Home Credit Project Proposal

DATA 403

Daniel Alvarez, James Irwin, Libby Brill

Data Summary

We will use information from Home Credit, an international consumer finance provider based in Europe and Asia. Their "application_train.csv" table, which is part of the "application" dataset, is broken into a training and testing set. Our dataset consists of over 30,000 observations, each representing a loan that has been taken out. There is information about the loan and the loan applicant at the time it was filled, such as loan status, the applicant's occupation, education level, and the size of their living area. The value of the response variable, "TARGET", is a 1 if the client is known to have payment difficulties in the past, and a 0 represents any other case. To ensure Home Credit best predicts the ability of its clients to repay their loans, we will perform a number of tasks. For example, we plan to deal with missing values by encoding them as a 0 if appropriate, such as for the basement area or car age, as well as including a string if possible, like saying "unknown" for an applicant's occupation type. We also plan to create new features that can condense our data as well as provide meaningful insight, such as collapsing income into categories (low income, medium, high), age into groups (applicants ages 20-30, 30-45, etc), as well as a variable for the living area size and age. Lastly, some valuable predictors we plan to use are age and gender (though we must proceed with caution), the type of loan taken out, annuity, and whether the applicant owns property or not.

Models and Metrics

Models

We will be fitting three different types of models: logistic regression, support vector machines (SVM), and linear discriminant analysis (LDA). While these are all classification models, they each have unique strengths and perform best under different conditions, making it worthwhile to evaluate all three. Logistic regression provides a stable, interpretable baseline for comparison. SVMs offer flexibility for capturing complex, nonlinear relationships. LDA, on the other hand, can be highly efficient when its distributional assumptions are satisfied. Comparing their performance allows us to determine which approach is most effective for our data's structure and patterns.

Testing

For model evaluation, we considered several approaches to splitting the data into training, validation, and testing sets. A completely random split is straightforward to implement and works well when the data are balanced across classes. However, when the data are imbalanced, a stratified split where the proportion of each class is preserved in all subsets helps prevent bias and ensures more reliable evaluation metrics. In our case, because the dataset only includes individuals who have already received loans, a non-random split may provide a more realistic assessment. By selecting test and validation sets that better reflect the characteristics of future applicants rather than approved borrowers, we can obtain performance estimates that more accurately represent how the models will behave once deployed.

Metrics

Conclusions and Deliverables

Ultimately, we will produce for Home Credit a model that accurately and fairly predicts whether potential clients will be able to pay back their loans. Because Home Credit is dedicated to providing banking opportunities for the unbanked population, this model will be inclusive of people who may not traditionally receive loan assistance. Additionally, we will provide a discussion about the different factors that are associated with loan repayment, a detailed description of our model selection and validation process, and a report of our calculated metrics. Finally, we will evaluate the fairness of our model with respect to protected attributes like gender and age. We appreciate the opportunity to present our proposal and look forward to working with Home Credit.