Home Credit Default Prediction Analysis

CSCI-556 Spring 2022: Group-11-HCDR

❖ TEAM INTRODUCTION AND RESPONSIBILITIES

Team Member	Email
Agalya Velusamy	agvelu@iu.edu
Amit Banerjee	ambaner@iu.edu
Pravallika Pentapati	prpent@iu.edu
Wesley Martin	wemarti@iu.edu

Roles and Responsibilities

Agalya Velusamy
GitHub repository setup, project timeline, EDA/feature engineering, pipelines, video editing

Amit Banerjee
Pipelines, machine learning algorithms, models

Pravallika Pentapati
EDA, hyperparameter tuning, feature engineering, submitting phase assignments

Wesley Martin
Metrics and evaluation, general machine learning engineering, pipelines, presentation

❖ ABSTRACT

The Project that we are working as a Team (Group-11-HCDR) is called the Home Credit Default Report project.

The problem arises from many people struggling to get loans due to insufficient or non-existent credit histories. Unfortunately, this population is often taken advantage of by untrustworthy lenders. The goal of our project is to use existing data to learn and classify new applications into two classes signifying whether the applicant has the capability to pay the loan (class 0) or would they default (class 1).
To reach our goal, we will utilize existing data available on such loan applications, including previous application and balance data with Home Credit, loan application, and balance data from other bureaus and credit card balance data. We will perform EDA on the data to determine necessary imputations and feature engineering. Subsequently we would apply and test several estimation methods: Logistic regression, Support Vector Machines, etc., to select the best model.
We aim to get an accurate selector to predict the right classes for a loan application. This would help the population without credit history in securing loans and also help Home Credit judge the risk associated with such a customer.

❖ DATA DESCRIPTION AND OVERVIEW

The dataset is provided by Home Credit which is a service dedicated to providing loans to the unbanked population. For the applicant data, we aim to predict if they are capable of repaying a loan. Given we have two classes in the target label, we will approach this problem as a binary classification problem.

➢ Dataset Background

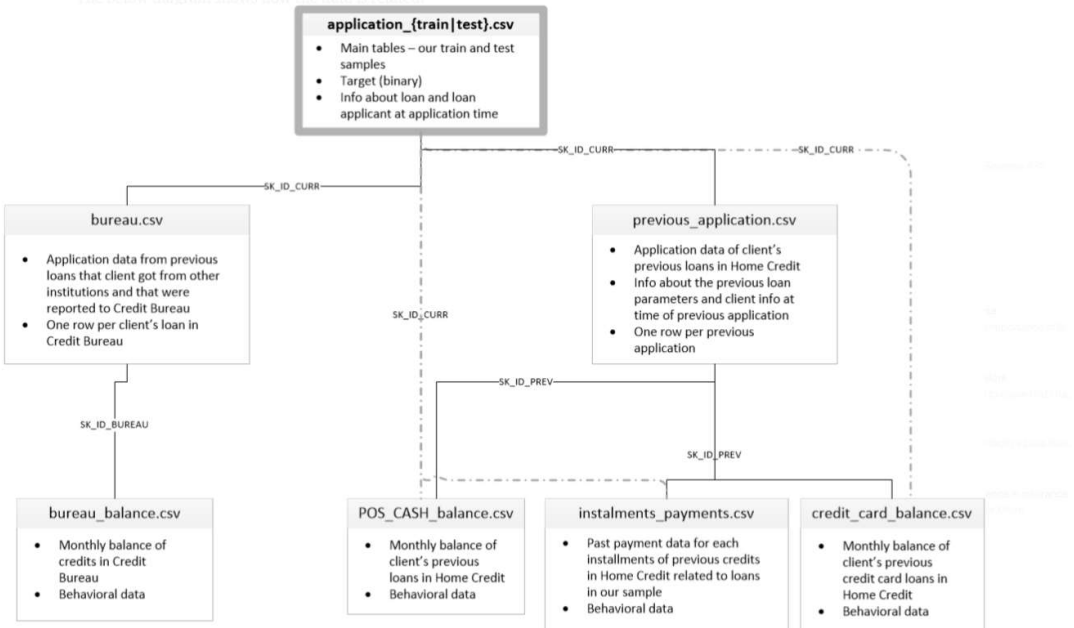
Founded in 1997 in the Czech Republic, Home Credit is a non-banking financial institution that operates in 14 countries including the United States, China, and India to name a few. The company focuses on lending primarily to borrowers with little to no credit history who would not obtain loans or potentially become victims of untrustworthy lenders.

Home Credit is currently using various statistical and machine learning methods to make predictions, Home Credit data scientists provided the dataset to Kaggle to help unlock the full potential of their data and ensure clients capable of repayment are not rejected for loans with a repayment calendar that will empower their clients to be successful.

There are seven different sources:

Dataset	Description
application_train/application_test	Main training data with information about each loan application at Home Credit. Every loan has its own row and is identified by the feature SK_ID_CURR. The training application data comes with the target indicating 0 or 1, the loan was repaid or not repaid respectively. The target variable defines if the client had payment difficulties indicating a late payment of more than X-days on at least one of the first Y installments of the loan. Such a case is marked as 1 while all other cases as 0.
bureau	Data concerning the client's previous credits from other financial institutions. Each previous credit has its own row in the bureau, but one loan in the application data can have multiple previous credits.

bureau_balance	Monthly data about the previous credits in the bureau. Each row is one month of a previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.
previous_application	Data of previous applications for loans at Home Credit of clients who have loans in the application data. Each current loan in the application data can have multiple previous loans. Each previous application has one row and is identified by the feature SK_ID_PREV.
POS_cash_balance	Monthly data about the previous point of sale or cash loans clients have had with Home Credit. Each row is one month of a previous point of sale or cash loan, and a single previous loan can have many rows.
credit_card_balance	Monthly data about previous credit cards clients have had with Home Credit. Each row is one month of a credit card balance, and a single credit card can have many rows.
installments_payment	Data of payment history for previous loans at Home Credit. There is one row for every made payment and one row for every missed payment



➢ Data Schema

➢ Data Dictionary

Homecredit_columns_description file contains descriptions for all columns in the various data files. please find the attached document for detailed information of columns

HomeCredit_columns_description.csv

➢ Main tables application train/test.csv information

application_train.csv	application_test.csv
122 columns with 307511 records	121 columns with 48744 records
Imbalanced target labels consisting of: 0 - (282686 total) 1 - (24825 total)	No target labels
Source for training models	Source for testing the performance of models

❖ MACHINE LEARNING ALGORITHMS USED

The team will use SciKit-Learn tools, along with NumPy and Pandas.

The home credit default risk prediction is a binary classification problem, as such, we plan to the following machine learning algorithms:

➢ Logistic Regression

This is a supervised machine learning classification algorithm generally used for a problem where the outcome is binary to predict probability by assigning observations to a discrete set of classes. We will use the algorithm and feature engineering to target the best model to predict if a customer can repay the loan. To measure the performance of this model we will use the Log Loss (CXE) function.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} -y_i log(h_\theta(x^{(i)})) - (1-y_i) log(1-h_\theta(x^{(i)})) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

➢ Support Vector Machines

SVM with stochastic gradient descent strives to attain a maximum margin hyperplane between positive and negative labels. Given the emphasis on boundary points, we will use the Hinge Loss function because we anticipate borrowers will be on the default risk boundary.

$$Loss_{SVM\_Hinge}(W) = \frac{1}{N}\sum_{i=1}^{N} Max(0, 1 - X_i^T W y_i)$$

➢ Random Forest

This will be implemented to improve the performance based on the results. We plan to start with the Decision Tree algorithm to produce a single decision tree with sequential outcomes based on the attributes. Afterward, we plan to extend to the Random Forest algorithm for multiple decision trees by randomly circulating the attributes.

➢ K-Nearest Neighbor

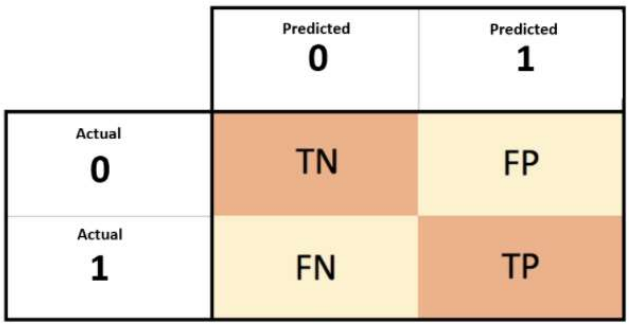
KNN classifier will be used to parse through the large dataset to partition the data into clusters and identify various features to determine whether borrowers belong to a repayment risk group.

❖ PERFORMANCE METRICS

Given the data is an imbalanced dataset, to measure accuracy, we will use the below metrics for evaluating model performance

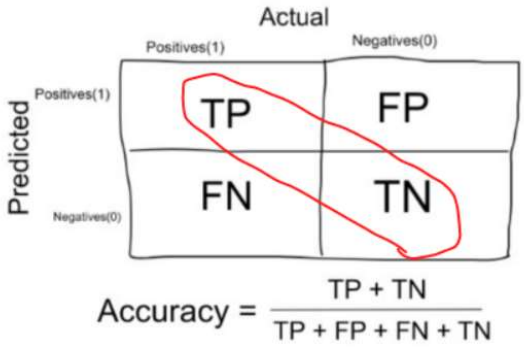
➢ Confusion Matrix

A confusion matrix will allow visualizing mistakes made by a model by detailing misclassifications for both classes. Using the confusion matrix, we will compute metrics to compare the classification performance. Below is a representation of confusion matrix:



➢ Accuracy Score

Accuracy is the number of correct predictions made by the model over all predictions.



➢ Precision Score

Precision measures how many predicted positives are actually positive.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

➢ Recall or Sensitivity Score

Recall calculates how many of the Actual Positives the model captures.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

➢ F-1 Score:

The F1 score measures the test accuracy by considering both the precision *p*, where *p* is the number of correct positive results divided by the number of all positive results; and the recall *r*, where *r* is the number of correct positive results divided by the number of all samples that should have been identified as positive.

The *F1* score is the harmonic average of the precision and recall, where an *F1* score reaches its best value at 1 (perfect precision and recall) and worst at 0.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

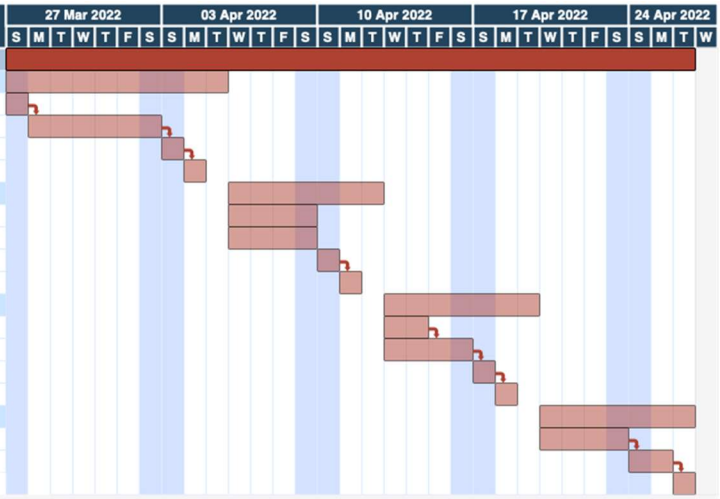
➢ ROC AUC Score

The Receiver Operating Characteristic Area under the curve (ROC AUC) is a metric that can be applied to imbalanced datasets since it does not generate 0 or 1 predictions, but rather a probability between 0 and 1.

It measures the false positive rate (x-axis) versus true positive rate (y-axis). The area under the curve (AUC) is the area between a model's ROC curve and the diagonal indicating a model with naive random guessing. If the ROC curve is more to the left/top of the diagonal, it indicates better performance with a higher ROC AUC.

❖ TIMELINE GANTT CHART

	Task	Duration	Start	ETA
1	Complete project execution	38 days	03/20/22	04/26/22
2	Phase 0	17 days	03/20/22	04/05/22
3	Kick-off meeting	1 day	03/20/22	03/20/22
4	Proposal requirements	14 days	03/21/22	04/02/22
5	Merge requirements	1 day	04/03/22	04/03/22
6	Submit proposal	1 day	04/04/22	04/04/22
7	Phase 1	7 days	04/06/22	04/12/22
8	EDA	4 days	04/06/22	04/09/22
9	Pipeline modeling	4 days	04/06/22	04/09/22
10	Prepare presentation	1 day	04/10/22	04/10/22
11	Submit presentation	1 day	04/11/22	04/11/22
12	Phase 2	7 days	04/13/22	04/19/22
13	Feature Engineering/Tuning	2 days	04/13/22	04/14/22
14	Pipeline modeling	4 days	04/14/22	04/16/22
14	Prepare presentation	1 day	04/17/22	04/17/22
15	Submit presentation	1 day	04/18/22	04/18/22
16	Phase 3	7 days	04/20/22	04/26/22
17	Data loss functions	4 days	04/20/22	04/23/22
18	Prepare final report	2 days	04/24/22	04/25/22
19	Submit final report	1 day	04/26/22	04/26/22



❖ MACHINE LEARNING PIPELINES USED

A sequence of data processing components and transformers, estimators and predictors is called a pipeline. Pipelines are a very useful tool since they combine the processing components and help in cross-validating across many different hyperparameters to test and select the best options. The key steps in our ML pipeline would be :

Data Transformation ->

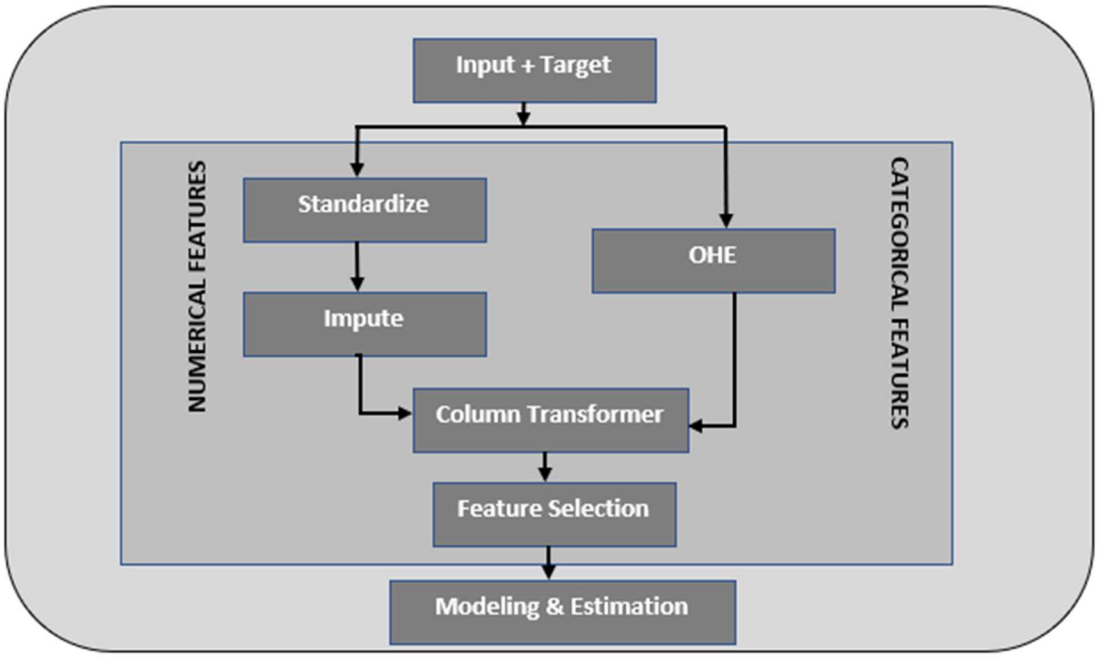
Standardize : This step would scale numeric data to make them into a uniform scale.
Impute : This step is a strategy for columns where data is missing and how to fill the missing values. We would pick a strategy like mean or median to impute missing values.
OHE : This step shows the One hot encoding of categorical data. In this step categorical data is split into as many columns as unique values and captured as binary flags.

Data Enhancement ->

Column Transformer : This step would enhance the data by adding additional derived features which contribute to the model.
Feature Selection : This step is where we remove features that do not contribute to the model and retain only contributing features.

Modeling and Estimation ->

This is the final step in our pipeline where an algorithm like Logistic Regression or SVM would be fitted to the data and then the model would be used to make predictions.



References:

https://www.kaggle.com/c/home-credit-default-risk/data

https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b