Udacity Machine Learning Capstone Project Proposal

Prepared for: Udacity
Prepared by: Damilola Omifare
June 2018.

Domain Background.

Home Credit Default Risk

This project is inspired by that fact that many people who deserves loan do not get it and ends up in the hands of untrustworthy lenders.

This project is a competition from Kaggle. Below is the link:
Kaggle | Home Credit Default Risk Competition

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Source : Kaggle

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a

variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.
While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

Problem Statement.

- Can you predict how capable each applicant is of repaying a loan ?

My analysis will be on how to predict how capable each applicant is of repaying a loan.

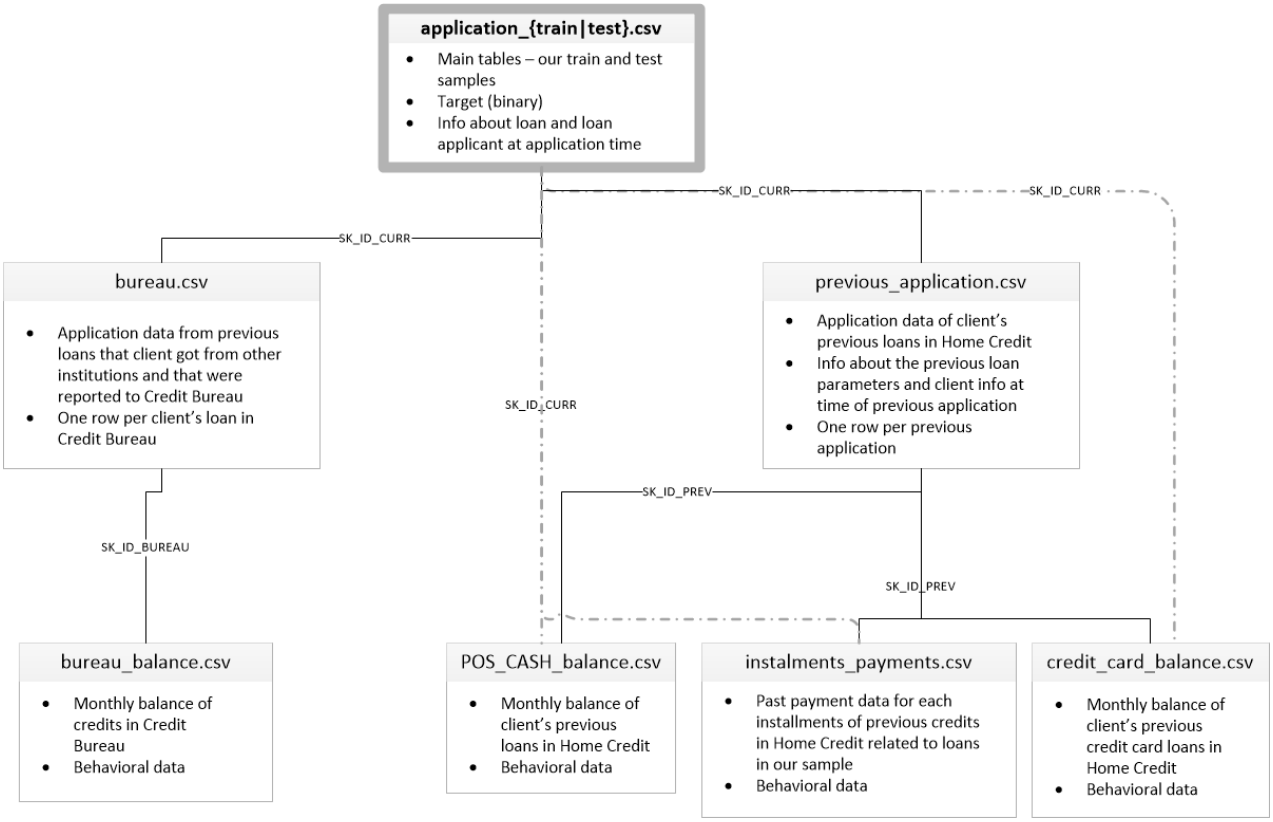
Datasets and Inputs.

The dataset for this project has been provided by Kaggle.
Date description is below :

- **application_{train|test}.csv**
  - *This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).*
  - *Static data for all applications. One row represents one loan in our data sample.*
- **bureau.csv**
  - *All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).*
  - *For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.*
- **bureau_balance.csv**
  - *Monthly balances of previous credits in Credit Bureau.*
  - *This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.*
- **POS_CASH_balance.csv**
  - *Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.*
  - *This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows.*
- **credit_card_balance.csv**
  - *Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.*

- *This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows.*

- **previous_application.csv**
  - *All previous applications for Home Credit loans of clients who have loans in our sample.*
  - *There is one row for each previous application related to loans in our data sample.*
- **installments_payments.csv**
  - *Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.*
  - *There is a) one row for every payment that was made plus b) one row each for missed payment.*
  - *One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.*
- **HomeCredit_columns_description.csv**
  - *This file contains descriptions for the columns in the various data files.*



Data description is taken from *Home Credit Data on Kaggle*

Solution Statement.

I cannot vividly say what the solution is at this time until I do the analysis. It is quite tricky to make a guess on the solution at the moment. But there is a lot in play here. There are a lot of factors to consider in order to predict the probability of whether an applicant will default on loan or not.

Our dependent variable is loan default. We do not know what the target variable but I will be predicting what the target should be.

Independent variables will be some financial variables like loan amount, borrower's credit grade, debt to income ratio ratio etc. After this, the dots will be connected together to make a solution statement.

Benchmark Model

I intend to use supervised machine learning algorithm since the data are labelled. However there are more papers and researches done on this topic. Few are : Machine learning in mortgage default prediction, Identifying signals for loan default in Text of Loan applications , *An analysis of default risk in mobile home credit* . There will be more research done on this during the course on the project.

Methodology that has been used on this are Linear Regression model, K Nearest Neighbours etc. I intend to experiment some of these algorithms with models like SVM, Decision Trees , XBoost amongst others.

Evaluation Metrics

One of the evaluation metrics I intend to use the confusion metrics and statistical measurement (i.e Recall and Precision) then I will set a threshold for the goal (which is if the applicant will default or not). Also I read about something called "Equality of Opportunity" that is equal chance of selection.

Project Design

The workflow of this project, is to load the data, understand the data, clean the data, do some visualisation for better understanding of the data, incorporate some theoretical some concepts about choosing which feature is best for the model, build the model and train it, validate the model and test it.

Based on the way Kaggle competition are structured, we do not have a target variable. So I will split the training into 70:20:10 that training, validation and testing set. The predicted variables are then compared against the actually ones in their database to see how good your prediction is.