Team 89 GitHub Repo

MGT 6203 Final Project Report
Team 89, Spring 2024

Analyzing the Probability of Loan Default for Home Credit

Kai Fung

Joyce Hu

Kevin Mach

Joncarlo Montenegro

Calvin TszFung Ng

Table of Contents

1 Project Overview

1.1 Choice of Topic

For many financial banking institutions, a client's credit score is essential in determining whether they qualify for a loan or not. However, there is a portion of the population that have an insufficient or non-existent credit history. As a result, they struggle to qualify for loans at traditional financial banking institutions.

One primary lender for the unbanked population, Home Credit, is looking into how they can use machine learning techniques on alternative data points to predict whether these individuals will default on their loans.

1.2 Business Justification and Objective

Utilizing machine learning techniques to make default predictions result in financial, marketing, and reputation gains for Home Credit. Financially, by reducing instances of default, we mitigate revenue loss and preserve the value of unactualized interest income, which bolsters our profitability in the long run. Marketing wise, we can tailor our marketing campaigns to resonate with high-value segments of the population by leveraging demographic data associated with successful loan repayment. Identifying reliable payees also enhances our organization's reputation and fosters trust among stakeholders by offering lower interest rates and favorable terms to creditworthy individuals.

1.3 Problem Statement

We will use Home Credit's data to determine if we can accurately predict the probability of an applicant defaulting on a loan without access to the most predictive variable, credit score. Our initial hypothesis includes believing that demographic variables, along with variables like loan amount and yearly income, will be the most predictive variables and the predictive model that will pick up the most signal is regularized logistic regression.

2 Exploratory Data Analysis

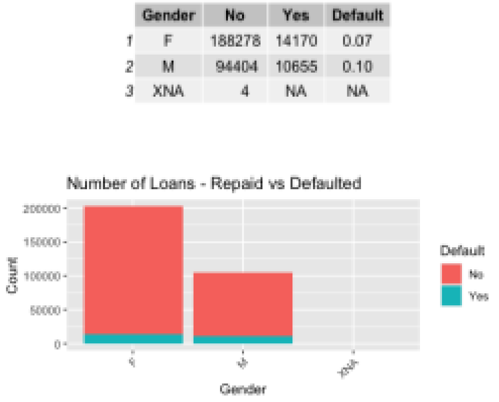
2.1 Combining the Data

Home Credit has generously given seven datasets relating to application data from current and previous loans that clients have applied for at Home Credit and at other institutions. One of the biggest challenges we faced was merging these datasets into one. For many of these datasets, there was a one-to-many relationship to the main (application) dataset. Due to these complications, we decided it was sufficient to use only the main dataset, which contained 124 predictor variables.

2.2 Exploratory Data Analysis

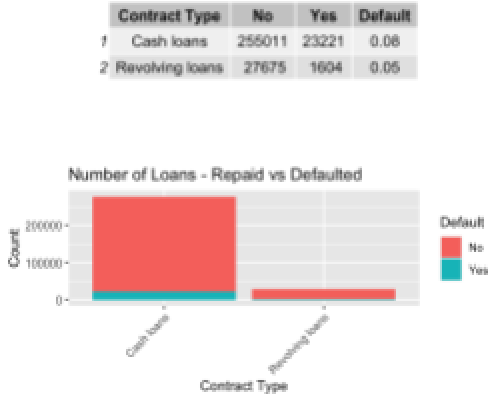
We conducted a univariate analysis on the categorical variables first to get a basic understanding of loan default behavior. Below are plots for the following: gender, loan type, own real estate flag, and education level. Based on Figure 1a, we see that females take out almost double the amount of loans compared to men, and are less prone to defaulting on their loans. In Figure 1b, we also observe that the majority of loans are cash loans, which makes intuitive sense considering revolving loans consist of items like credit cards and Home Credit's clients tend to have insufficient or non-existent credit history. A very interesting observation shown in the own real estate flag plot in Figure 1c is that around two-thirds of the clients own real estate, but there isn't a significant difference in the default rate between the two levels. Although it is surprising to see so many clients own homes but have insufficient credit history, the fact that the default rate between the two groups is equivalent leads us to believe that this variable will not be a predictor in loan default rate. The last categorical variable shown is education level (Figure 1d) and the insights drawn are extremely intuitive; the higher the education level of our client, the lower the default rate. As expected, this will be a crucial predictor in determining loan default rate.

Figure 1a



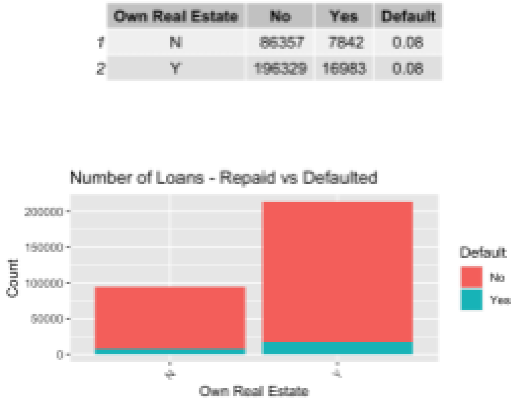
	Gender	No	Yes	Default
1	F	188278	14170	0.07
2	M	94404	10655	0.10
3	XNA	4	NA	NA

Figure 1b



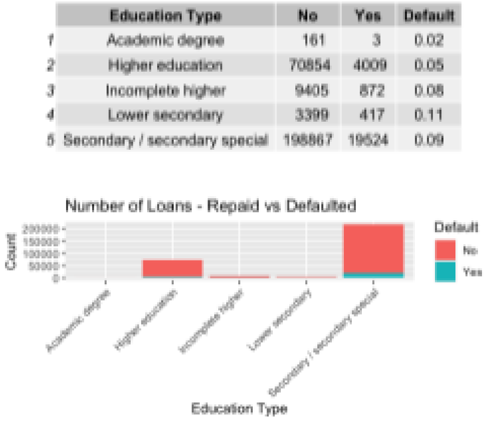
	Contract Type	No	Yes	Default
1	Cash loans	255011	23221	0.08
2	Revolving loans	27675	1604	0.05

Figure 1c



	Own Real Estate	No	Yes	Default
1	N	86357	7842	0.08
2	Y	196329	16983	0.08

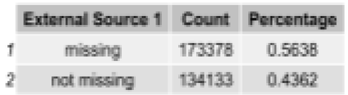
Figure 1d



	Education Type	No	Yes	Default
1	Academic degree	161	3	0.02
2	Higher education	70854	4009	0.05
3	Incomplete higher	9405	872	0.08
4	Lower secondary	3399	417	0.11
5	Secondary / secondary special	198867	19524	0.09

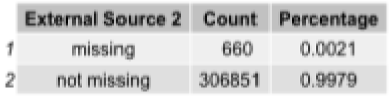
Based on the univariate analysis on the continuous variables for external source 1 and external source 2, we are seeing very different probability density functions between the two classes. We aren't given much context about these external sources other than the scores being normalized. For external source 1 in Figure 2a, even with 56% of the data missing, we see a significant difference between the two classes, which indicates that this variable is likely to be extremely predictive.

Figure 2a

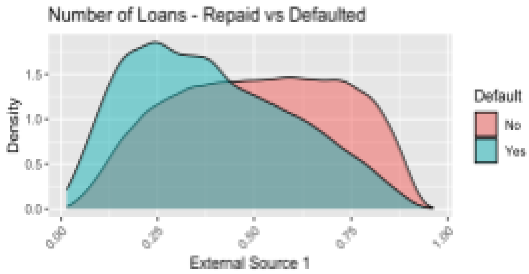


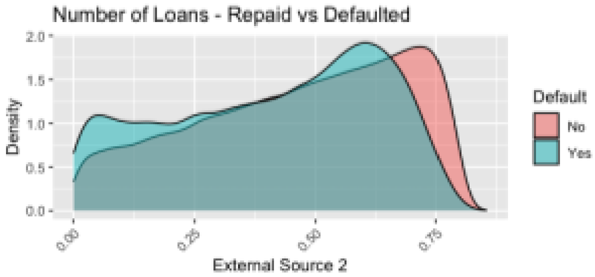
	External Source 1	Count	Percentage
1	missing	173378	0.5638
2	not missing	134133	0.4362

Figure 2b



	External Source 2	Count	Percentage
1	missing	660	0.0021
2	not missing	306851	0.9979





When binning the number of days employed variable, we noticed there were some values with ‘365243’ which equals 999.9808 years employed. This didn’t make any sense, so digging further we found that of those values 22 were ‘unemployed’ and 55,352 were ‘pensioners.' After binning by years, we noticed there to be a high correlation between lower years of employment with propensity to defaulting (Figure 3a). When looking at occupation type, we observe there to be 90,113 NAs (there are 22 unemployed, 55,352 pensioners, 3,787 state servants, and 24,920 working) with a baseline default rate of 0.07. We can see a pattern in the types of work correlated with default rate - more “white collar” occupations with a lower default rate and “blue collar” occupations with a higher default rate. When binning incomes, we see that the rate of default is about the same from 0-100k and 150k-250k but we do observe a slightly higher default rate in individuals at 100k-150k, possibly due to the fact that this income bracket has the highest number of applicants (Figure 3b). We do see that applicants 250k+ onwards show a lower default rate.

Figure 3a

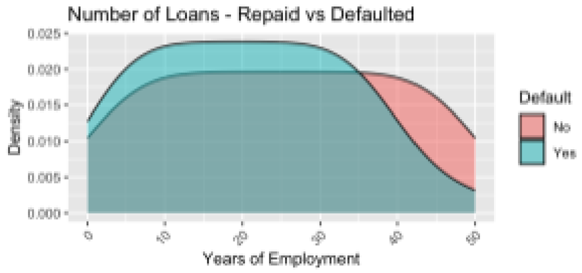
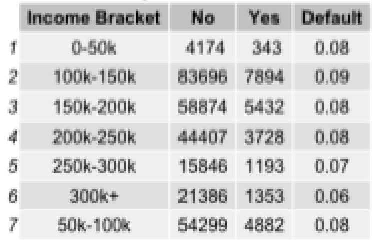


Figure 3b

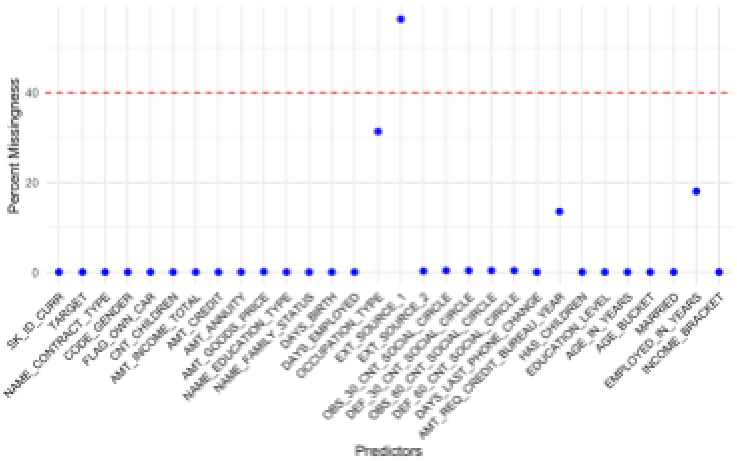


	Income Bracket	No	Yes	Default
1	0-50k	4174	343	0.08
2	100k-150k	83696	7894	0.09
3	150k-200k	58874	5432	0.08
4	200k-250k	44407	3728	0.08
5	250k-300k	15846	1193	0.07
6	300k+	21386	1353	0.06
7	50k-100k	54299	4882	0.08

2.3 Cleaning Data

Figure 4

Percent Missing in Columns



A sizable portion of the variables in our dataset contained missing values, and we decided to remove variables with over 40% missingness, as these would be difficult to salvage even with imputation (Figure 4). We next looked at a series of variables starting with “FLAG_”, which contained generic information about whether a client provided email/documents. We ran a tetrachoric correlation on our response variable against every flag and found no meaningful relationship, thus adding these flags to the chopping block (Juras, 2006, p. 1). The rest of the variables were manually appraised via the documentation file of column attributes until we agreed on the variables to remain. For example, variables containing information such as number of elevators in a building were deemed not particularly useful and then omitted.

As the imputation of any missing observations on variables we intended to keep would benefit greatly from deducing its imputed value from observations containing similar characteristics, the application of hot-deck imputation proved a natural choice, as hot-deck imputation allows for imputation to pull from “donors,” whose characteristics are left to our choosing (Myers, 2011, p. 297). We selected key demographic variables (age/income/gender/married/education/parental status) to serve as donor variables. Four observations with missing gender values were omitted due to our lack of ability to reasonably impute these, and the “mice” package successfully imputed the rest of our variables with the exception of occupation type, a categorical variable. Properly imputing this would require access to the Census’ NIOCCS system, which is beyond the scope of a student project.

2.4 Feature Engineering

In order to analyze the interaction between several continuous numeric variables relating directly to the amount of credit offered to an applicant, we utilized feature engineering to quantify these relationships. In particular, we wanted to quantify how the amount of credit related to the applicant’s income, the annuity of the loan, and the total price of goods purchased with said loan. Three variables were constructed: CREDIT_TO_INCOME_RATIO, CREDIT_TO_ANNUITY_RATIO, and CREDIT_TO_GOODS_PRICE_RATIO, which are made by dividing the credit amount by applicant income, loan annuity, and goods price respectively. We figured that these constructs may better predict the probability of repayment more than any of the untouched raw inputs.

2.5 Data Transformation

Certain variables present in our application data were stored in a manner less conducive to predictive modeling and its subsequent interpretation of analysis. These required conversion into more intuitive categories. We converted days since birth into the standard age in years, days employed into years employed, the number of children into a binary HAS_CHILDREN flag, and education attainment categories into three easy to interpret categories of incomplete highschool education, complete highschool education, and complete college education. We initially collapsed various family statuses into a binary MARRIED flag, but subsequent exploratory data analysis revealed differences in the distributions of loan repayment in the various family statuses, leading us to remove the binary marriage flag entirely in favor of the original categories. We made sure to avoid including any of the original variables with their collapsed constructs within a given model, as failure to do so would ultimately lead to multicollinearity. For example, we would test a model with AGE_IN_YEARS vs a model with a further collapsed age group variable AGE_BUCKET (a pretty standard grouping: "18-25", "26-45", "46-64", "65+"), in order to assess which performed better.

2.6 Variable Correlations

A critical aspect of our analysis involved examining the correlations between various variables. Understanding these relationships aids in constructing a predictive model by highlighting influential factors affecting the probability of loan default.

2.6.1 Categorical To Categorical

We utilized Cramér’s V statistic to measure the association between categorical variables. This approach allowed us to pinpoint significant correlations that could impact loan default rates. For example, our analysis revealed a strong correlation between gender and occupation type (Cramer’s V = 0.405). This intriguing finding suggested that occupation-related credit risk might not be gender-neutral.

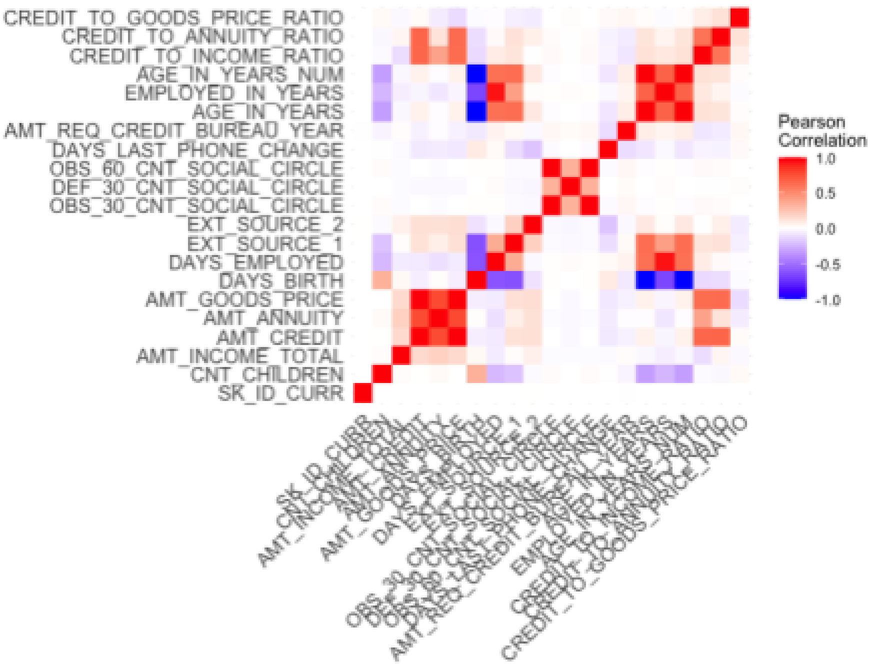
The moderate Cramér's V value of 0.421 between occupation type and education level uncovered in our dataset suggests an engaging narrative: individuals’ careers are somewhat intertwined with their educational backgrounds. This connection goes beyond a simple correlation; it implies that certain jobs might require specific educational attainments, or that people with certain degrees are attracted to certain fields. This insight is crucial, as it might reflect on an individual’s earning capacity and job security, which are direct indicators of their ability to service debts. Consequently, this moderate yet significant association between occupation and education demands attention in our predictive modeling, potentially serving as a beacon for assessing credit risk more accurately.

2.6.2 Continuous To Continuous

For the continuous variables, it felt like staring into a kaleidoscope as we observed the varying degrees through our heatmap. The resulting heatmap in Figure 5 illustrated the interdependencies. The column AMT_GOODS_PRICE and AMT_CREDIT are highly correlated, suggesting that as credit amounts increase, so does the value of the goods for which the credit is taken.

Figure 5

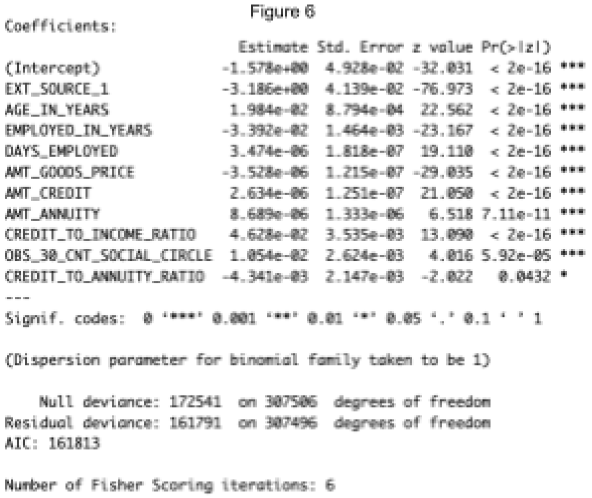
Correlation for Continuous variables



We focused on variables that exhibited high intercorrelations to understand their collective impact on the binary response variable. Using a correlation matrix as a guide, we identified variables with Pearson correlation coefficients above a specific threshold. This informed our selection of variables for inclusion in a logistic regression model using backward selection, forward selection, and a combined stepwise approach to identify significant predictors of the binary response variable. The models converged on the same set of predictors with the high correlations of continuous variables, indicating robustness in the variable selection process.

All predictors in the final model shown in Figure 6 are statistically significant, with p-values well below the conventional alpha level of 0.05, indicated by the significance codes. This suggests strong evidence against the null hypothesis for these coefficients, pointing to a meaningful association between each predictor and the response variable. The AIC of 161,813 provides a measure of the model's quality, balancing fit and complexity. When comparing models, lower AIC values are generally preferable.

Figure 6



```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.578e+00  4.928e-02 -32.031  < 2e-16 ***
EXT_SOURCE_1              -3.186e+00  4.139e-02 -76.973  < 2e-16 ***
AGE_IN_YEARS               1.984e-02  8.794e-04  22.562  < 2e-16 ***
EMPLOYED_IN_YEARS         -3.392e-02  1.464e-03 -23.167  < 2e-16 ***
DAYS_EMPLOYED              3.474e-06  1.818e-07  19.110  < 2e-16 ***
AMT_GOODS_PRICE           -3.528e-06  1.215e-07 -29.035  < 2e-16 ***
AMT_CREDIT                 2.634e-06  1.251e-07  21.050  < 2e-16 ***
AMT_ANNUITY                8.689e-06  1.333e-06   6.518 7.11e-11 ***
CREDIT_TO_INCOME_RATIO     4.628e-02  3.535e-03  13.090  < 2e-16 ***
OBS_30_CNT_SOCIAL_CIRCLE  1.054e-02  2.624e-03   4.016 5.92e-05 ***
CREDIT_TO_ANNUITY_RATIO   -4.341e-03  2.147e-03  -2.022   0.0432 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 172541  on 307506  degrees of freedom
Residual deviance: 161791  on 307496  degrees of freedom
AIC: 161813

Number of Fisher Scoring iterations: 6
```

2.6.3 Categorical To Continuous

We conducted ANOVA tests to explore how continuous variables vary across different categories. One revelation came when examining the connection between HAS_CHILDEN and AMT_CREDIT. Contrary to expectations, the p-value was a high 0.87108, suggesting that the presence of children did not sway amount of credit significantly. It seems that when it comes to credit amounts, parental status takes a back seat.

We also apply it with our response variable and other variables together. When we looked at the pairing of TARGET with AMT_INCOME_TOTAL, we found a promising p-value of 0.027237. This low p-value suggests the possibility of a subtle yet significant relationship between income totals and loan defaulting – a lead worth investigating further.

3 Modeling Methodology

3.1 Methodology

Home Credit's goal to figure out how to classify whether a loan applicant is likely to default or not is a binary classification problem. We employed several predictive modeling techniques to solve this binary classification problem: regularized logistic regression, random forest, K-nearest neighbors (KNN), gradient boosting machines (GBM), probit regression, and support vector machines (SVM). Unfortunately, two of these predictive models did not work out; probit regression is only good for normally distributed data and the SVM model was too computationally expensive to complete.

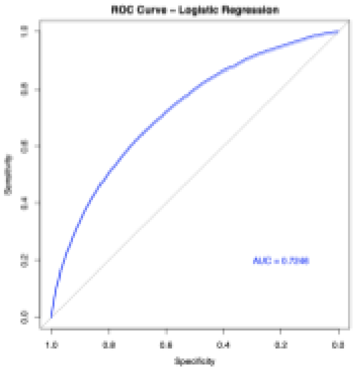
The data was split into 80% for our train and validate dataset and 20% for our test dataset. To measure which model performed the best on the test dataset, we evaluated the model's Area under the Receiver Operating Characteristic Curve (AUROC). The reason why AUROC was our selected performance metric rather than a metric like accuracy is because Home Credit's response variable, TARGET, has a class imbalance (Jeni, 2013, p. 1). In this case, we want a performance metric that is more sensitive towards the minority class, those more likely to default on their loans, because misclassifying these clients will be much more financially costly to Home Credit. In contrast, a metric like accuracy is not sensitive towards the minority class.

3.2 Modeling

3.2.1 Regularized Logistic Regression

The logistic regression model was built with the predictor variables that remained after our exploratory data analysis and stepwise regression research. As mentioned in 2.5 Data Transformation, there were still some variables that were heavily correlated with each other due to being transformed versions of their original variable. To test which variable to keep, we would keep all other predictor variables constant within the model and only change the variable we wanted to test. To determine which variable to keep, we inspected which model returned the highest AUROC in the k-fold cross validation dataset, with k = 5. This process of elimination procedure was repeated for multiple correlated variables, and the result was a final model with twenty-two predictor variables. The final selected variables consisted of some original variables, some transformed variables, and some feature engineered variables.

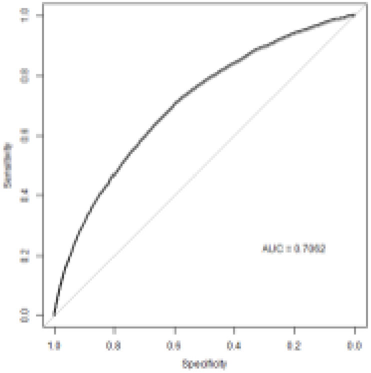
Figure 7



The hyperparameters α and λ were optimized by allowing each hyperparameter to test five different inputs. As a result, there were twenty five different hyperparameter combinations tested, with the model returning the highest train AUROC with $\alpha = 0.1$ and $\lambda = 0.0001973627$. This model was then used on the test data and returned an AUROC = 0.7246, as shown in Figure 7. The process of elimination procedure used to deal with highly correlated variables also proved to be effective as our final logistic regression model did not have any predictor variables regularized out.

3.2.2 Random Forest

Figure 8 - Forest ROC Curve

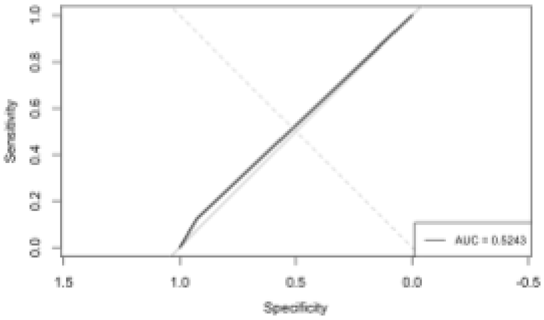


We attempted to model our response variable against the various predictors we decided to keep, using the “randomForest” package. In order to address some of the class imbalance in the response variable between those who had difficulties repaying the loan (a very small proportion in the data) vs those who had no payment difficulties (the overwhelming majority), we weighted them differently. Of the weights tested in our random forest model, a 20 to 1 ratio seemed to perform the best, although the extreme run-time prohibited us from testing more than a few weighting schemes. This selected model included 54,469 nodes and 500 trees. This model resulted in an AUROC of 0.7062 (Figure 8) and overwhelmingly predicted values of “no payment difficulties” in the test data, akin to all other models. Our forest AUROC score indicates a slightly worse prediction power than the logistic model. The run-time also proved to be an issue when attempting to apply the “steprf: Stepwise Predictive Variable Selection for Random Forest” package in R. Based upon our investigation of this package, a proper stepwise implementation could run far in excess of 30 hours.

3.2.3 K-Nearest Neighbors (KNN)

When tuning the *k* parameter, smaller values of *k* (e.g., < 5) can lead to higher variance in the performance estimate because the evaluation is based on fewer data points. Conversely, larger values of *k* (e.g., > 10) can lead to higher bias in the estimate because each fold contains a smaller portion of the data. We experimented with values of *k* ranging from 2 to 10. We also attempted oversampling the minority class and undersampling the majority class to achieve a more “balanced” dataset, however according to Blagus et al. (Blagus, 2015), the risk of overestimating the predictive accuracy is greater when oversampling techniques are used, and so we quickly abandoned this method. We then calculated the AUROC values for each model and found that *k* = 5 yielded the best AUROC value (0.5249). However, an AUROC of 0.5 is equivalent to random guessing, so this result raised concerns about the model’s performance. When we tested the KNN model with *k* = 5 on the test dataset, we obtained an AURUC of 0.5243 (Figure 9). One significant limitation we encountered with KNN was the inability to apply weights to the specific class of false negatives, which we were particularly interested in avoiding. This made KNN a less robust model compared to the other predictive models we developed.

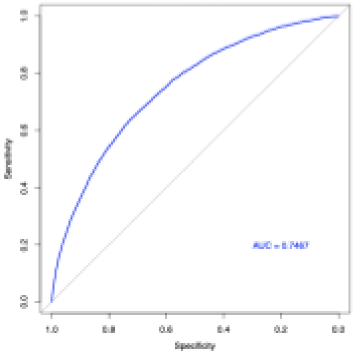
Figure 9
ROC Curve k: 5



3.2.4 Gradient Boosting Machines (GBM)

For gradient boosting machines, we also employed the same formula used for regularized linear regression. To select the hyperparameters, we also used the same logic applied to our regularized linear regression model. Two hyperparameters, learning rate and the minimum number of observations required to create a terminal node (leaf) in each tree, were kept constant at 0.1 and 10, respectively. The remaining two hyperparameters, the number of trees and interaction depth, were given five different levels to test. As a result, there were twenty five different hyperparameter combinations tested, with the model returning the highest train AUROC having number of trees = 250 and interaction depth = 5. With the optimal interaction depth being greater than one, this implies that our GBM found some meaningful interactions between our predictor variables. This model was then used on the test data and returned an AUROC = 0.7467, as shown in Figure 10.

Figure 10
ROC Curve – GBM



4 Results

Based on the test AUROC metric returned from all models, it was easy to eliminate KNN from consideration. We also decided against the random forest model due to not only the high computational resources it requires but also its AUROC was noticeably lower than both the regularized logistic regression and GBM models.

Although the GBM returned the highest test AUROC, we selected the regularized logistic regression model as the predictive model that Home Credit should implement to detect the probability of a client defaulting. The main reason why the regularized logistic regression model was selected over the GBM is because it is more explainable. Being more explainable means that it is easier to get the approval of both internal and external stakeholders, therefore the model can be implemented and pushed into production. With Home Credit being in the financial industry, there are regulatory concerns that the company has to take into consideration when deciding which model to select. Logistic regression models are easier to interpret and explain and much more widely accepted throughout many industries.

5 Conclusions

We began this project aiming to better understand the various relationships that impact credit payment difficulties, and in that regard, we succeeded. Our investigation into this particular domain of credit loan displayed a fundamental characteristic: the overwhelming majority of applicants in the data paid off their loans without difficulty. Within any model we tested, our predicted probability of defaulting never exceeded 25% for any given individual (and even these are very rare). Our findings indicate that Home Credit can mitigate their financial losses by implementing and using the regularized logistic regression model to determine whether a new client is likely to default on their loan payments. Home Credit can generously offer loans with the statistical knowledge that the overwhelming majority will be paid without issue. A promising area of interest to our aim of identifying loan default rates would be to impute an applicant's occupation status when it's missing from the data. Imputing occupation status can be quite difficult for many models to assess, and our dataset contained a large proportion of missings, which we converted to values of "Unknown." Successful imputation or upcoding of these cases via a method such as the NIOSH Industry and Occupation Computerized Coding System (NIOCCS) would improve model accuracy and predictive power. Given more time, we would have liked to explore combining the supplementary datasets using a one-to-one relationship to see if any of those variables add any additional predictive power to the machine learning models we used. Further investigation within the industry is of course necessary, and the members of our group may very well revisit this topic in the form of a future career.

6 Citations: (APA 6th Edition)

1. Blagus, R., Lusa, L. Joint use of **over- and under-sampling techniques** and cross-validation for the development and assessment of prediction models. BMC Bioinformatics 16, 363 (2015). https://doi.org/10.1186/s12859-015-0784-9
2. Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing **imbalanced data**--recommendations for the use of performance metrics. IEEE: 245-251. https://laszlojeni.com/pub/articles/Jeni13ACII.pdf
3. Juras, J. i Pasarić, Z. (2006). Application of **tetrachoric and polychoric correlation** coefficients to forecast verification. Geofizika, 23 (1), 59-82. Preuzeto s https://hrcak.srce.hr/4211
4. Myers, T. A. (2011). Goodbye, Listwise Deletion: Presenting **Hot Deck Imputation** as an Easy and Effective Tool for Handling Missing Data. Communication Methods and Measures, 5(4), 297-310. https://doi.org/10.1080/19312458.2011.624490