexicon versus machine learning”, *Journal of Consumer Marketing*, vol. 34, August 2017 [Online]. Available: <https://www.emerald.com/insight/content/doi/10.1108/JCM-03-2017-2141/full/html> [Accessed on: Dec. 01, 2020].
11. IIBA, “Chapter 11: The Agile Perspective”, *BABOK 3 ONLINE - A Guide to the Business Analysis Body of Knowledge*, page 374, 2015 [Online] Available: <http://babokonline.org/> [Accessed on: Dec. 01, 2020].
12. A. Azevedo & M. F. Santos, “KDD, semma and CRISP-DM: A parallel overview”, *IADIS European Conference on Data Mining 2008*, p. 182, July 24-26, 2008, Available: https://www.researchgate.net/publication/220969845\_KDD\_semma\_and\_CRISP-DM\_A\_parallel\_overview [Accessed on: Dec. 01, 2020].
13. Vorhies, W., “CRISP-DM – a Standard Methodology to Ensure a Good Outcome”, Data Science Central, July 26, 2016 [Online] Available:  [https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome/](http://babokonline.org/) [Accessed on: Dec. 01, 2020].