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

1st Teresa Ventaja, 2nd E. S.

Higher Diploma in Science in Data Analytics

School of Computing, National College of Ireland

Email: 1st x@x.com, 2nd e.s@x.ie

*Abstract*— Credit risk management plays an important role in preventing losses of financial institutions, but it also has significant consequences on the global economy. Not limited to risk prevention, this research presents a wide scope of innovative operational strategies involving multiple stakeholders, aiming to demonstrate how data analytics can be helpful on business process improvement in the banking industry. A dataset containing assorted real financial information and demographics has been used as a fictional IT consultancy project.

First, an exploration on previous research on the topic has provided a theoretical framework to work with. As per the analysis, a Mann-Whitney U test determined that there is a genre bias on credit allowance, and demographical trends about credit allowance and risk of default were further explored in a graphical way. To reach a finer level of detail on credit allowance, a prediction estimation tool was created using a multiple linear regression.

Turning to default prediction and prevention, five algorithms were trained to compare and select the one providing the greatest level of accuracy: Decision Trees (C5.0, rpart, train from caret package and bagging) and Random Forest (randomForest). Kappa and accuracy were displayed in confusion matrices to compare results, and 30% of the data was used to test the model. Random Forest performed best, with 99.32% accuracy and 0.98 Kappa. These CART models were also used to generate training materials for Risk Prevention agents. To conclude, 979 tweets were used for sentiment analysis and topic modeling on the theme of “Credit Score”, to get an understanding of what the public is discussing about and create an effective marketing campaign. Suggestions were given as to what words will perform better on advertising, and what topics to focus on.

Keywords—Machine Learning, Default Prediction, Risk Management, Credit Prediction, Multiple Linear Regression, Decision Tree, Random Forest, Sentiment Analysis, Topic Modeling.

# INTRODUCTION

## Problem Statement

This project is based on the “Default Payments of Credit Card Clients” [1] dataset. It aims to simulate a real business scenario based on the fictional assumption that it is a consulting project from a client (a bank) that wants to leverage the application of data analytics to drive innovative changes. The goals of the projects are:

1. To renew the global strategy of credit card (CC) 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.

The design has been thought based on practicing new techniques (over 80% are used here for the first time) and for gaining exposure to data analytics in the banking industry, in alignment with my career objectives. The original scope of the project has been amplified, adding sentiment analysis, as it has been identified that it does not makes a big difference in terms of time resources, but it adds valuable insights worth being included.

## Operational Objectives

This analysis aims to provide a realistic simulation, assuming that it is data about clients from different global locations, collected in 2019, the currency being US dollars. The bank is investing 15,000 USD for 240h of IT consultancy work. It will result on a comprehensive report, and the implementation of the technology into the back-office system will have an extra cost of 2,000 USD. The bank will manage by its own means the regulatory and risk management plan, as well as any other non-technical requirements.

The requirements resulting from elicitation activities have been included in [Appendix 1](#_Appendix_1:_Requirements), which translate into the following operational targets (grouped by business division):

1. **Management Board**: enhancing the global strategy of credit risk prevention and implementing innovative changes in multiple departments (HR, credit services, risk operations and marketing)
   1. Defining default patterns using CART algorithms: decision trees/random forests.
   2. Analyzing demographic trends on CC default and credit allowance
2. **Risk Prevention**: creating a model that will be used to predict CC default.
   1. Comparing the performance of different CART algorithms and selecting the best
   2. Creating a flowchart about credit default prediction, used to train agents on detecting default patterns.
3. **Credit Services (CS):** creating tools capable to produce reports with metrics on CC default/credit allowance and track performance.
   1. Building a dashboard for CS agents, including trends on CC default.
4. **Human Resources**: improving onboarding materials for new employees so genre biases are not reflected in their business decisions.
   1. Identify if credit allowance genre bias exists using a statistical test
5. **Marketing**: enhancing the Community Management credit campaign
   1. Understanding how credit is discussed in Twitter by performing sentiment analysis and topic modelling

# RELATED WORK

## Risk Management on Credit Card Debt

The Covid-19 pandemic is boosting the transition into a cashless society, in a world where card transactions accounted for 67% of the total number of payments in Ireland in 2018 [2]. Consumers use credit when they believe that they may need fungibility in the future, when credit might be unavailable, and they use it even if interest rates makes it costly [3]. Banks’ credit card profit is mainly coming from interchange fees, but cardholders known as revolvers play an important role. They usually carry a balance on their card, borrow at high interest rates, but will eventually repay their loans. The interest on the amount that is rolled over to the subsequent periods accounts for 70% of card revenue, making revolvers bank’s best customers [4].

With over 80% of banks’ balance related with risk management, credit risk is the main cause of banks insolvency. In particular, solvency risk is the breach of credit contract clauses by cardholders, which can be recovered partially or delayed, or lead to a definitive loss of capital [5]. Lenders’ practices such as “run-up”, i.e., cutting or freezing credit on accounts that seem to go into default, could avoid an increase on the due balance. However, they might also cut the credit of non-defaulted accounts, loosing profitable lending opportunities. Accurately forecasting the risk of default will decrease false positives, allowing cutting credit lines only on the bad accounts and saving costs [6].

On the top of the benefits of reducing losses in banks, risk prevention has an impact on the global economy. The global financial crisis in 2008 was mainly due to unrestricted growth of credit, limited credit quality and deficient credit risk management. Consequently, a new international regulatory framework called “Basel III” was created, aiming to incentivize the use of sophisticated models for calculating credit risk. Accurately assessing credit risk, financial institutions contribute to avoiding the devastating effects of recessions [7].

Global losses from payment fraud has tripled from 2011 to 2020 and are expected to continue increasing by 25% in 2027, representing a cost of $40.62 billion [8]. In CC debt, defaulting is failing to make a payment on a debt by the due date. Traditional assessments of credit risk included sociodemographic and credit application information but including transaction behavioral data improves the predictions. For example, the repayment capacity is a signal of the ability to make debt payments in the future, while the rate of bad debt is the main measure of credit risk [7].

Advancements in technology present an invaluable opportunity to use innovative methods in risk prevention, such as classification algorithms. C. Zhang et al. [9] compared 11 algorithms, and they concluded that decision trees and random forests were ranked in the top 5 [10]. Testing algorithms, ensuring the business need for using the data or the algorithm, and monitoring the performance of the model over time are among the best practices for successful implementation. On the other hand, only using automated resources to make lending decisions might have negative ethical consequences. To reduce them, regulators have an interest in introducing human discretion, documenting the reasoning for such decisions [11].

## Machine Learning & Analytics applied to financial analysis

This section covers the definition of some key technical and analytical methods used in the field of finance. Machine learning is the process by which complex computers create analytical models used to make decisions. The computer determines which data elements to use in predicting whether to allow or deny credit. As the system ingests more data, the model update itself, selecting which variables are important [11].

* **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 [12]. Multiple linear regressions are useful to predict an outcome from two independent variables. An estimation of Credit Limit will be calculated with this method [13]. Since the explanatory power of this analysis is limited, demographics will be explored with graphical methods, visualizations with Tableau.
* **Data Mining & Classification Algorithms**: CART algorithms such as Decision Trees (DT) and Random Forests (RF) involve segmenting associations into a set of splitting rules, making their interpretation simple and useful [14]. They will be used for discovering credit risk patterns. RF reduce overfitting of DT by using a subset of the observations to build many individual trees. The algorithm averages predictions over many individual trees, built on bootstrap samples rather than the original sample. DT are more easily interpretable than RF, but they tend to predict better [15].
* **Text Mining**: there are over 500 million new tweets each day. Using automated methods, such as a lexicon of weighted words, customers’ feelings and opinions are classified into positive and negative. Sentiment analysis is challenging when applied to unstructured text, because of the informality and instantaneity of communication, free-flowing text, casual in word and grammar usage, typically including abbreviations, misspellings, emoticons, emojis, use of SMS-like syntax and the 140-character limitation [16]. For a more comprehensive analysis, this project is including topic modelling.

# METHODOLOGY

## Research Design

The development of this analysis is based on the Agile Scrum technique[[1]](#footnote-1) [17]. CRISP-DM is the selected methodology due to its strong focus in the business and data understanding, and the flexibility that its cyclical feedback system involves [18]. It consists on the following phases:

Diagram

Description automatically generated

Figure 1: CRISP-DM Phases [19]

“Business Understanding” was covered in [Sections I](#_INTRODUCTION) & [2](#_RELATED_WORK), and [this section](#_METHODOLOGY) covers “Data Understanding” and “Data Preparation”. They will be separated into different sub-sections, because this project uses different models and analysis methods, and so the data pre-processing needs are. [Sections IV](#_TREND_ANALYSIS) and [V](#_MODELING) will cover “Modeling” and “Evaluation”.

The following technological resources have been selected to perform the analysis:



Figure 2. Technological Tools

In [Section IV.2](#_Credit_Allowance_-)., evaluation will be measured by interpreting the p-value on a 95% confidence level, with an effect size of 0.2. In [Section V.1](#_Multiple_Linear_Regresion)., the R value will determine the predictive power of the model. Finally, in [Section V.2](#_Decision_Trees_&)., precision and kappa will evaluate the performance. Kappa statistic measures performance relative to random classification, and above 0.6 represents substantial performance. 30% of the dataset will be used for testing. Assessing the error rate of this data, that did not participate on training the classifier, will test the performance on new data [20].

This research uses anonymized data, complying with GDPR and privacy by design and by default rules. Since information privacy also depends on users’ and employees’ behavior, to avoid revealing personal data, it is recommended that appropriate training is given [21].

## Data Description & Preparation

The original dataset contains 30,000 observations and 24 variables[[2]](#footnote-2). No missing values were found, but some categories not containing meaningful information have been transformed for simplification[[3]](#footnote-3):





Figure 3. Dataset Description

[**Section IV.1.**](#_Dashboard_with_Demographics)

Multiple numeric fields have been converted into categorical[[4]](#footnote-4) for an easier manipulation of data. The dependent variable is “Pay\_Def” and the independent variables are all others used.

[**Section IV.2.**](#_Credit_Allowance_-)

The dataset has 18,112 observations for women and 11,888 for men. This test requires samples of equal size. The sample size was calculated with “G\*Power” software. For an effect size of 0.2, capable of detecting small differences, and power of 0.95, the result indicates a size of 682 for each genre[[5]](#footnote-5).

## Statistical Analysis

[**Section IV.2.**](#_Credit_Allowance_-)

The dependent variable is “Amount of Given Credit” (LIMIT\_BAL) and the independent variable is “Sex”. Selecting the appropriate statistical test requires testing normality. Histograms have been displayed in SPSS, being negatively skewed and platykurtic, and Q-Q Plots suggest a significant deviation from normality:

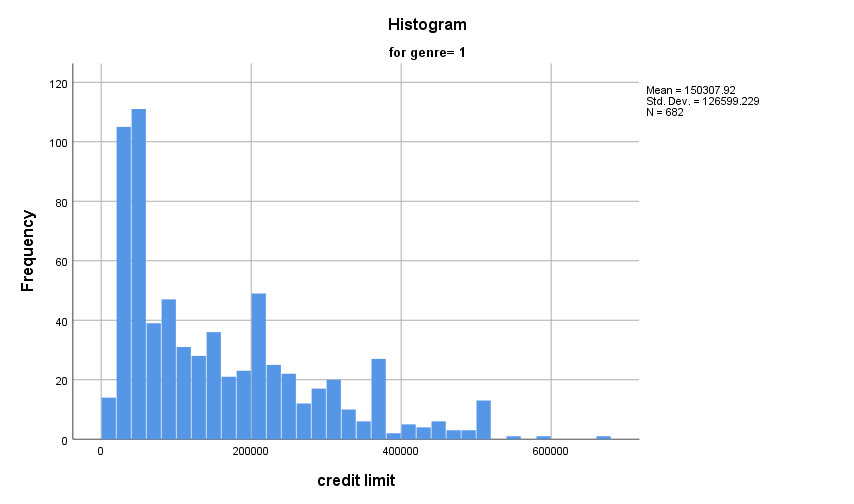


Figure 4. Histogram (Male)

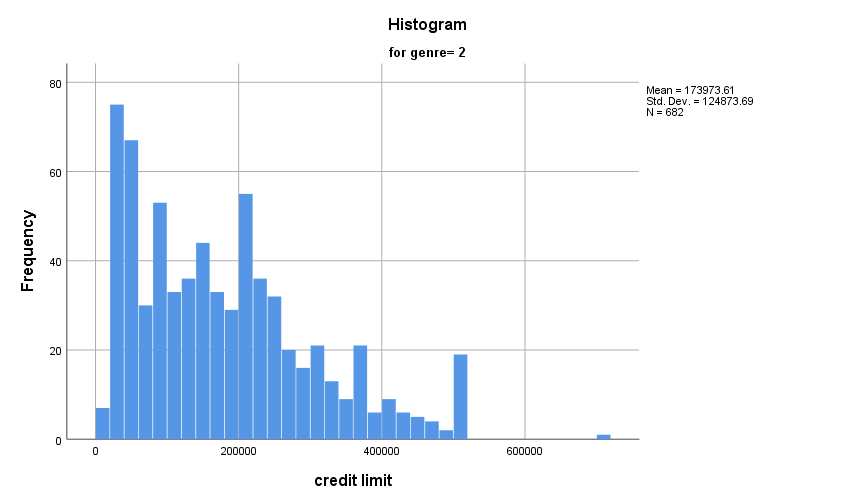


Figure 5. Histogram (Female)

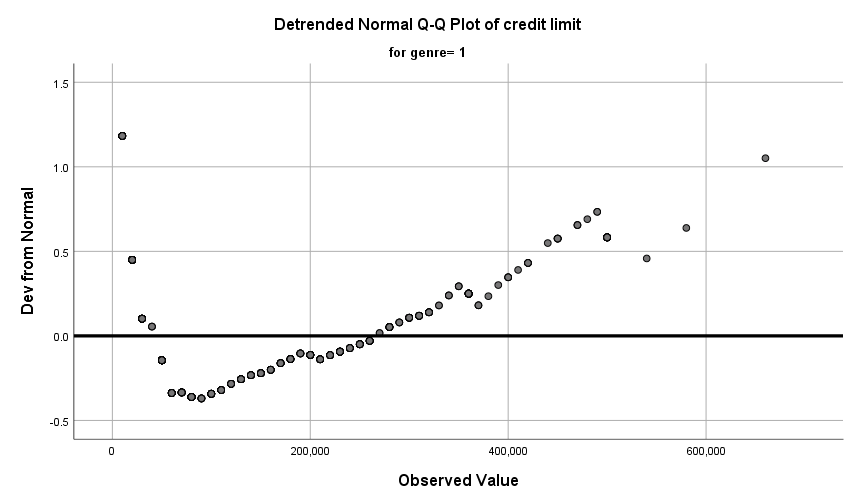


Figure 6. Detrended Q-Q Plot (Male)

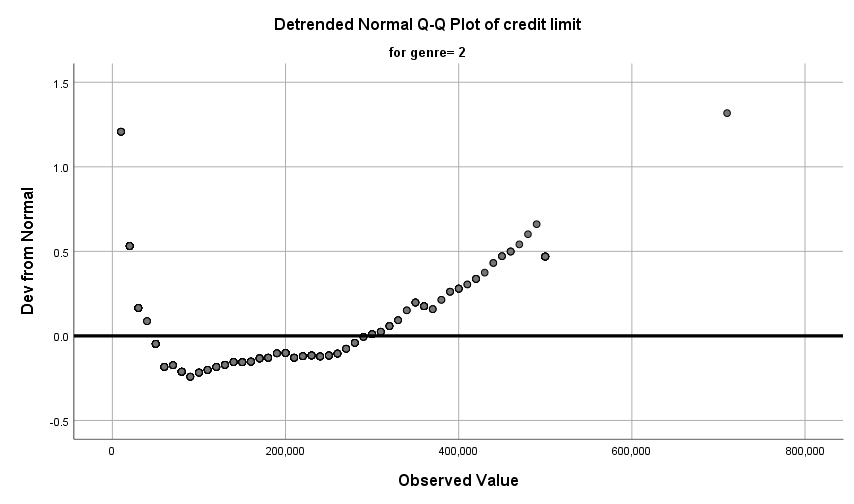


Figure 7. Detrended Q-Q Plot (Female)

A Kolmogorov-Smirnov test[[6]](#footnote-6) was performed in SPSS, resulting in an alpha value smaller than 0.05 (95% confidence), so the data is not normally distributed:

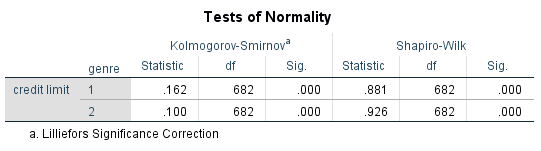


Figure 8. Kolmogorov-Smirnov Normality Test

[**Section V.1.**](#_Multiple_Linear_Regresion)

The dependent variable is the “Amount of Given Credit” (LIMIT\_BAL). To find the best independent variables, a bivariate correlation matrix has been performed in SPSS ([Appendix 2](#_Appendix_2:_Bivariate)). “Repayment Status in August” and “Amount of Bill Statement in September” were found to be the best predictors.

# TREND ANALYSIS

## Dashboard with Demographics & Trends

Figures 9 and 10 show the percentage difference between default and non-default. The lower they are, the greater the rate of default. University and High School educated holders present higher rates of default, and single people seem to default less than users with other marital statuses.

A picture containing chart

Description automatically generated

Figure 9. % Differ. between Default & Non-Default (by Education Level)

A picture containing graphical user interface

Description automatically generated

Figure 10. % Differ. between Default & Non-Default (by Marital Status)

In terms of genre and age, 21-40 female present slightly lower default rates (over the total percentage), and in higher ages the rate default/non-default is relatively similar in both sexes.

Chart, bar chart

Description automatically generated

Figure 11. Default percentage by Sex & Age

The “Repayment Status in September” is one of the best predictors of credit risk[[7]](#footnote-7). A payment delay of 3 months or more is a strong indicator of default risk, followed by a delay of 2 months and 1 month, subsequently.

Chart, bar chart

Description automatically generated

Figure 12. Default percentage by Repayment Status in September

An interactive dashboard has been created presenting “Demographics on Default of Credit Card Debts” (See [Appendix 3](#_Appendix_2:_Dashboard) for a large image). It will be integrated into the Credit Services Specialists’ CRM software. This information can be used to make informed decisions about credit allowance and management.

## Credit Allowance - Genre Bias Detection

A Mann Whitney U Test[[8]](#footnote-8) is performed in SPSS to test if genre bias in credit allowance exists.

= the median credit limit value is equal in both sexes

= there is a significant difference between the “male” and the “female” median credit limit value



The assumptions of the test are met, because the observations are independent and the sample distribution of the independent variable for each category have approximately the same shape [22]:

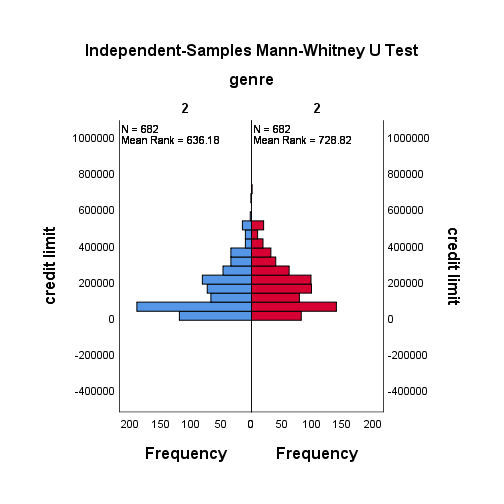
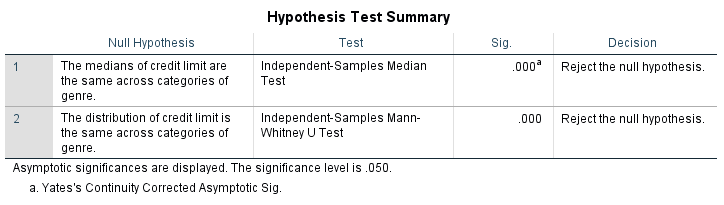


Figure 13. Sample Distribution by Genre

The test summary indicates that there is a statistically significant difference in the median value of both categories (U=264,149.5, p=0), therefore  is rejected in favor of . The evidence confirms that there is a genre bias in credit allowance.



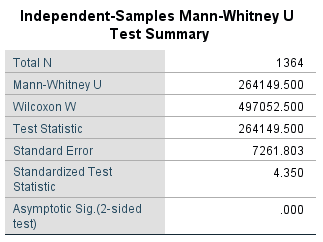


Figure 14. Mann-Whitney U Test Results

The following figure provides details of this bias. There is a higher percentage of men on the lowest credit limits (10,000 to 90,000) and the highest (350,000 to 1,000,000), and a higher percentage of women with credit limit between 100,000 and 340,000.

Chart, bar chart

Description automatically generated

Figure 15. Genre Distribution of Credit Limit

Figure 16 shows that the older the user, the higher their credit limit tends to be, and vice versa. As per Figure 17, university educated customers lead the low-medium range of credit, while school educated users abound in the highest credit limit ranges. That is consistent with age groups, because older generations were not as highly educated as nowadays, but their income is usually higher due to the long years work experience.

A picture containing timeline

Description automatically generated

Figure 16. Credit Limit by Age

Chart, bar chart

Description automatically generated

Figure 17. Credit Limit by Education Level

# MODELING

## Multiple Linear Regresion - Prediction of Credit Limit

The dependent variable is the “Amount of Given Credit” (LIMIT\_BAL) and the independent variables are “Repayment Status in August” and “Amount of Bill Statement in September”[[9]](#footnote-9).

= there is no relationship between the independent variables and the dependent variable



= there is a relationship between the independent variables and the dependent variable



A data sample containing these 3 variables have been loaded into SPSS and, as a result,  can be rejected in favour of  at 95% confidence level, since the Significant F Change value is smaller than 0.05. Therefore, there is a relationship between the set of independent variables and the dependent variable.

The predictive capacity of the model is weak, since R value is between 0.2 and 0.4, and a 11.7% of the variance of the predicted values can be explained by the independent variables (R Square) [23]:

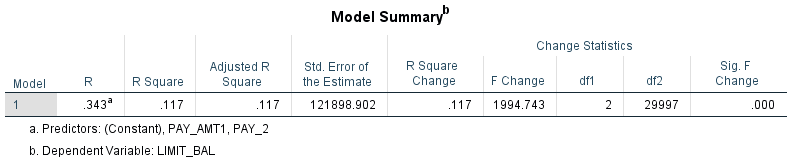


Figure 18. Regression Outcome

Shown below is the regression formula. It can be added as a featured credit score value on the back-office CRM of the bank, as a quick indicator of the overall credit health of the account[[10]](#footnote-10). Configuring it colour-coded (green to red for lower and higher credit limits) will help to interpret the score value in a sight.

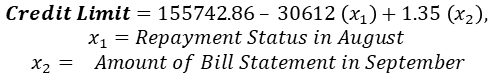




Figure 19. Regression Coefficients

## Decision Trees & Random Forests, testing and evaluation

This section will build a model that will be used to predict CC default and create training materials for Risk Prevention agents. Designing the model will follow this structure:

Diagram

Description automatically generated

Figure 20. Design of the Predictive Model

RF are considered state-of-the-art in terms of prediction performance, but DT are more easily interpretable. Using both types of algorithms will maximize the accuracy of the model and the extraction of insightful meaning. Specifically, C5.0, rpart, train (caret package), bagging and randomForest have been selected. The dataset has been split into 70% training (to build the model) and 30% test (to evaluate it).



Figure 21. Training Accuracy Results

The training results show that the highest accuracy was reached by RF algorithms, train(caret) and randomForest, with 98.5 and 99.32% accuracy, respectively. They also present high Kappa values. Bagging, C5.0 and rpart achieved significantly lower results.







Figure 22. Confusion Matrices of the 5 Algorithms

The confusion matrices display the actual predicted values. The column on the left refers to predicted and the top row refers to the actual values[[11]](#footnote-11). Newly, it is evident how train(caret) and randomForest outperform the others.



Figure 23. Test Accuracy Results

The test results present lower values than the training, overall[[12]](#footnote-12). This is due to the sample size, the greater it is, the better the predictions are. These results indicate that randomForest is the best default prevention model in terms of cost-savings and predictive power.

Turning now to knowledge materials for agents, Figure 24 shows variables indicating risk patters, ranked by importance. The “Repayment status in September” stands out as the most relevant attribute, followed by 15 predictors with a relative high importance and 7 with a lower one. Regarding Figure 25, it can be used as a flowchart for training purposes, but [Appendix 4](#_Appendix_4:_Flowchart) is preferred due to its greater level of detail.

Chart, scatter chart

Description automatically generated

Figure 24: Variable of Importance Plot

A picture containing diagram

Description automatically generated

Figure 25: Flowchart “Default Prediction”

## Twitter Text Scrapping, Sentiment Analysis & Topic Modeling

Marketers play an essential role in attracting creditworthy users while avoiding credit card losses [24]. Since this section is intended to create a solid marketing campaign on credit products, selecting an appropriate search query for Twitter scrapping is crucial. “Credit Score” returned a corpus of 979 tweets with rich content[[13]](#footnote-13). A calculator of the sentiment associated with the topic has been built with Bing lexicon. The topic is associated with a neutral sentiment, with a very slight negative incline.

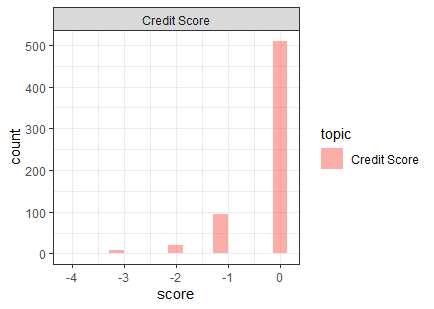


Figure 26. Overall Sentiment towards “Credit Score” topic

Words that are present over 25 times have been plotted (Figure 27) to understand what topics related with “Credit Score” are being discussed. “Finance”, “Mortgage”, “Home”, “Loan” and “Check” are topics that people like talking about.

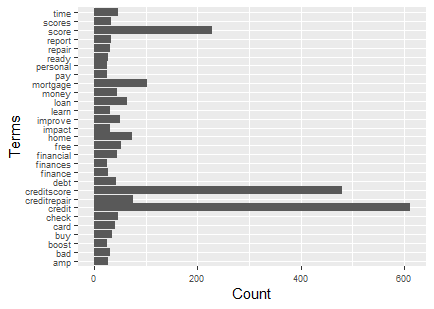


Figure 27. Word frequency of the Tweets corpus

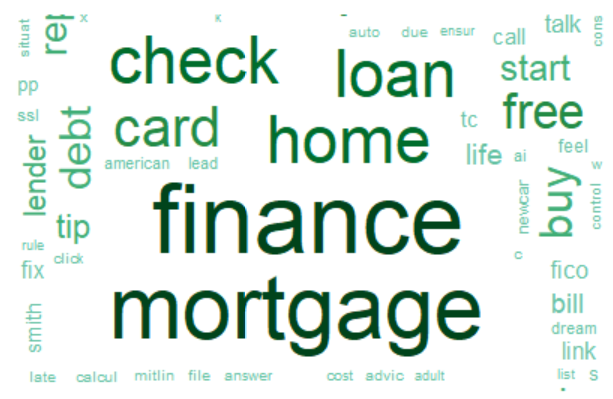


Figure 28. Word cloud “Most Discussed Topics”

As a general recommendation, using the words “Free”, “Improve”, “Ready”, “Boost”, “Affordable”, “Top” and “Easy” in the campaign will be associated positively with credit products[[14]](#footnote-14). Also, in our modern society, with consumers exposed to vast amounts of information, adding scientist figures or facts to the campaign will help building brand trust [25].

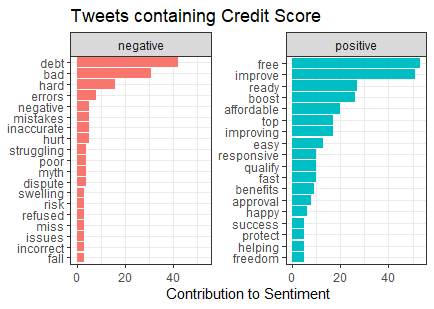


Figure 29. Words contributing to sentiment

LDA Topic Modelling has been built using 4 topics[[15]](#footnote-15). The meaning for each seems clearly differentiated, constituting a single entity, while presenting enough granularity and level of detail.



Figure 30. Topic Modelling Results: Top 10 Terms

It is encouraged to talk about the following topics in the marketing campaign. Topic 2 suggests being referred to personal finance and using consultancy services to make informed decisions. Topic 3 seems to talk about improving financial and time resources by using credit instruments (card, check, loan, mortgage). Lastly, Topic 4 appear to refer to long term investment decisions, such as buying a house or using bank lending services[[16]](#footnote-16).

# CONCLUSIONS

This project has successfully generated new tools and resources for 5 different departments, proving that outsourcing 240h of consultancy work by roughly 17,000 USD, a financial institution can implement significantly innovative changes. Risk Prevention agents have an interactive dashboard and new training materials, helping the early identification of payment fraud; Credit Services agents have a new color-coded feature on the CRM which will estimate the credit health status of the account; the Marketing team have clear indications on what content to focus the campaign in; and the bank will use a state-of-the-art machine learning algorithm to accurately predict payment default.

And most importantly, the reputation of the bank will be positively affected, as the Management Board is taking decisions based on insights analysis instead of intuition. Finally, the implementation of a new onboarding plan to educate employees on avoiding genre bias in their business decisions is showing a strong ethics and corporate social responsibility.

##### References

1. I-C. Yeh & C. H. Lien, “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. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[Accessed on: Dec. 01, 2020].

1. 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]
2. I. A. Telyukova & R. Wright, “A model of money and credit, with application to the credit card debt puzzle”, *Federal Reserve Bank of Cleveland*, Working Papers (Old Series) 0711, revised 2007. [Online]. Available: <https://www.clevelandfed.org/en/newsroom-and-events/publications/working-papers/working-papers-archives/2007-working-papers/wp-0711-a-model-of-money-and-credit-with-application-to-the-credit-card-debt-puzzle.aspx> [Accessed on: Dec. 02, 2020].
3. 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]
4. 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]
5. 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].
6. 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].
7. PwC, “Fighting fraud: A never-ending battle. PwC’s Global Economic Crime and Fraud Survey”, 2020 [Online] Available: <https://www.pwc.com/fraudsurvey> [Accessed on: Dec. 04, 2020].
8. 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].
9. 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].
10. 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].
11. N. J. Salkind, “t(ea) for Two (Again) Tests Between the Means of Related Groups”, *Statistics for People Who (Think They) Hate Statistics*, SAGE, page 288, 2017, 6th Edition.
12. N. J. Salkind, “The More Predictors the Better? Maybe”, *Statistics for People Who (Think They) Hate Statistics*, SAGE, page 376, 2017, 6th Edition.
13. G. James, D. Witten, T. Hastie & R. Tibshirani, “Lab: Decision Trees”, *An Introduction to Statistical Learning with Applications in R*, Springer page 323, 2017
14. 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].
15. 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].
16. 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].
17. 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].
18. W. Vorhies, “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].
19. I. H. Witten, E. Frank, M. A. Hall & C. J. Pal, “Credibility*”, Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann p. 181, 2017, 4th Edition
20. K. Bednara, S. Spiekermanna & M. Langheinrichb, “Engineering Privacy by Design: Are engineers ready to live up to the challenge?” *Information Society*, vol. 35 Issue 3, p. 122 - 142. May/Jun 2019 [Online] Available: <http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=bsu&AN=136089737&site=ehost-live> [Accessed on: Dec. 04, 2020].
21. Laerd Statistics, “Assumptions of the Mann-Whitney U test”, 2018 [Online] Available: <https://statistics.laerd.com/statistical-guides/mann-whitney-u-test-assumptions.php> [Accessed on: Dec. 04, 2020].
22. Statistical Consulting Group, “Regression Analysis | SPSS Annotated Output”, *Institute for Digital Research and Education*, 2016 [Online] Available: <https://stats.idre.ucla.edu/spss/output/regression-analysis/> [Accessed on: Dec. 04, 2020].
23. J-H. Oh & W. J. Jhonston, “Credit lender–borrower relationship in the credit card market – Implications for credit risk management strategy and relationship marketing”, International Business Review, Vol. 23, Iss. 6, Pages 1086-1095, 2014, ISSN 0969-5931 [Online]. Available: <https://doi.org/10.1016/j.ibusrev.2014.06.010>. [Accessed on: Dec. 06, 2020].
24. J. Brown, “Customers are more informed than ever – are you?”, *The future of consumer engagement and experience,* 2018 [Online] Available: <https://www.the-future-of-commerce.com/2018/08/13/customers-are-more-informed-than-ever/> [Accessed on: Dec. 06, 2020].

## **APPENDICES**

# Appendix 1: Requirements Elicitation

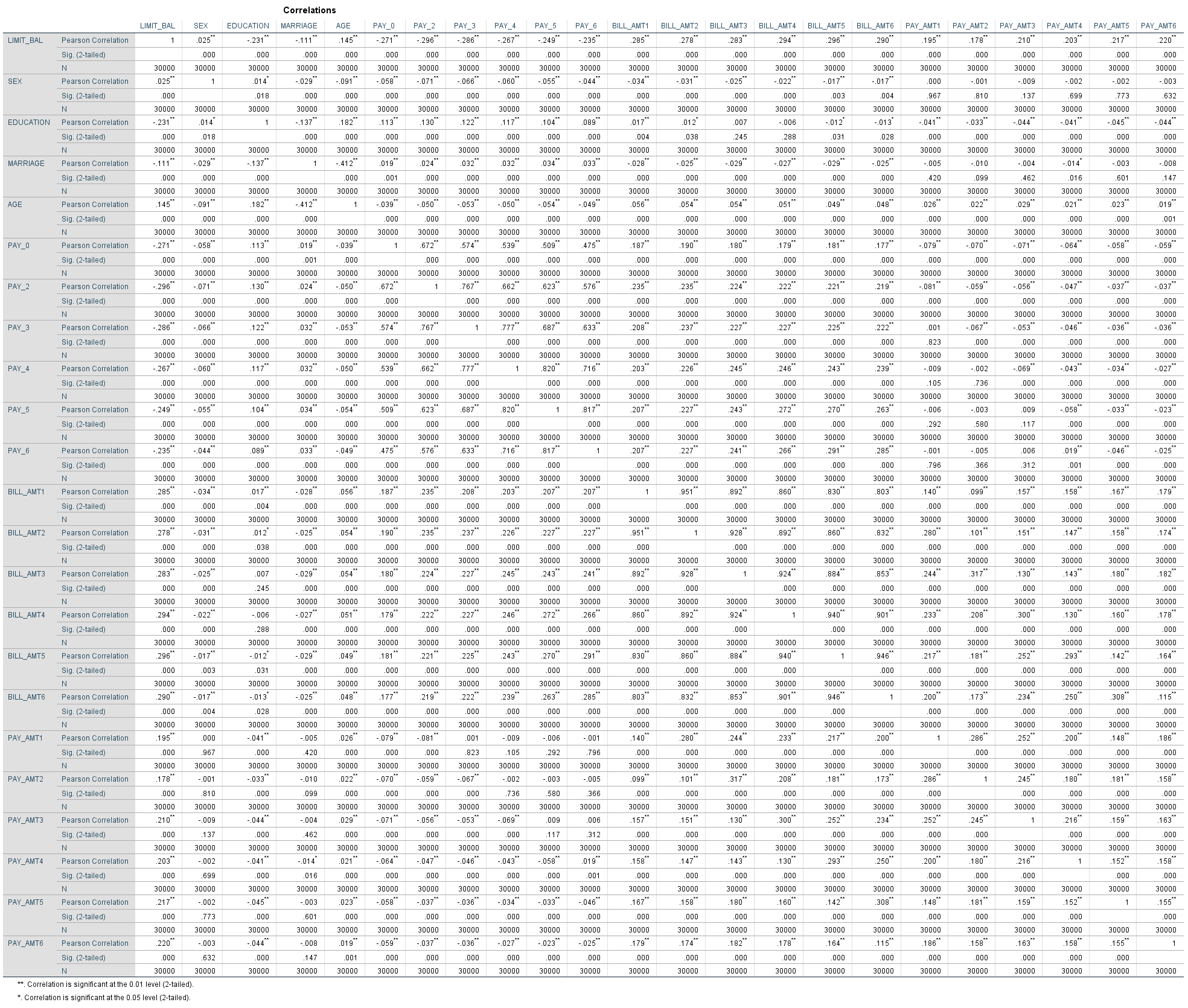
To reach an agreement on the services needed, a series of elicitation methods have been conducted:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

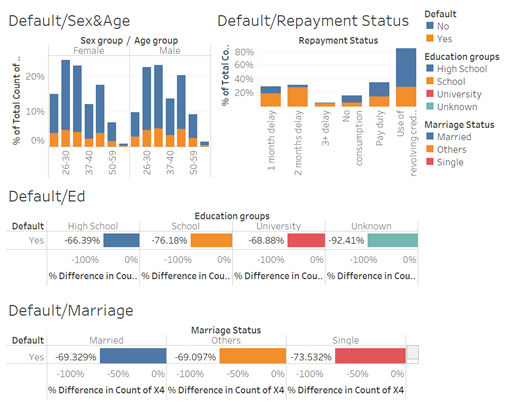
After these activities have been performed, the following requirements have been identified:



# Appendix 2: Bivariate Correlation Matrix (SPPS output)



# Appendix 3: Interactive Dashboard “Demographics on Default of Credit Card Debts”



# Appendix 4: Flowchart “Prediction of Payment Default”

Diagram

Description automatically generated

1. Sections I and III have been designed on the first sprint, followed by highly complex modelling (Section V.B). On a third sprint section II was built, leaving the complexion of simpler tasks for the latest stage [↑](#footnote-ref-1)
2. The first column is a unique identifier for each observation. There are two headers, of which the least descriptive has been deleted for an easier manipulation. [↑](#footnote-ref-2)
3. “Education” categories 0, 5 and 6 has been converted into 4, since they all represent “unknown” category. Equally, “Marriage” category 0 has been transformed into 3 (unknown). [↑](#footnote-ref-3)
4. “Sex”, “Education”, “Marriage”, “Pay\_0” and “Pay\_Def”. Also, “Age” and “Limit\_Bal” have been grouped into 5 and 7 categories, respectively. [↑](#footnote-ref-4)
5. These 2 simple random samples have been created in Excel, creating a list of random numbers, ordering them, and organizing the sample observations in accordance with that order. [↑](#footnote-ref-5)
6. This normality test is adequate for a sample size of 682. [↑](#footnote-ref-6)
7. See Figure 24. [↑](#footnote-ref-7)
8. The findings of non-normality have been discussed in [section III.3](#_Statistical_Analysis). [↑](#footnote-ref-8)
9. As discussed in [Section III.3](#_Statistical_Analysis). [↑](#footnote-ref-9)
10. The relatively weak prediction power of the model suggest that the actual credit capacity overall should be evaluated in conjunction with other indicators. [↑](#footnote-ref-10)
11. For example, for C5.0, 15,583 observations were correctly predicted as non-defaulted, while 2,962 were incorrectly predicted as defaulted. [↑](#footnote-ref-11)
12. Except for C5.0 and bagging, where they remain similar [↑](#footnote-ref-12)
13. To select the search query, multiple terms have been tried tentatively. “Debt” or “Default” have been found to be less useful, since they clearly have negative connotations. [↑](#footnote-ref-13)
14. Negative words such as “Debt”, “Bad”, “Hard” or “Errors” can be used to indicate users that using the bank’s credit they will not face problems associated with these words. [↑](#footnote-ref-14)
15. Combinations of top 10 terms for 2, 3, 4, 5 and 6 topics have been tested, to check which one is presenting a more consistent result (4 in this case). [↑](#footnote-ref-15)
16. Topic 1 seems to refer to technology. It is expected because the tweets have been extracted from an IT company, but it is not relevant for credit, the topic discussed here. [↑](#footnote-ref-16)