D. Chris Young
Machine Learning Engineer Nanodegree

Capstone Proposal

1. Domain Background

Consumer finance companies provide loan options to individuals to make every day purchases such as home appliances, personal computer or electronics. Credit risk is the risk that borrowers fail to make the required payments and the company recognizes a loss on the loan.[1] Credit policies and guidelines can establish minimum lending standards and models can be used to quantify the risk of default.

People often struggle to get loans due to insufficient or non-existent credit history but often are the very consumers that need financial assistance. Home Credit Group is a company that focuses on responsible lending primarily to people with little or no credit history. The company has posted the Home Credit Default Risk Kaggle competition to see if the community can help predict if a customer will have payment difficulties using their proprietary dataset.

I have spent the majority of my professional career working in the financial services industry and witnessed first-hand one of the worst real estate market crashes of all time in 2008. This project will utilize my domain expertise and machine learning skills to help Home Credit build a stronger model to identify credit worthy consumers.

2. Problem Statement

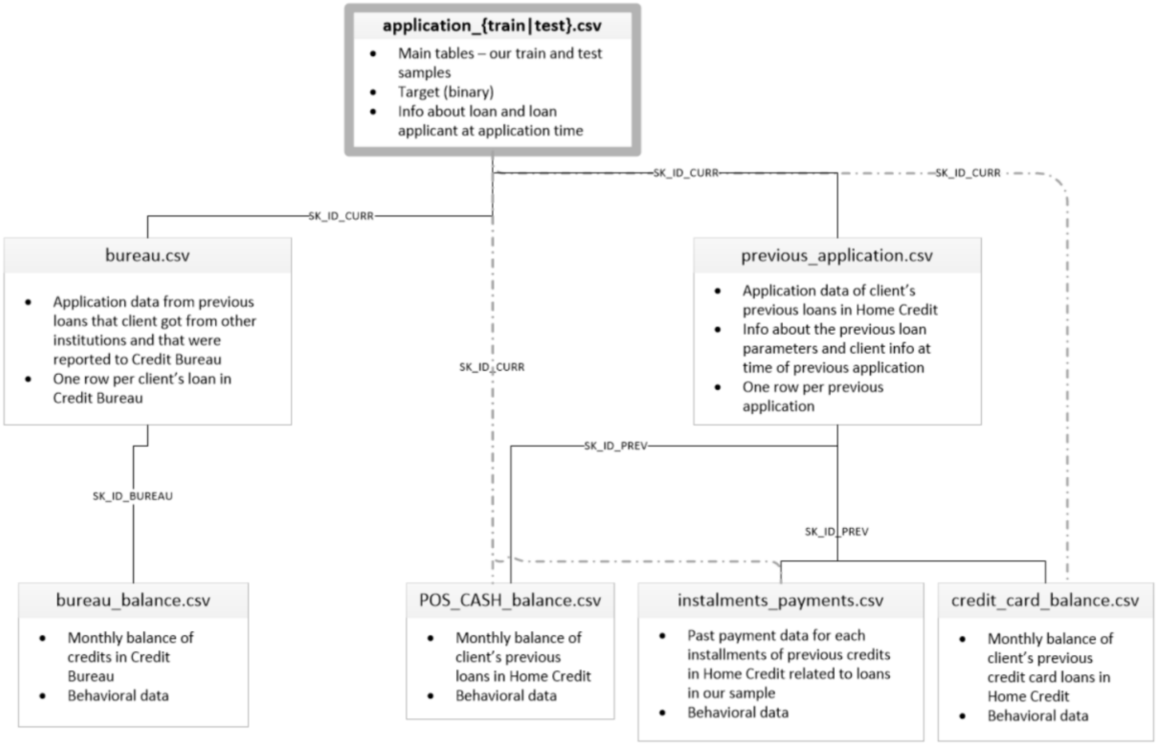
The problem that needs to be solved is reducing the likelihood of Home Credit rejecting consumers that are capable of repaying their loans. The inputs are general applicant statistics for current and prior loans with Home Credit; loan, credit card balance and payment history data for prior loans with Home Credit as well as credit bureau and balance data for loans from other lenders. This is a supervised learning task based on the binary classification on the target variable. The output is the predicted probability that the consumer will have payment difficulty.

3. Datasets and Inputs

The Home Credit Default Risk dataset for the competition can be found here. The dataset is unique because it includes actual application, balance, point of sale purchase and payment

[1] *Credit Risk Wikipedia*

statistics data for Home Credit customers. The training data is labeled with a binary variable Target and is the dependent variable in the machine learning task. The dataset also includes credit bureau application and balance statistics for loans customers have with other lenders. There are 221 columns across 7 files. Below is a snapshot of the data model.



application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

Figure 1: Home Credit Default Risk data model [2]

The dataset relates to the problem statement because demographic information, payment history, purchase patterns and credit balance features should provide insights into the characteristics of customers that have defaulted on their loans. These customers have higher credit risk and the types of customers Home Credit should not offer new loans.

4. Solution Statement

The solution for this problem is a supervised learning task. The classifier must be probabilistic that outputs a probability distribution for the predicted class. Some suitable supervised learning models for this problem are:

- Logistic Regression
- Decision Trees
- Random Forest

[2] Home Credit Default Risk Data

- Gradient Boosting

The predict_proba method returns the probability estimates for the prediction task. This output can be used to measure model performance and to create the competition submission file.

Models will be created for each of these classifiers and evaluated to determine the model that has the best initial results. The best model will be selected and tuned for final model evaluation and prediction.

5. Benchmark Models

A random predictor would have area under the curve (AUC) score of .5. This is represented by the diagonal dashed line in the plot below and provides the benchmark if a model is useful. A model that produces an AUC score of 1 means the predictions are perfect.[3]

The orange solid line in the plot below is the AUC score of 61.8% for a Decision Tree Classifier with parameters max_depth = 10 and max_features = sqrt. The model was created with minimum data preprocessing, no feature engineering and only basic application features were included. This establishes the benchmark for a model that is slightly better than random guessing.

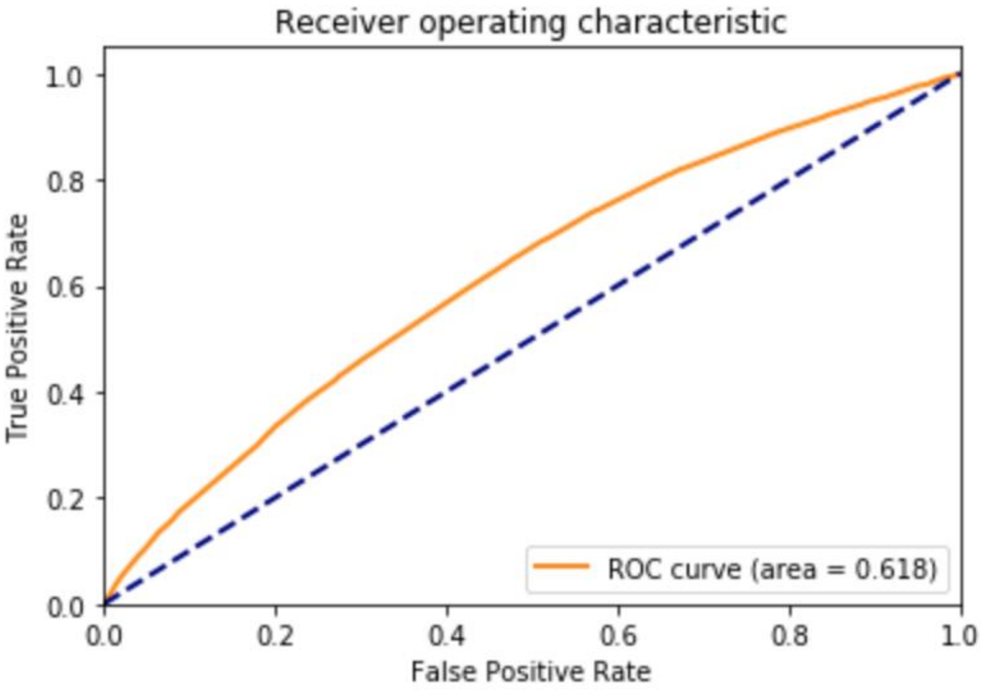


Figure 2: Receiver operating characteristic curve

[3] What does AUC stand for and what is it

6. Evaluation Metrics

Evaluation is based on the area under the receiver operating characteristic (ROC) curve between the predicted probability and the observed target. The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The higher the AUC the stronger the classification model is performing on the target variable.

The two key metrics used to calculate AUC is the True Positive Rate (TPR) and False Positive Rate (FPR) from a confusion matrix:

- True Negative (TN) is when a negative class is predicted, and the actual class is negative.
- False Negative (FN) is when a negative class is predicted, and the actual class is positive.
- False Positive (FP) is when a positive class is predicted, and the actual class is negative.
- True Positive (TP) is when a positive class is predicted, and the actual class is positive.

TPR is defined as $\frac{TP}{TP+FN}$ and FPR is defined as $\frac{FP}{FP+TN}$. The ROC curve is created by plotting the TPR against the FPR at various threshold settings.[4]

7. Project Design

Step 1:

Conduct exploratory data analysis. Due to the large feature space in the data model this will be an important step to identify relationships between the features and target variable.

Step 2:

Perform data preprocessing to transform skewed features and normalize numerical features as appropriate.

Step 3:

Engineer new features and prepare any categorical features for training. Engineering new features will be important due to the data model design.

[4] *Receiver Operating Characteristic Wikipedia*

Step 4:

Build and train several different classification models.

Step 5:

Evaluate and choose best model.

Step 6:

Tune model to improve performance results.

Step 7:

Make predictions and prepare submission file.