****

**November 6, 2023**

**Submitted By:**

Andrew Kerr - adkerr@calpoly.edu

Nathan Hill - nhill05@calpoly.edu

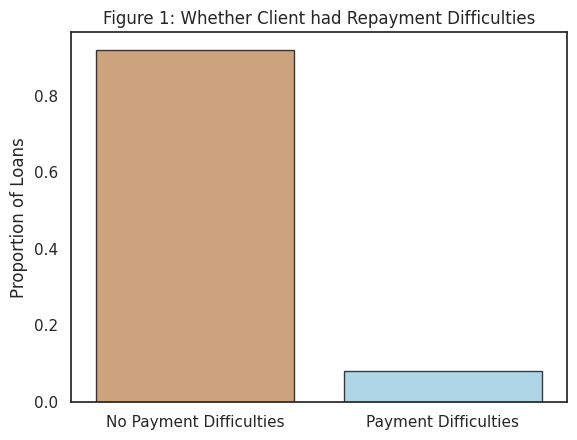
Jadyn Ellis - jellis13@calpoly.edu

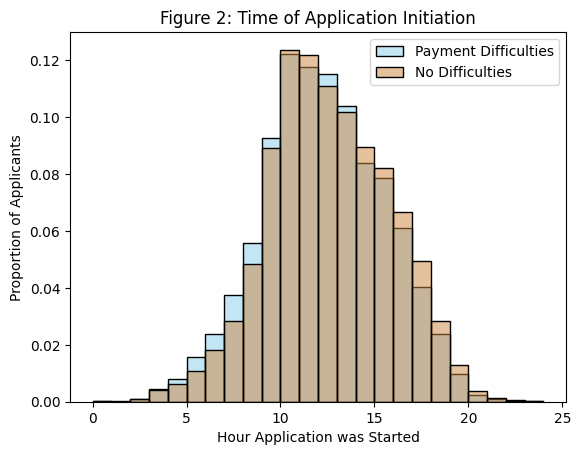
Jamie Luna - jluna28@calpoly.edu

Aditi Gajjar - agajjar@calpoly.edu

Home Credit is interested in predicting whether an applicant will repay a loan without difficulties based on information provided on their loan application. In order to achieve this goal, we will investigate and compare logistic regression models, support vector machines, and linear discriminants. These models will use various information provided on the loan application to classify the applicant as more likely to have payment difficulties or not to have difficulties. While we are devoted to creating the most accurate model we can, it is important to maintain a high standard of ethics and fairness in all of our processes.

**Data Collection**

To predict whether or not a client is able to repay a loan without difficulties, we will analyze data on 307,511 unique loans from past Home Credit clients[[1]](#footnote-0). As shown in Figure 1, initial investigation into this data shows that only 8% of clients had difficulty paying back their loan - only 8% of clients had a late payment more than X days on at least one of the first Y installments of the loan (as reported in the documentation). After thoroughly cleaning the collected data, we will create features for possible factors that could segregate the clients into three groups: those who will definitely repay without difficulty, those who might have difficulties, and those who will definitely have difficulties.

We are motivated to include features related to the client's job and income stream because we expect these to be strongly influential in predicting an applicant’s ability to pay back a loan. The employment and income feature information are strong indicators of an applicant’s financial resources. It is logical to conclude that applicants with more financial resources should have less difficulty repaying their mortgages. Furthermore, we theorize that the stability of their work and family life could reflect the applicant’s ability to make regular payments in a timely manner. 

Another important feature we are motivated to include is when the client submits their loan application. As shown to the left, in Figure 2, clients who submitted applications in the early morning tended to have difficulties repaying their loan, while clients who submitted applications in the afternoon and evening tended not to have difficulties. Thus, we believe the application submission time may reflect different types of clients.

**Modeling Process**

We plan to fit three different models: *logistic regression, support vector machines,* and *linear discriminant*. The goal of each of these three models is to determine whether or not a person is able to pay back a given loan. For each of our models, we will use the same subset of determined features in order to compare our given models. Specifically, we are interested in exploring different subsets of the following features:

| **Factor of Interest** | **Description** |
| --- | --- |
| Total Income | The amount of monthly total income, self-reported by the applicant |
| Total Amount of Loan | The total credit amount of the loan being applied for |
| Occupation Type | The broad category of work done by the applicant |
| Education Level | Whether the applicant has a college degree or not |
| Credit Bureau Enquiries | The total number of inquiries made about the applicant to the Credit Bureau |
| Gender | Gender of applicant |
| Commute | Whether or not the applicant lives in the same city they work in |
| Application Details | Time and day the applicant began their process |
| Marriage Status | The status of the applicant’s relationship |
| Organization Type | The type of industry the applicant is employed in |
| Days Employed | Number of days the applicant has been with current job |
| Loan Type | Whether the loan is for a lump-sum of cash or revolving credit |

*Table 1. Of-interest feature information*

Furthermore, we want to utilize the previous loan histories of the applicant. We have information regarding previous loans from Home Credit and other companies that could prove to be useful if there is a pattern in repayment history. Some of the factors we are interested in include previous rejections, payment methods, loan amounts, and when previous loans were taken out.

During our model selection process, we will utilize K-Fold Cross Validation to assess each model’s predictive classification capabilities on whether a person has had difficulty paying back a loan or not. Furthermore, we will use k-fold cross-validation to determine values of hyperparameters, such as lambda penalties for our logistic regression model. We will use different methods of splitting our train and test data, such as random split or stratified split, in order to investigate and determine the optimal splitting method for validation. During our validation process, we will compare several metrics, such as accuracy or F1 score.

**Project Outcomes**

Our investigative process will employ sophisticated approaches to model selection, with the end goal of creating the most accurate and informative set of models in order to determine the risk of providing a loan to given borrowers. We will be able to provide you with informative information on borrower capability of paying back a loan, characterized as *will pay back, will not pay back,* and *could go either way*. During our model selection process, we will apply advanced techniques in order to retrieve the optimal set of predictors to utilize in each of our models. Further, we will perform fairness explorations on a selected set of protected attributes to ensure our created model does not exhibit high levels of bias. Our resulting findings will be presented in an informative, spoken presentation as well as a detailed discussion in a written report. You will find professionally formatted visualizations such as charts, tables, and model output along with descriptive, in-context interpretations and model summaries to reach all members of your team, from analytics experts to executives.

Our model shall determine the most optimal set of features to use across our three model specifications in order to compare methodologies and employ different techniques to classify borrowers. Further, we will perform an exhaustive fairness investigation, so as to not employ uninformed model bias in the situation that our decisions can negatively impact a potential borrower. Once developed, our model shall accurately determine the risk of providing a loan to a borrower, given the values of our selected features. Our goal is to foster a trustworthy partnership with Home Credit and provide you with informed, validated, and tested models to aid you in your risk-discovery efforts, while also making certain you still have full autonomy in your final loan borrowing decisions. We hope you will consider us as your partner moving forward in your data-directed business decisions.

1. https://www.kaggle.com/competitions/home-credit-default-risk/overview [↑](#footnote-ref-0)