Final Report: Home Credit Default Risk

Daniel Alvarez, Libby Brill, James Irwin

Introduction

Our primary goal for this project was to help Home Credit, an international consumer finance provider, predict whether or not a loan applicant would repay their loan. Being able to determine default risk is crucial because inaccurate decisions may lead to significant harm not only to the company but also to prospective borrowers. Granting money to applicants who are more likely to default results in the company losing revenue, while denying money to someone who would've successfully repaid a loan is detrimental to customer relationships and can negatively impact the company's reputation. Given the complex relationship and numerous factors that go into default risk, predicting whether or not a loan applicant will repay their loan poses a complex and multivariate predictive problem. To tackle this problem, our team built a variety of classification models: Logistic Regression, Linear Discriminant Analysis (LDA), and Support Vector Classification (SVC), using Home Credit's database. Then, to determine the predictive power of each model, we considered metrics such as precision, recall, and ROC-AUC to assess model performance.

Data Overview

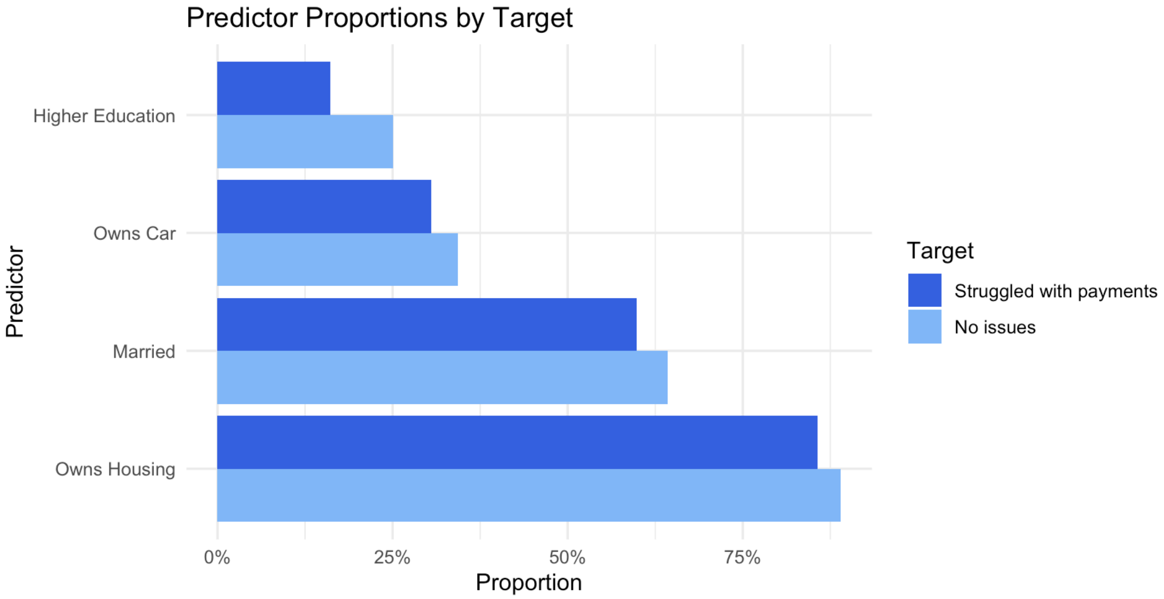
We used a variety of data sources throughout this project, with our main efforts focused on data about Home Credit's applications, various information on completed loan applications and the applicants themselves. Each row in this dataset is a single loan. Examples of features in this data set include the income and education levels of the applicant, the number of children the applicant has, and whether the loan is cash or revolving. Additionally, we incorporated information from bureau data, where each row corresponds to a single month of history for a past credit line. This file provided an in-depth view of the applicant's money-borrowing history, repayment patterns, and credit use over time.

Feature Engineering

To supplement the application and bureau data, we engineered three new features. These features are the number of supplementary documents that the client provided, the number of previous loans to the client that were reported to the credit bureau, and of those previous loans, the average number of days before the deadline that application was submitted.

An Initial Look

The following chart illustrates the proportion of certain demographics from the application dataset, broken down by the outcome of their loan repayment.



From the plot, the distributions of the attributes across outcomes are fairly similar, with only slight differences. For example, about 25% of loan applicants who have had no issues repaying their loans in the past have completed higher education, compared with only 15% of those who struggled with payments. However, we must note that these differences are small, so we can not infer predictive power from these variables alone.

Goals and Metrics

Throughout the course of this project, we wanted to ensure that our work aligned with Home Credit's company goals:

- Serving the unbanked: empowering people to enter the broader global economy
- Financial inclusion: focusing on delivering greater access to financial services through financial inclusion efforts across our markets
- Global partnerships: working with dynamic e-commerce and manufacturing partners across our markets
- Enabling growth: we have been offering affordable, accessible financial products and services for decades

Keeping these values in mind was especially important when it came to choosing metrics on which to optimize our models. The obvious choice was precision: of the people we predicted would pay back their loan, how many actually did? This metric makes sense because we didn't want Home Credit to lose money on defaulted loans. However, we also decided to focus on recall: of the people that will pay back their loan, what proportion did our model find? Using this metric aligned with the company values of serving the unbanked and financial inclusion - we wanted our model to successfully identify a high proportion of all people who will pay back their loans. Therefore, we chose to use precision as a satisfying metric (meaning our model needed to reach a certain threshold of precision, 0.6 in this case), and recall as an optimizing metric (meaning of the models that met that precision threshold, we chose the ones with the highest recall).

Splits and Balancing Data

To determine an appropriate train/validation/test strategy for our loan-repayment prediction model, we evaluated several potential data-splitting approaches. We considered a standard random split, a stratified split based on the target variable, and a few non-random splits designed to mimic potential future data distributions (training on older loans and testing on newer loans, training on medium and high-income borrowers while testing on low-income borrowers, etc.).

After comparing these options, we selected a stratified split on the target variable, allocating 60% of the data for training, 20% for validation, and 20% for testing. A stratified split was particularly appropriate for our data because the target outcome was highly imbalanced: only 8% of loans had repayment issues. Stratification ensured that each subset contained the same proportion of positive and negative cases as the full dataset, which helped guarantee that the validation and test sets included enough minority-class observations for reliable evaluation. To further address the imbalance during model training, we also applied 50/50 sampling within the training set so the model could learn to recognize the minority class more effectively.

Model Selection

Once we determined the correct split for the data, we decided upon a model selection process. First, we fit and trained numerous models of each type (Logistic, LDA, and SVC) on the training data. Then, for each model type, we selected the model fit that had the highest recall and also met the satisficing precision metric. At this point, then, we had three models: the Logistic, LDA, and SVC models that performed best on the training data. Next, we evaluated each of these models on the validation set and chose the model that had the highest recall as our final selection. This model was a logistic regression model. Finally, we evaluated this logistic model on the test set and reported these metrics.

It is important to note that, because no SVC model met the satisficing precision metric, we did not evaluate a SVC model on the validation set. Therefore, we only considered the performance of Logistic and LDA models on the validation set.

Preliminary Model

Our initial goal was to identify the best-performing model for predicting whether a loan would be repaid without issues. After experimenting with multiple model specifications, our strongest performing model was a logistic regression with a set of carefully selected predictors. These predictors were:

- Owning a car
- Owning a house/flat
- Credit amount of loan
- Loan annuity
- Price of goods (for consumer loans)
- Highest education level
- Married/single status
- Housing type
- Days employed
- Documents flagged
- Number of credits
- Mean days credit

On the test set, this model achieved a precision of 0.95 and a recall of 0.57.

Fairness Metrics

Throughout model development, we intentionally avoided including protected characteristics (such as gender or age) as predictors. However, we know that correlated features can act as proxies, introducing unintended bias. To assess this risk, we evaluated three fairness metrics: Demographic Parity, Equal Opportunity, and False Positive Rate Parity.

Demographic Parity measures whether different demographic groups were predicted to pay back loans successfully at similar rates. If one group is systematically approved at a higher rate, this may indicate inequity in access to credit. Equal Opportunity compares the recall across groups - how often the model correctly identifies people who have no payment issues. This ensures that we are not over or under-approving qualified borrowers. False Positive Rate Parity compares the rate at which the model incorrectly predicts “no issues” for borrowers who actually do have issues. Together, these metrics help identify biases that could create unequal opportunity, which is highly relevant to Home Credit with its goals centered on financial inclusion and serving unbanked or underserved populations.

Fairness Results for Preliminary Model

We are looking for our fairness metrics to be relatively similar for different levels of protected classes. Across the genders prevalent in our data, the initial model showed reasonably balanced performance.

Metric	Female	Male
Demographic Parity	0.5699	0.5257
Equal Opportunity	0.5836	0.5470
False Positive Rate Parity	0.3914	0.3314

Across age groups, however, the differences were substantial. Some metrics differed by up to 40 percentage points, indicating a largely contrasting impact.

Metric	< 35	35-45	45-55	55+
Demographic Parity	0.4068	0.5360	0.5411	0.7709
Equal Opportunity	0.4259	0.5513	0.5555	0.7766
False Positive Rate Parity	0.2504	0.6685	0.3649	0.3587

These results did not align with both legal expectations and our company values focused on equitable lending. We investigated the model predictors and determined that much of the age-related disparity stemmed from the “days employed” feature.

Final Model

We refit the model using the same predictors except for days employed:

- Owning a car
- Owning a house/flat
- Credit amount of loan
- Loan annuity
- Price of goods (for consumer loans)
- Highest education level
- Married/single status
- Housing type
- Documents flagged
- Number of credits
- Mean days credit

Removing this variable had minimal impact on performance: precision remained 0.95, and recall decreased only slightly (going from 0.57 to 0.56)

Fairness Results for Final Model

Gender fairness improved slightly (an unintended but positive outcome).

Metric	Female	Male
Demographic Parity	0.5490	0.5354
Equal Opportunity	0.5627	0.5556
False Positive Rate Parity	0.3704	0.3508

However, age fairness improved considerably. Although not perfect, the disparities were far smaller and more acceptable given business and ethical goals.

Metric	< 35	35-45	45-55	55+
Demographic Parity	0.4460	0.5867	0.5781	0.5774
Equal Opportunity	0.4656	0.6030	0.5936	0.5848
False Positive Rate Parity	0.2852	0.4043	0.4043	0.4452

The improvements in fairness (especially in age-related metrics) far outweighed the negligible 0.01 decrease in recall. As a result, we selected this updated logistic regression as our final model.

Interpreting the Final Model

The coefficients for several key features in the model provide insight into their varying influence on loan repayment success and risk. Adjusting for all other predictors, both owning a car ($\beta = -0.0596$) and owning a house/flat ($\beta = -0.0192$) are associated with a lower probability of having issues repaying the loan, although the effect for owning a house/flat is smaller than that of owning a car. This suggests that both car and realty ownership are a sign of financial stability. On the other hand, submitting one additional document (increasing the number of the documents flagged variable by one) is associated with a higher probability of having issues repaying the loan ($\beta = 0.1272$). Similarly, an increase in the number of previous credits is also associated with a higher probability of having issues repaying the loan ($\beta = 0.0281$), although with a smaller impact.

Takeaways

In our efforts to help Home Credit determine whether or not a loan applicant will default or not, we built three classification models: Logistic Regression, Linear Discriminant Analysis, and a Support Vector Classifier, with the intent of using metrics such as precision and recall to evaluate the predictive power and reliability of each. Upon trying a variety of splitting methods, we decided that a stratified split on the target variable was appropriate, provided the immense class imbalance between positive and negative target variable observations. Furthermore, undersampling the target class in the training set ensured that we saw an equal number for each class. After exhausting many possible combinations of models and predictors on our split (i.e., fitting all models on LDA, fitting only four predictors on SVC), we got our best LDA, our best Logistic Regression, and our best SVC model in terms of recall. Since no SVC model we fit met the satisficing precision score of 0.6, we moved on to our two best models: LDA and Logistic Regression. We chose to use a stratified split and a balanced training set, and experimenting with multiple specifications yielded our best model: a Logistic Regression fit with 12 predictors.

Although predictive power was important in our selection of the best model, a key aspect of our assessment of default risk was to employ fairness. This meant excluding protected classes like age and gender in our final model. To ensure our current "best" model was not unintentionally biased, we evaluated fairness metrics Demographic Parity, Equal Opportunity, and False Positive Rate Parity on gender and age in particular. Initially, our best model performed reasonably well across gender but not age. However, upon realizing that the number of days an applicant has been employed is indirectly tied to age, we omitted the variable as a predictor in our model, establishing it as a proxy variable while simultaneously improving age fairness. Our model's recall decreased by 0.01 percentage points. Notwithstanding this almost insignificant decrease, we established this as our best model, one that considers both performance metrics as well as fairness.