Factors affecting the likelihood of defaulting on a loan

Big Data Technologies Coursework

Author names withheld as requested

CS982 Big Data Technologies

Computer and Information Sciences

University of Strathclyde, Glasgow

November 8, 2023

This page intentionally left blank

Contents

List of Figures

List of Tables

List of Tables

Chapter 1

Introduction

In March of 2022, a major credit firm published a worrying press release labelling five million British people as 'credit invisible', a term referring to 'people with little or no credit history, greatly reducing their access to mainstream financial services' (Experian 2022).

'Credit invisibles' come from many walks of life, including: young people, older people, and recent immigrants, all without credit backgrounds. This is a major problem for loan applicants, as most UK-based lenders rely on credit scores provided by credit rating agencies (CRA) to assess applicant suitability (Responsible Finance 2023), a heavily criticized system that first emerged around the 1950s (Cashfloat no date). Applicants without proper credit history are often assessed on other factors such as their postcode, a practice deemed even more unfair than the use of credit history by respondents to a 2023 opinion poll (Responsible Finance 2023).

While the specifics of how CRAs calculate credit scores within the UK is a guarded secret, it is widely known that credit history is a major part of the formula (Barclaycard no date). This report will cover an investigation conducted into the main contributory features in the process of determining a loan applicants likely hood to default on their loan (by checking the correlation of top features). The results of simplifying predictive models by removing some controversial features will also be explored for the purposes of increasing fairness and making loans more inclusive.

Chapter 2

Dataset

2.1 Source

All of the data used in this study is freely available as part of the 'Home Credit Default Risk' competition data-set, which can be found on the website 'Kaggle' after registering for a free account. Of the ten available files, only 'HomeCredit_columns_description.csv' and 'application_train.csv' were used, with the former serving only to provide contextual information during exploratory data analysis regarding the contents of columns in the latter.

2.2 Rationale

The providers of the data-set, Home Credit, didn't specify how or where the data was collected from, or if it was even real. We can however assume that since the data came from a paid kaggle challenge that was more than likely a direct effort by Home Credit to improve their own algorithms, it is likely to be either real anonymized data from past customers of Home Credit or simulated data based on past experience. It is for this reason, and the fact that the dataset is of substantial size, that our team decided to use it to carry out our investigation.

2.3 Description

The 'application_train.csv' and 'HomeCredit_columns_description.csv' are both '.csv' files. Post cleaning/processing, the 'application_train.csv' file is found within the 'df_clean' DataFrame, having 109 columns, each representing a feature (except for the 'TARGET' column which is our dependant variable) and 304'525 rows, each representing a loan from HomeCredit.

2.4 Engineered Features

The final 'clean_applications' DataFrame contains a number of engineered features as described in Table 2.1 below.

Table 2.1: Engineered Features

Feature	**Description**
EXT_MEAN	Mean of EXT_SOURCEs 1, 2 and 3 (Normalized external credit scores).
EXT_STD	Standard deviation of EXT_SOURCEs. Aggregated information is significant as mentioned.
GOOD_CREDIT	AMT_GOODS_PRICE to AMT_CREDIT ratio. Indicates purchase's credit financing portion.
ANNUITY_INCOME	AMT_ANNUITY to income ratio. Shows income percentage for loan repayment.
CREDIT_INCOME	Credit to income ratio. Compares credit size to customer's income.
CNT_FAM_INCOME	Income per family member. Higher values suggest more resources per person.
INCOME_EMPLOYED	Income to DAYS_EMPLOYED ratio. Compares earnings to work experience.
INCOME_BIRTH	Income to age ratio. Reflects financial stability and planning.

Chapter 3

Features by Category

Due to high dataset dimensionality, features have been grouped into categories and preliminary analysis is provided for each.

3.1 Document Flag

The features in this category are all binary indicators of whether or not the client provided a certain document when applying for their loan, and mostly exhibit weak correlations meaning high feature Independence as seen in Figure 3.1

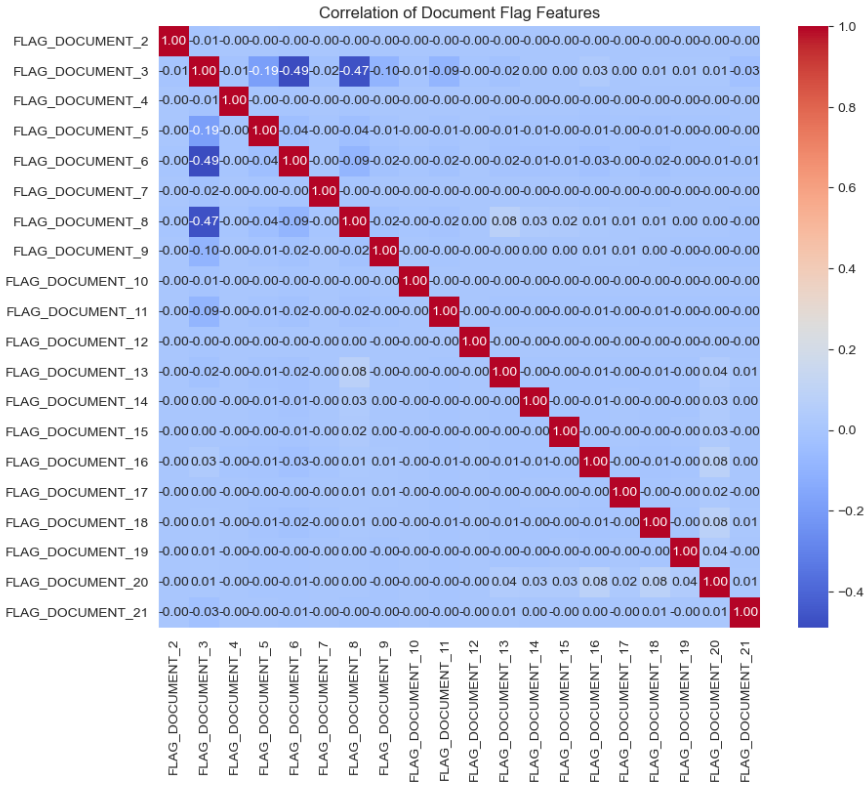

Figure 3.1: Document Flag Interdependencies

Figure 3.2 Shows the feature importance of each document flag in terms of their correlation with the target variable. Interestingly, the document 3 flag exhibits a strong interrelationship with the target variable, far beyond that seen with any other document.

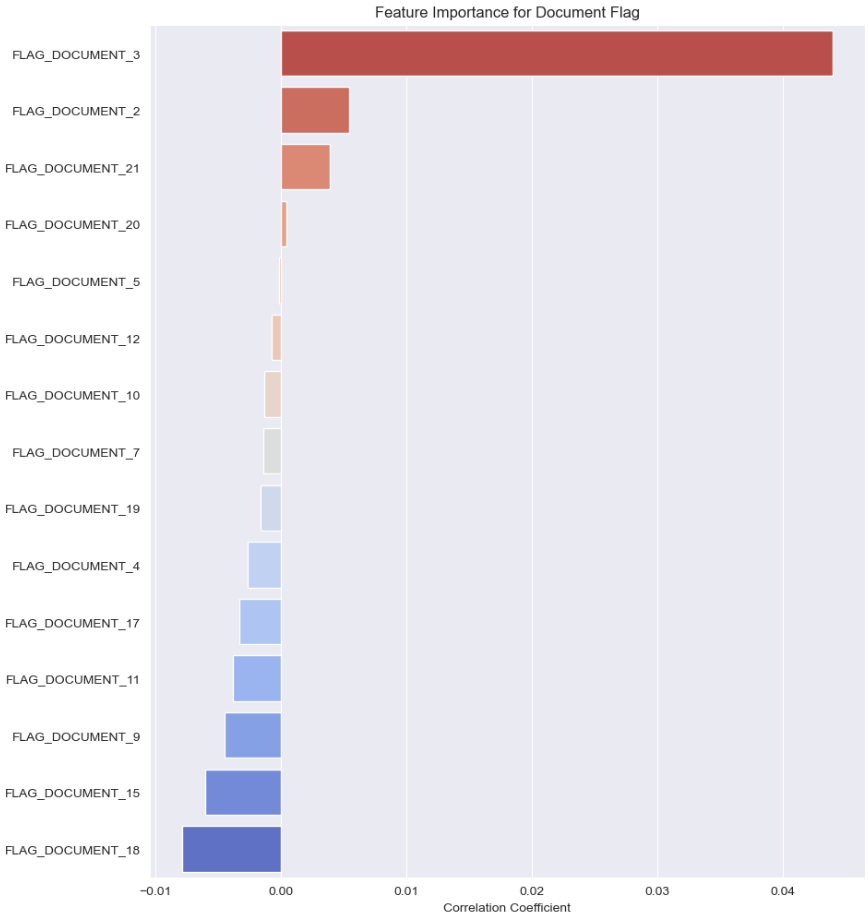

Figure 3.2: Document Flag Feature Importance

3.2 Client Personal

This category describes personal details about the client (gender, age etc..). Some features have high correlations such as CNT_CHILDREN and CNT_FAMILY_MEMBERS which makes logical sense. This can all be seen in Figure 3.3.

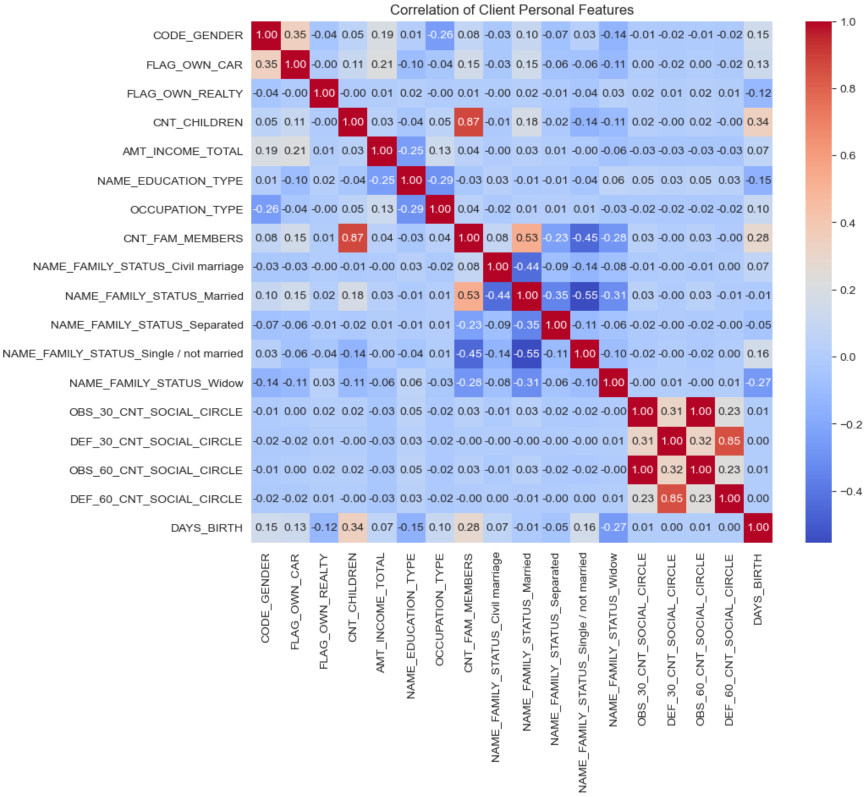

Figure 3.3: Client Personal Interdependencies

In Figure 3.4 we can see that the strongest relationships between personal features and the target variable are found in a clients age and education type.

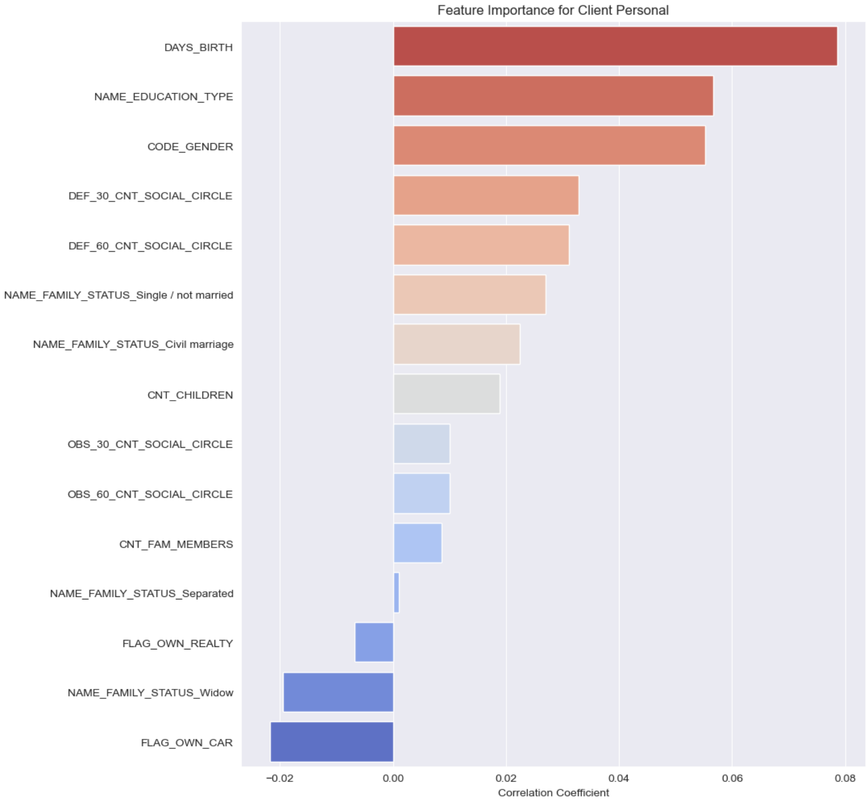

Figure 3.4: Client Personal Feature Importance

3.3 Credit

This category relates to credit history and shows weak Interdependencies as seen in Figure 3.5. Some notable exceptions include the flag for 'Revolving Loans' and 'Cash Loans'. This perfect correlation is due to the fact that a client may not have more than one loan type. We also see strong positive correlations between a number of engineered features, and the features from which they were engineered, such as EXT_MEAN and EXT_SOURCE_2.

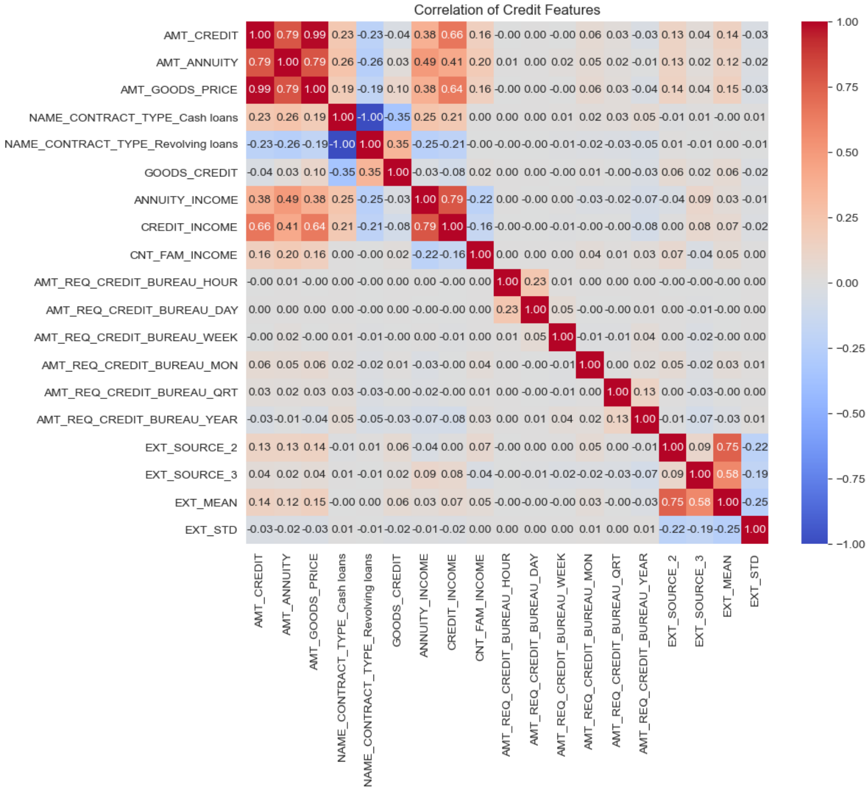

Figure 3.5: Credit Interdependencies

Unsurprisingly, as seen in Figure 3.6, we can see that the strongest correlation for credit related features comes from the standard deviation of the external source, which is a credit rating from an external agency.

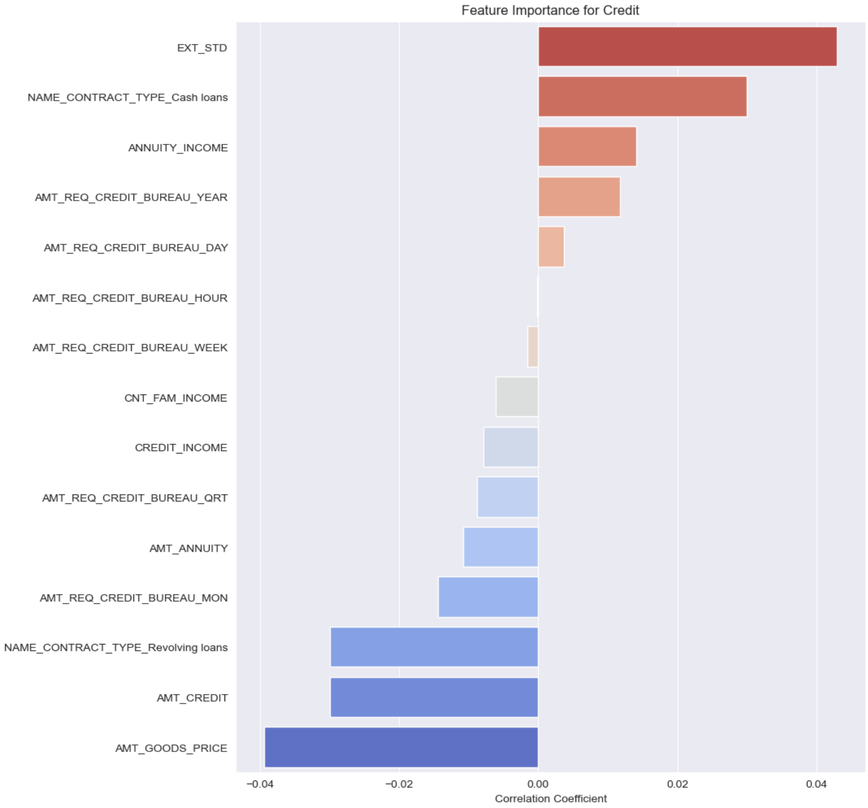

Figure 3.6: Credit Feature Importance

3.4 Contact

Features in this category are ’flag’ columns for the most part and exhibit no significant levels of correlations across the board. Interdependencies here are generally very weak as can be seen in Figure 3.7

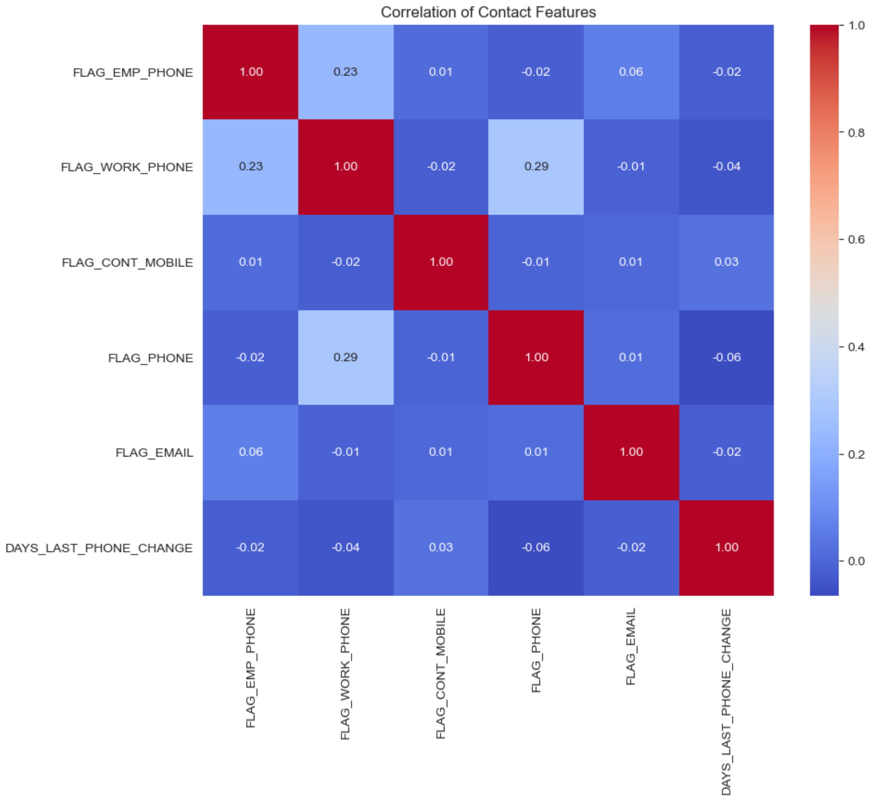

Figure 3.7: Contact Interdependencies

Figure 3.8 Shows a surprisingly strong relationship between the day the client last changed their phone and the target variable.

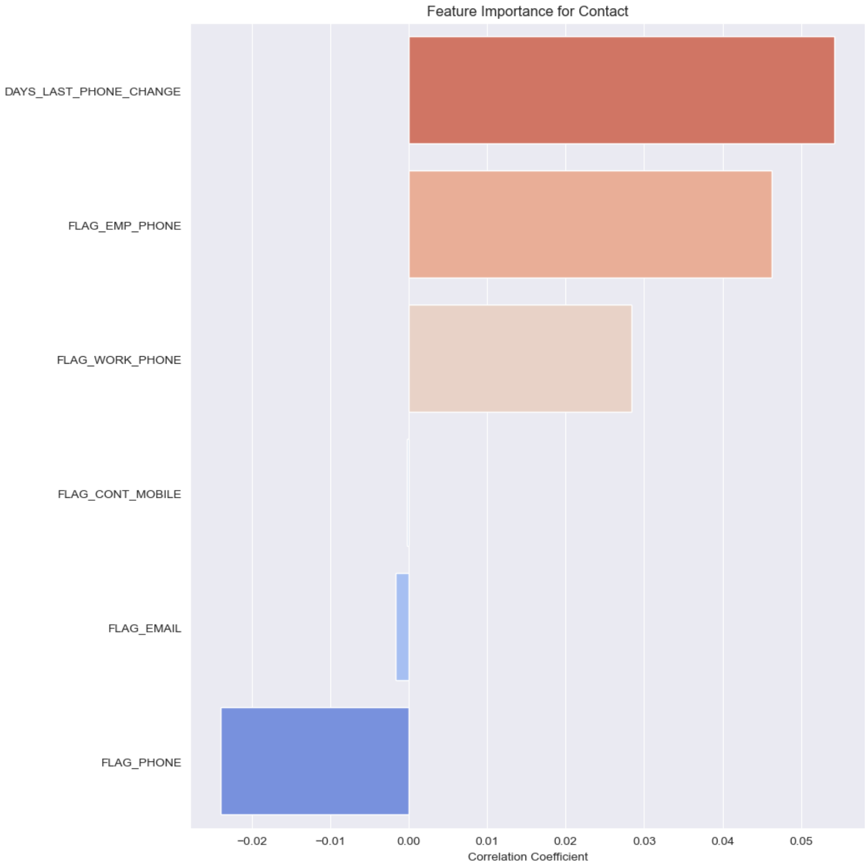

Figure 3.8: Contact Feature Importance

3.5 Employment

These employment related features exhibit weak Interdependencies as seen in Figure 3.9, with the exception of the relationship between ORGANIZATION_TYPE and DAYS_EMPLOYED. This correlation is likely due to the fact that turnover rates differ between organizations.

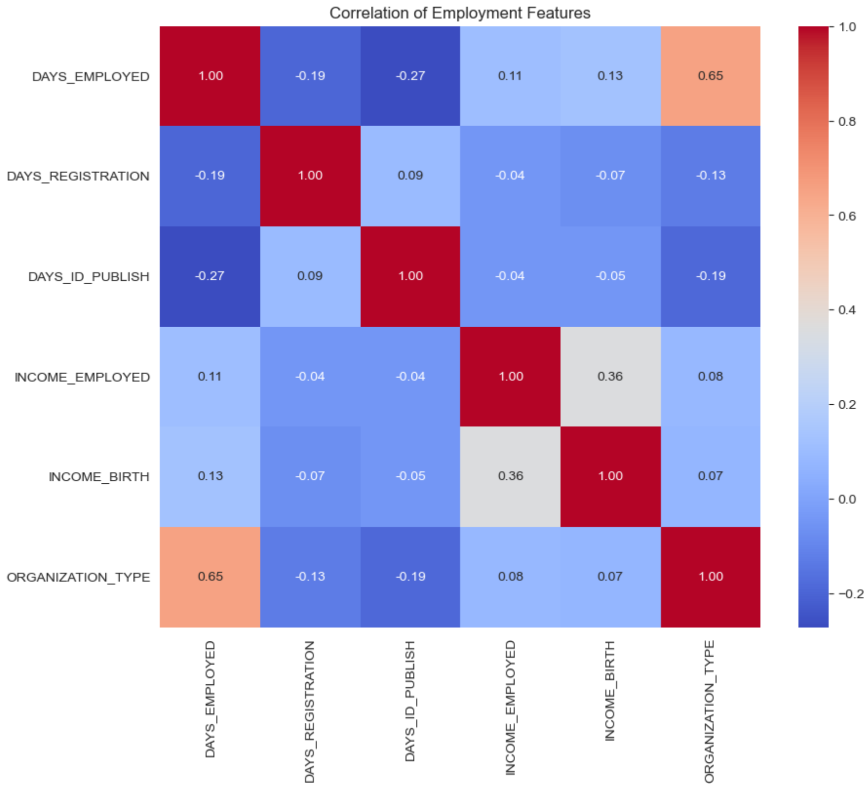

Figure 3.9: Employment Interdependencies

Figure 3.10 shows a reasonably strong negative relationship between a clients organization type and the target variable.

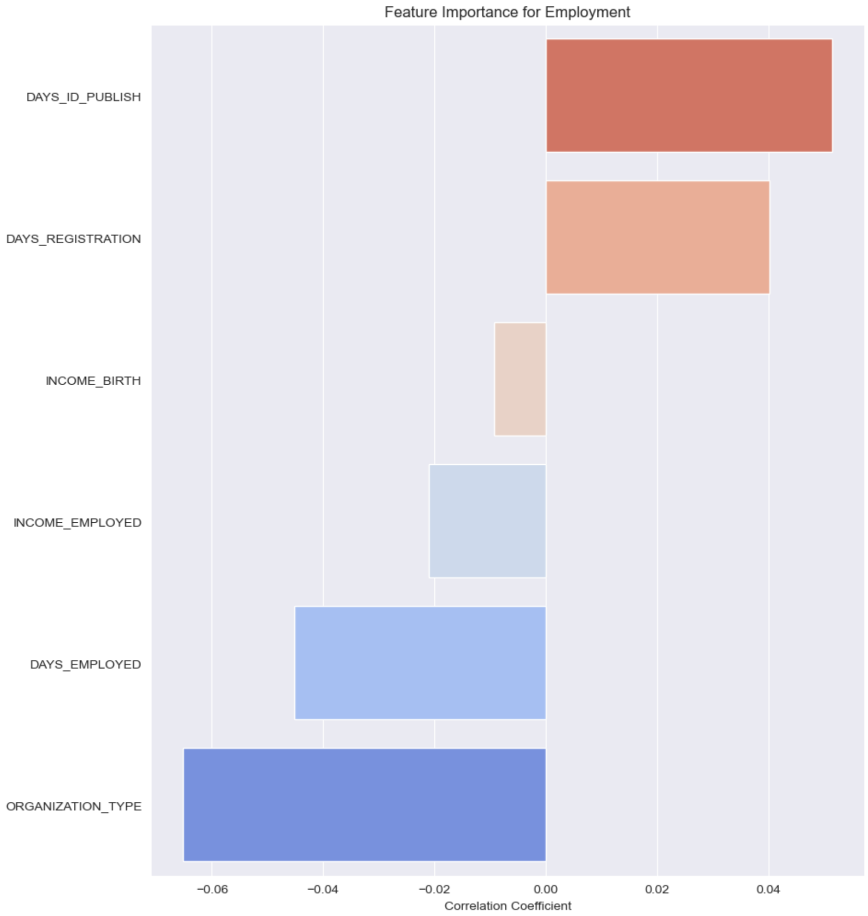

Figure 3.10: Employment Feature Importance

3.6 Address

Address related features exhibit several strong correlations are notable when looking at Figure 3.11, including the reasonably strong relationship between living with parents and living in an apartment. There is also a visible negative correlation between the rating of a city and its population.

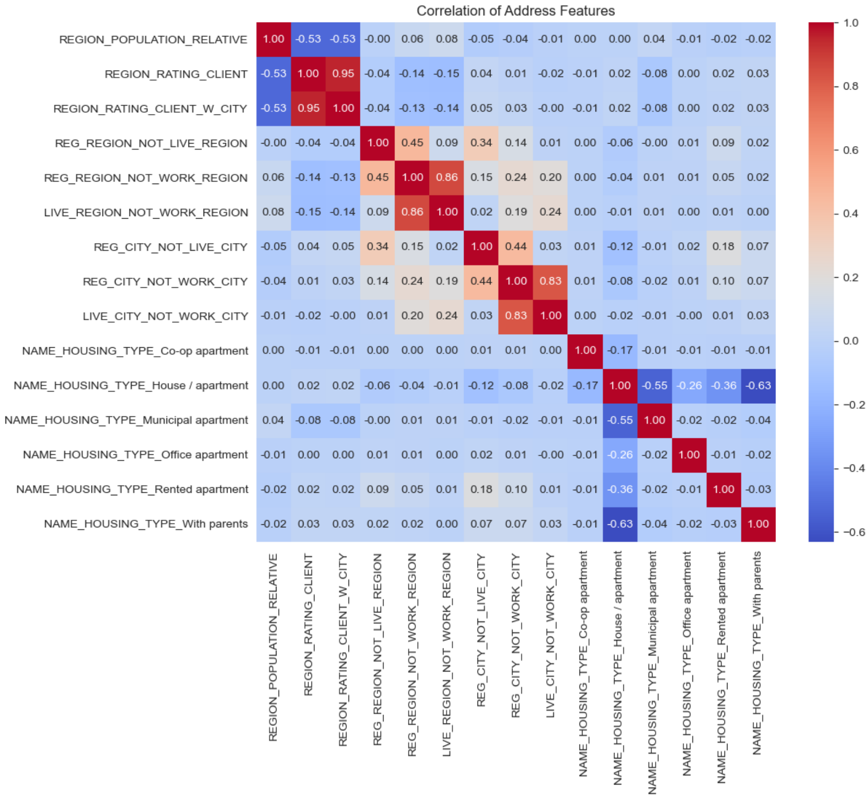

Figure 3.11: Address Interdependencies

The strongest correlation observed so far can be seen in Figure 3.12 in the relationship between the rating of a clients city and the target variable.

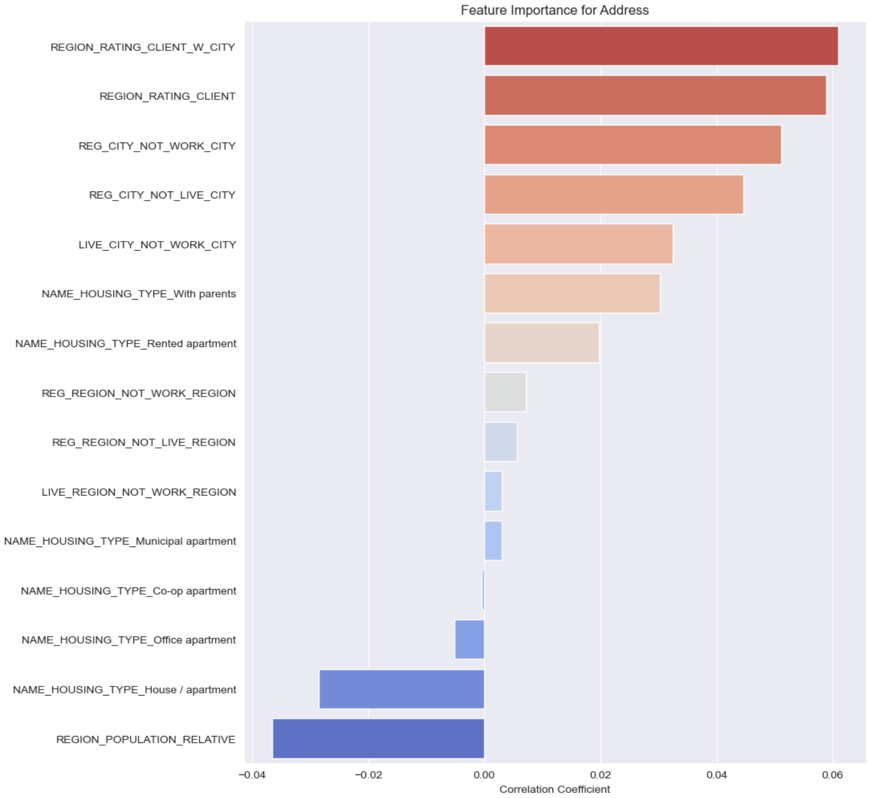

Figure 3.12: Address Feature Importance

3.7 Loan Application

Features in this category contain information about a clients loan application. There are unusually high correlations visible in Figure 3.13 for this category and this is likely due to the fact clients might tend towards conducting their non work related business on a set number of days each week.

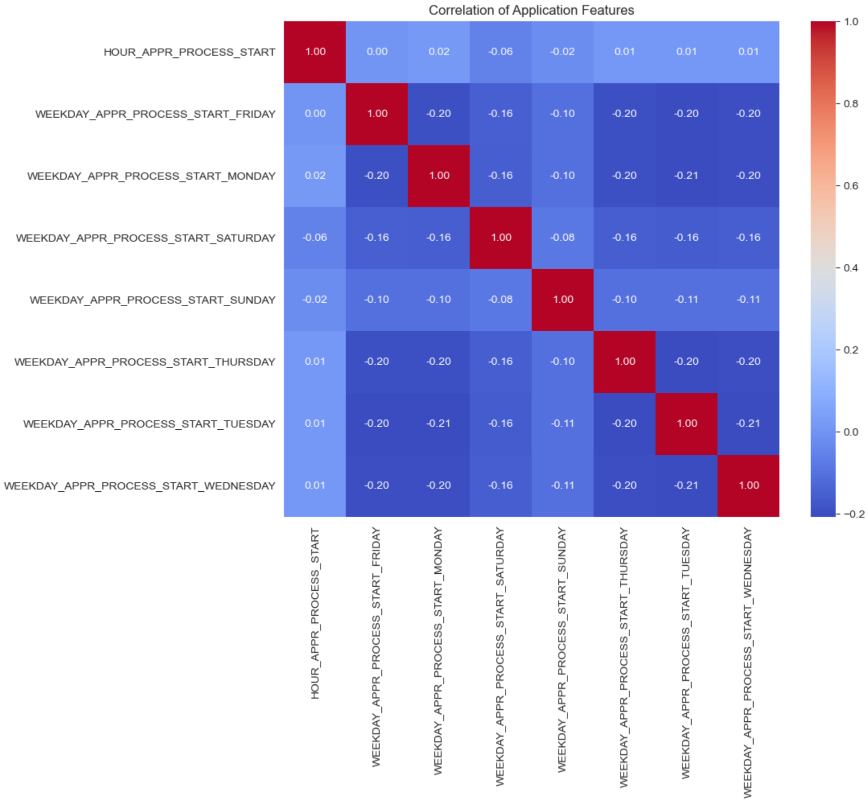

Figure 3.13: Loan Application Interdependencies

Very weak feature importance is observed across the board in Figure 3.14, with the highest levels being observed in the hour the application process was started.

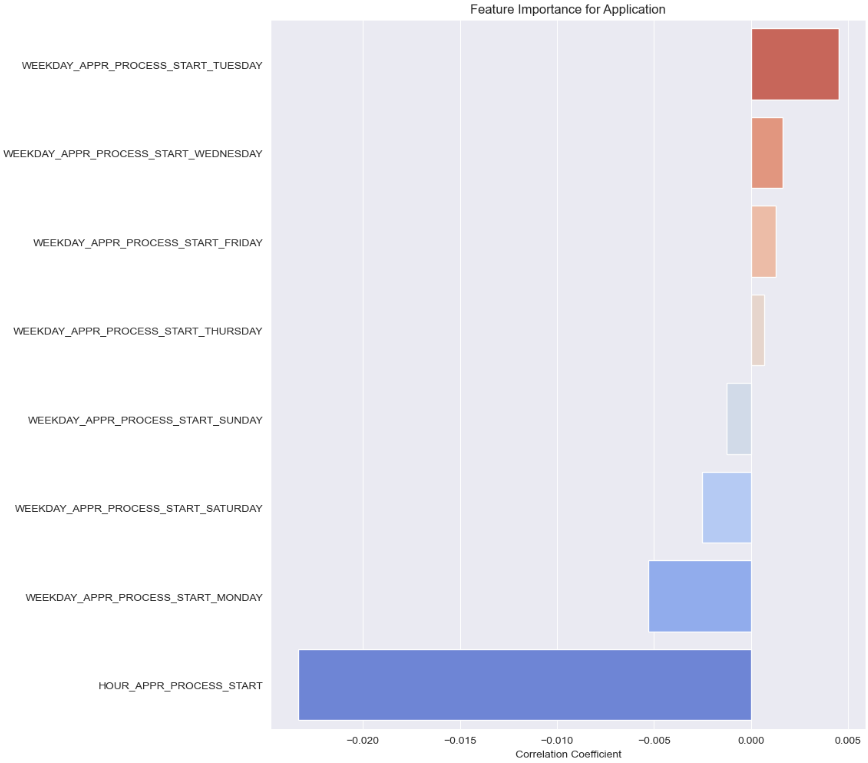

Figure 3.14: Application Feature Importance

3.8 Top features

The Interdependencies between the top 15 features, as well as their feature strength can be seen below in Figures 3.15 and 3.16 respectively.

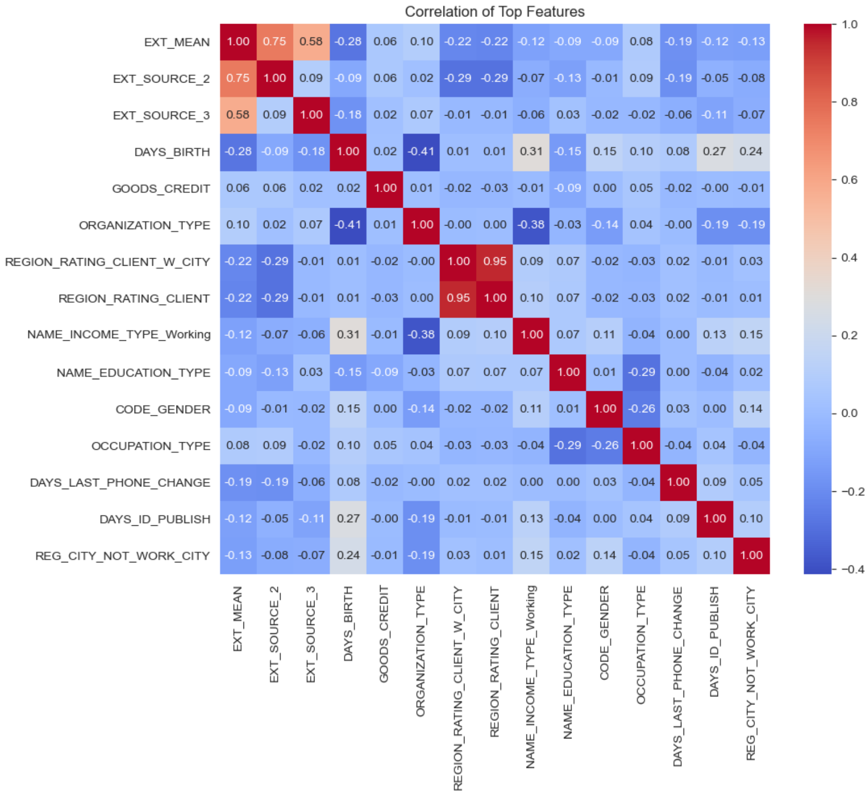

Figure 3.15: Top Features Interdependencies

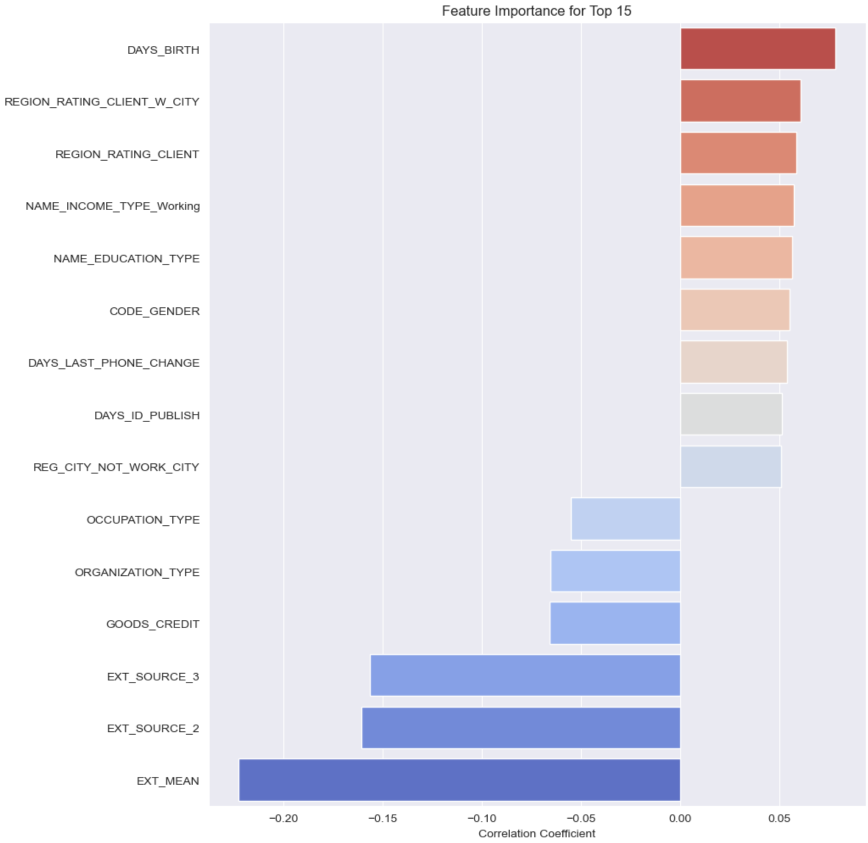

Figure 3.16: Top Features Importance

Chapter 4

Unsupervised Analysis

4.1 Method Selection and Rationale

While the dataset is already labeled so using unsupervised learning may not be as beneficial but we can still perform EDA on our dataset. As a binary classification problem, we already know there are 2 clusters in our dataset, divided by the TARGET feature. From this understanding, we will use k-means to perform clustering on different sub set of the dataset to see if any clear distinction between the 2 group can be drawn. Since the dataset is labeled, we will compare the label generated from k-means clustering to the actual labels and calculate Completeness and Homogeneity score, this way we will have a quantitative method to validate the clustering result.

4.2 Application and Results

With the full dataset including more than 300.000 rows, the process could take a lot of time so we will only use an under-sampled subset of 14800 rows. From this, we will run K-means using:

- All 108 of the features (Figure 4.1)
- Only the top 15 most correlated features to TARGET (Figure 4.2)
- Using only the top 2 most correlated features: EXT_MEAN and DAYS_BIRTH (Figure 4.3)

When performing clustering on a high dimension dataset in scenario (1) and (2), we will apply PCA to reduce the result down into a 2 dimension result for purpose of visualization.

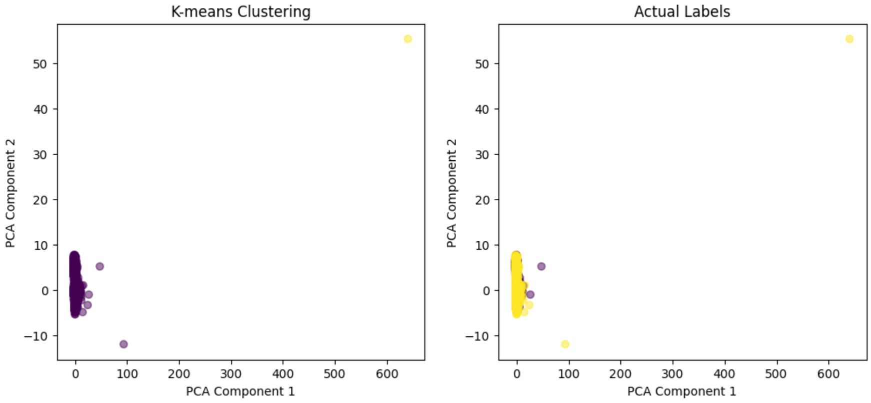

Figure 4.1: All 108 of the features clustering

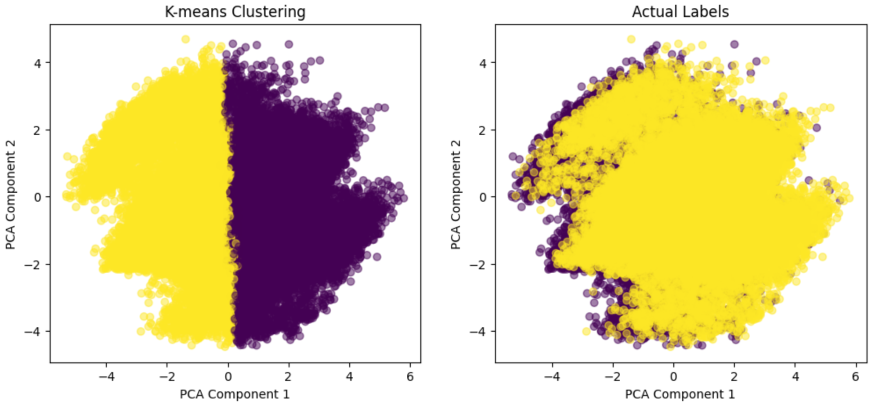

Figure 4.2: Top 15 most correlated features clustering

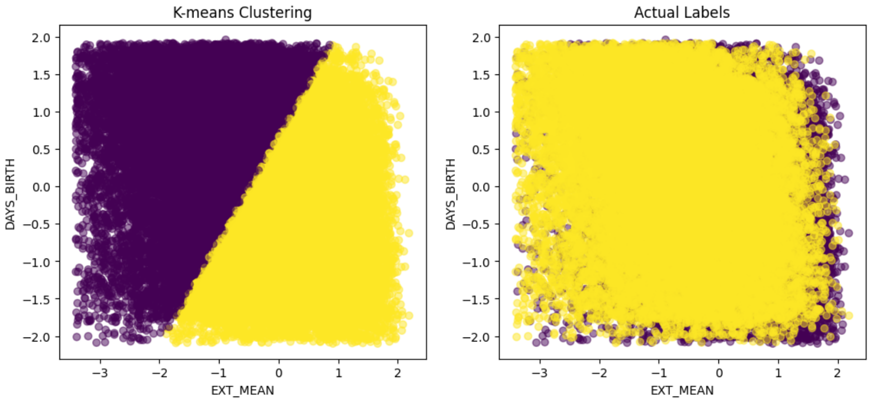

Figure 4.3: Top 2 most correlated features clustering

From figures 4.1, 4.2 and 4.3, no distinction is seen between the default and non-default group. When comparing subset scores in Table 4.1, having all the features appears to further confuse the clustering. Having selected features is evidently better.

For clustering, we want results to be as close to 1 as possible, but in reality, even our top performer, using the top 15 correlated features, does not return with a score above 0.1. As conclusion, there does not seem to exist a simple distinction between the default and non-default loans.

Table 4.1: Completeness and Homogeneity score of the clustering

	All features	Top 15 correlated features	Top 2 correlated features
Completeness score	0.05871	0.07810	0.06272
Homogeneity score	2.0271e-05	0.07783	0.06265

Chapter 5

Supervised Analysis

5.1 Method Selection and Rationale

As can be seen in figure 5.1, the data-set is highly imbalanced, with most loan being non-default (TARGET = 0). This make intuitive sense since a high default rate will result in bankruptcy.

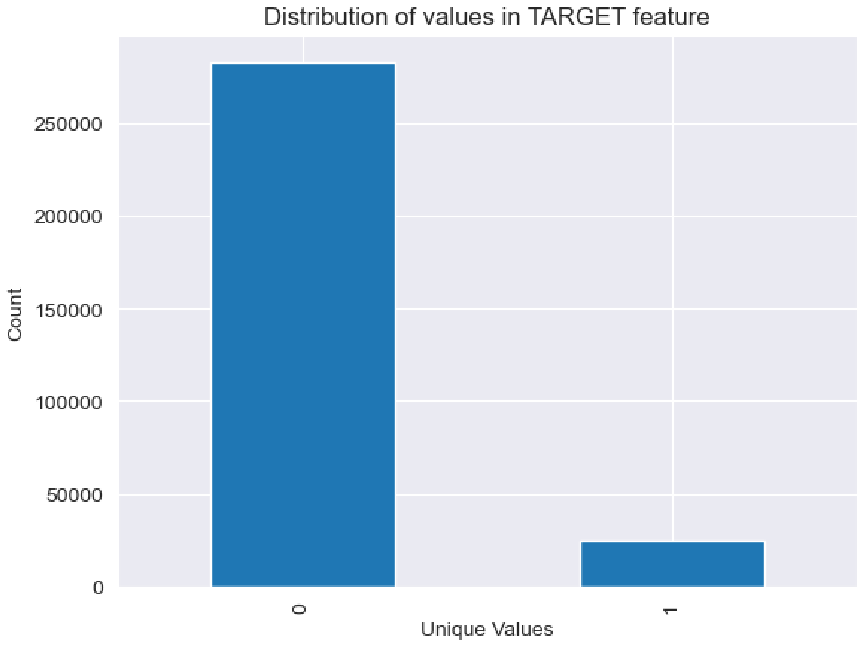

Figure 5.1: Imbalance dataset with mostly non-default loan

With an imbalanced dataset, the model will most likely classify everything as the majority resulting in high accuracy but completely missing minority which is what we are more concerned with. To mitigate this, we consider 2 approaches (Mastery no date):

- Ovesampling: randomly reselect rows from the minority class until the dataset balanced
- Undersampling: randomly remove rows from the majority class until the dataset balanced

In case the relationship between TARGET and other independent features is non-linear, we consider raising the dataset to a degree through polynomial features. Logistic regression is chosen for machine learning (Tensorflow no date). The output of a logistic model is probability of an input belong to the positive class, and mapped in the (0,1) range so the result is intuitive and easy to explain. We will also try other machine learning methods, including Decision Tree, K-Neighbors, Gaussian, and Random Forest and compare the performance. Finally, we will try 2 simplified scenarios in the applications process:

- Removing any features related to external sources, in case of first time borrowers
- Removing features relating to borrowers address and postcode, to avoid discrimination

5.2 Model Training and Validation

For model validation, we will use:

- Precision: classifier ability not to label as positive a sample that is negative (scikit learn no date, a).
- Recall: classifier ability to find all the positive samples (scikit learn no date, b).
- F1-score: harmonic mean of the precision and recall (scikit learn no date, c).

- ROC AUC score: Area Under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores - ROC score base line is at 0.5 (complete random) and peak at 1.0

We will also measure the elapsed time of different methods but this will be secondary as each methods will have different tuning options and set a max number of iteration to 5000, limiting runtime.

5.3 Results and Insights

First we observe the important features from the Logistic (Forecastegy 2023), Decision Tree and Random Forest Classifier (Medium 2020) (Figures 5.2. 5.3 and 5.4)

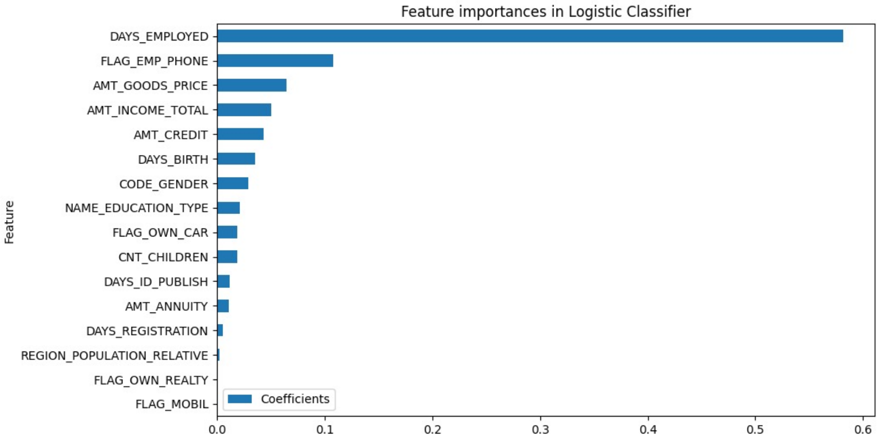

Figure 5.2: Feature importance in Logistic Classifier

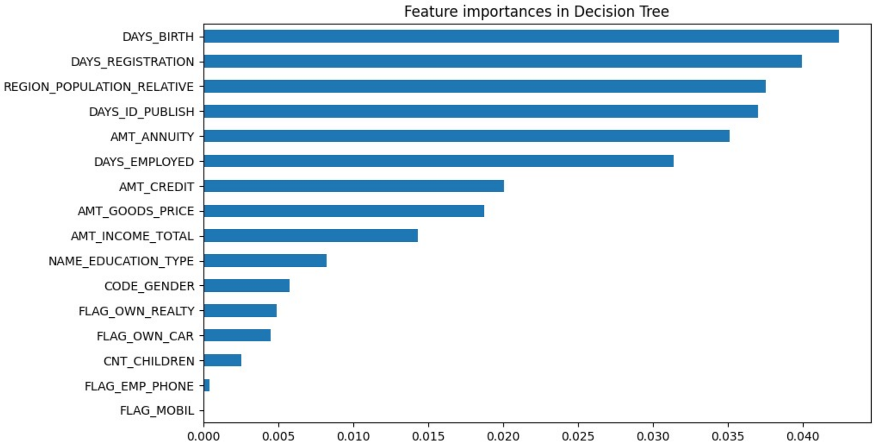

Figure 5.3: Feature Importances in Decision Tree classifier

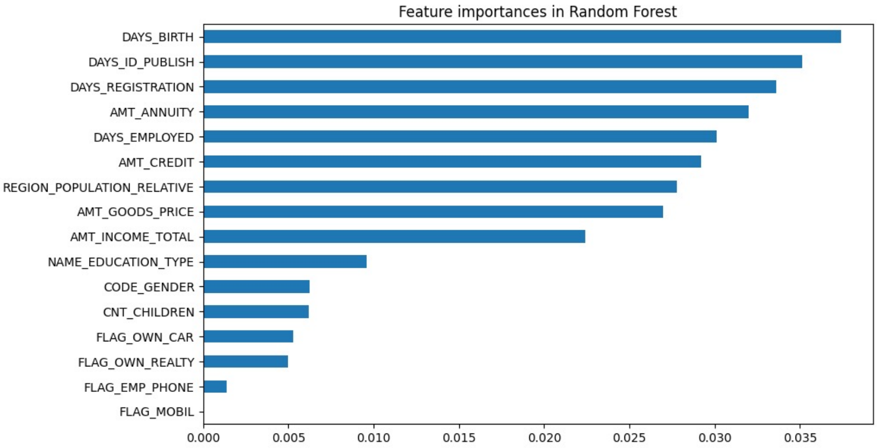

Figure 5.4: Features Importances in Random Forest Classifier

To find defaults, we will only show the result of the model for capturing TARGET = 1

Table 5.1: Performance of different machine learning methods and transformation

Subset	**ML method**	**Precision**	**Recall**	**F1 score**	**ROC AUC**	**Time (s)**
Full dataset	Logistic	0.5	0.01	0.02	0.5	20.12
Oversampled	Logistic	0.68	0.68	0.68	0.68	97.32
Undersampled	Logistic	0.68	0.68	0.68	0.68	2.7
Undersampled +Top 15 correlated +Polynomial features	Logistic	0.68	0.67	0.67	0.68	98
Undersampled	DecisionTree	0.58	0.59	0.59	0.58	5.63
Undersampled	KNeighbors	0.59	0.59	0.59	0.59	19.6
Undersampled	GaussianNB	0.56	0.84	0.67	0.58	0.12
Undersampled	RandomForest	0.69	0.66	0.67	0.68	19.3
Undersampled -EXT_SOURCE	Logistic	0.63	0.63	0.63	0.63	2.45
Undersampled -ADDRESS	Logistic	0.69	0.69	0.69	0.69	2.53

From the above table, key learning can be draw:

- Using the dataset without resampling will result in a useless model
- Logistic Regression perform relatively well across the board
- Oversampling and undersampling result in equal results but since undersampling mean a smaller dataset and faster training time (97s comparing to 2.7s) so later model will use the undersampled dataset
- Gaussian is actually the fastest model, but among the worst classifier
- Adding polynomial and interaction features does not improve the classifier
- Removing features from external sources result in a weaker model
- Removing features relating to address slightly improve the model performance
- Random Forest, an ensemble approach to Decision tree, really result in a better classifier

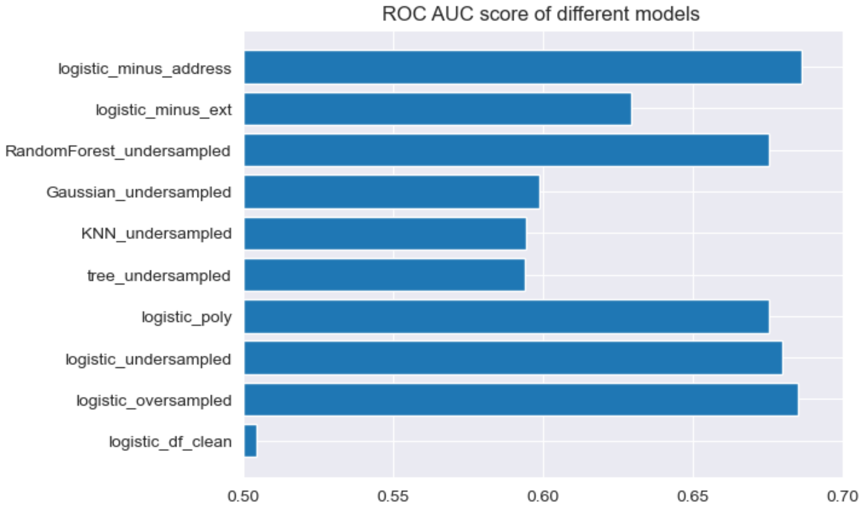

Figure 5.5: ROC AUC score of different classifier (baseline at 0.5)

Chapter 6

Reflection and Discussion

6.1 Reflection on Methods

6.1.1 Feature engineering

3/9 of the engineered features, including EXT_MEAN, OCCUPATION_TYPE, ORGANIZATION_TYPE are in the top 15 most correlated so the planning for features engineering seem to be on the right path

6.1.2 Logistic Regression

Logistic perform really well in this binary classification problem. But the resulting classifier are too reliance on a small number of features, despite we already scaled the dataset.

6.1.3 Decision Tree and Random Forest

Despite poor performance, decision tree provides a simple, highly interpretable model. It is also possible to visualize the tree as seen in Figure 6.1 below.

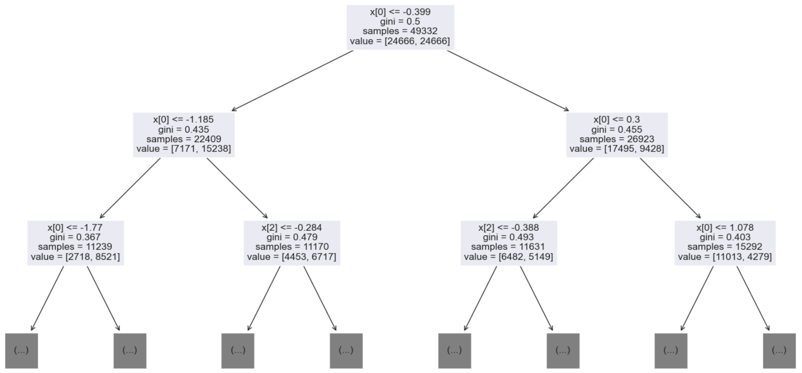

Figure 6.1: Visualizing the output from Decision Tree classifier

On the other hand, RandomForest is a ensemble learning method that combines multiple decision trees to produce a more robust and accurate model. But the increased in performance may also reduce the level of interpretability

6.1.4 Removing features from external source and address

Externally sourced features play key roles in the training process, but even without them, the classifier performs. Hence, we can consider first time borrower as riskier but accessible, and consider alternative loan plan or lower credit limit As for features relating to address, the classifier does not suffer at all, this suggest we could consider removing these features from application process

6.2 Limitations and further improvement

Top 'Credit kaggle Competition' performers have ROC AUC score around 0.81 so there is room for improvement. We didn't utilise other application processing datasets, including credit card history and past applications. As for machine learning methods, we have not include any hyperparameter tuning or ensemble learning in our process.

Appendix A

Development Details

A.1 Integrated Development Environments Used

A.1.1 PyCharm

PyCharm 2023.2.3 (Professional Edition) — Licensed to Christopher Turnbull — Subscription is active until October 18, 2024. — For educational use only. — Runtime version: 17.0.8.1+7-b1000.32 amd64 — VM: OpenJDK 64-Bit Server VM by JetBrains s.r.o. — Windows 11.0 — GC: G1 Young Generation, G1 Old Generation — Memory: 4066M — Cores: 24 — Registry: ide.experimental.ui=true

A.1.2 Google Colab

Google Colab

A.2 Packages Used

- os
- time
- zipfile
- numpy
- pandas

- seaborn
- copy
- sklearn - version 1.2.2 (! pip install scikit-learn==1.2.2)
- imblearn (! pip install -U imbalanced-learn)
- matplotlib

A.3 Python

- Python 3.10.12
- Python 3.11

Bibliography

Barclaycard (no date). *How is credit score calculated?* https://www.barclaycard.co.uk/personal/money-matters/credit-score-basics/how-is-your-credit-score-calculated. [Online; accessed 3-November-2023].

Cashfloat (no date). *The Origins of Credit Scoring Systems.* https://www.cashfloat.co.uk/blog/personal-finance/origins-credit-scoring-systems/. [Online; accessed 1-November-2023].

Experian (2022). *Meet the 5 million credit-invisible Brits still at risk of exclusion from the financial system.* https://www.experianplc.com/newsroom/press-releases/2022/meet-the-5-million-credit-invisible-brits-still-at-risk-of-exclusion-from-the-financial-system. [Online; accessed 1-November-2023].

Forecastegy (2023). *How To Get Feature Importance In Logistic Regression.* https://forecastegy.com/posts/feature-importance-in-logistic-regression/https://forecastegy.c
importance-in-logistic-regression/. [Online; accessed 29-October-2023].

Mastery, Machine Learning (no date). *Random Oversampling and Undersampling for Imbalanced Classification.* https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/. [Online; accessed 29-October-2023].

Medium (2020). *Random Forest on Titanic Dataset.* https://medium.com/analytics-vidhya/random-forest-on-titanic-dataset-88327a014b4dhttps://medium.com/analytics-vidhya/random-forest-on-titanic-dataset-88327a014b4d. [Online; accessed 5-November-2023].

Responsible Finance (2023). *UK public think credit scores unfair as organisations call for reform of credit information market.* https://responsiblefinance.org.uk/2023/04/uk-public-think-credit-scores-unfair-as-organisations-call-for-reform-of-credit-information-market/. [Online; accessed 2-November-2023].

scikit learn (no date, c). *sklearn.metrics.f1_score.* https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.htmlhttps://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1$_{s}core.html$. [Online; accessed 29-October-2023].

— (no date, a). *sklearn.metrics.precision_score.* https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.htmlhttps://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision$_{s}core.html$. [Online; accessed 29-October-2023].

— (no date, b). *sklearn.metrics.recall_score.* https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.htmlhttps://scikit-learn.org/stable/modu
[Online; accessed 29-October-2023].

Tensorflow (no date). *Logistic regression for binary classification with Core APIs.* https://www.tensorflow.org/guide/core/logistic_regression_core#logistic_regression.htmlhttps://www.tensorflow.org/guide/core/logistic$_{r}egression_{c}orelogistic_{r}egressio$
[Online; accessed 1-November-2023].