**Credit Card Approval Prediction**

**Project Proposal**

Team Member

Aiwen Peng

Xintian Zhang

Jianghui Li

Jinchao Zhao



IST 718: Big Data Analytics

Dr. Emory Creel

Syracuse University

New York

October 2022

# **Project Objective**

Trust is the basis of human interaction, and it is also a moral purpose and means. Credit is a part of people's financial strength and has a significant role in their lives. Good credit may be the detail that makes or breaks your ability to get a credit card, car loan, or student loan. At the same time, bad credit will make it challenging to apply for a credit card or get a low-interest loan. Even if some people don't need a loan, good credit can have a significant impact. For example, landlords, insurance companies, and employers try to use credit information as a basis.

Credit scoring is a standard risk control tool for banks or companies. Personal information and historical data about the applicant are used to predict whether the applicant will default or have bad debts in the future. Banks can decide whether to offer applicants a loan or credit card by credit scoring. This can help the bank to reduce the risk.

**Data Description**

* Overview/Description

Two datasets will be used for this project: the application record and the credit record. The application record dataset includes customers’ personal information, and the credit record dataset contains customers’ monthly credit status information.

* Number of rows and columns

For the application record dataset, it has 438557 rows and 18 columns. The columns are ['ID','CODE\_GENDER','FLAG\_OWN\_CAR','FLAG\_OWN\_REALTY','CNT\_CHILDREN','AMT\_INCOME\_TOTAL','NAME\_INCOME\_TYPE','NAME\_EDUCATION\_TYPE','NAME\_FAMILY\_STATUS','NAME\_HOUSING\_TYPE','DAYS\_BIRTH','DAYS\_EMPLOYED','FLAG\_MOBIL','FLAG\_WORK\_PHONE','FLAG\_PHONE','FLAG\_EMAIL','OCCUPATION\_TYPE','CNT\_FAM\_MEMBERS']. For the credit record dataset, it has 1048575 rows and 3 columns. The columns are ['ID','MONTHS\_BALANCE','STATUS'].

* Sample predictors

The application's personal information (we will determine which are useful columns that can be used to make predictions as we further explore the dataset) and credit history is used to predict whether the applicant is a 'good' or 'bad client.

* A link to the dataset

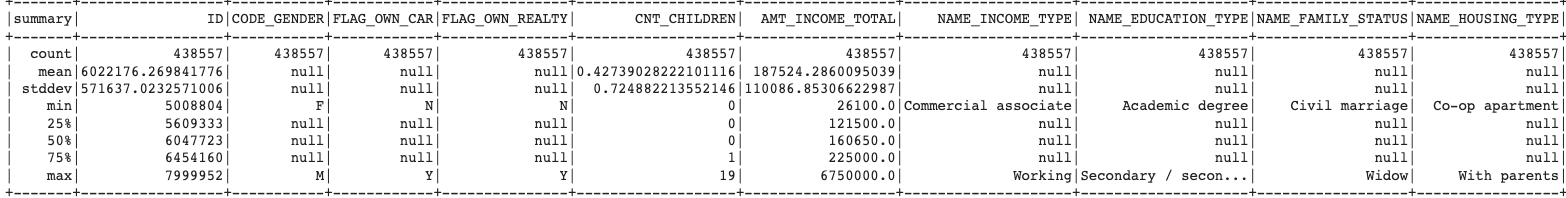
The dataset is download from Kaggle, <https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction?select=application_record.csv>.

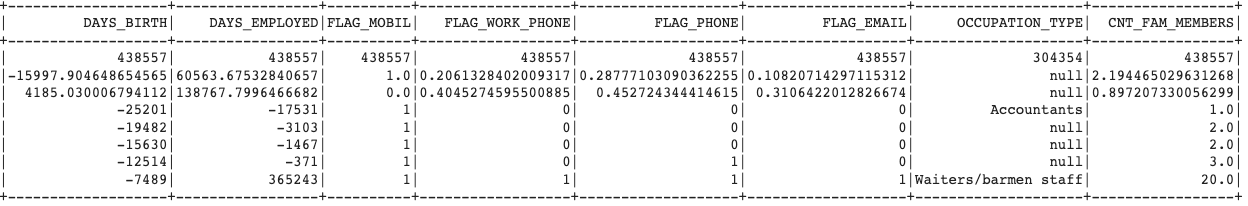
* Interesting or surprising about the data

The dataset contains the information of over 400,000 customers, so we expect the customers’ gender would proportion equally, but only about one third of the customers are male, and the rest are female. Also, by looking at the min and max values, we found one not married female customer has 19 children.

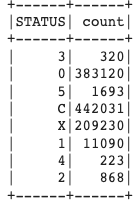
**Preliminary Data Exploration**

The plot below shows a summary for application record dataset, although for some columns (like categorical columns), the results are meaningless. We still can see that, the average children number is 0.427, the maximum children is 19; the average amount of income is $187,524, the minimum is $26,100, and the maximum is $6,750,000. The average age of application is about 44 years old ( the -15997 in column ‘DAYS\_BIRTH’ means the person was born 15997 days ago), the oldest applicant is 69 years old, and the youngest applicant is 20 years old. The average number of family members is 2.19, and the minimum is 1, the maximum is 20.





About the credit record dataset, we can see the different ‘STATUS’ and count in the plot below. C means load paid off, X means No load, 0 means 1 month past due, 1 means 2 months past due… As we can see, most people pay their load on time or no load. Only a small number of people are overdue by more than a month.



Both the application record dataset and credit record dataset need to be merged, and the ‘STATUS’ column in the credit record needs to be annotated in a machine learning friendly way (Ex: C, X, 0, 1 represents a customer has a good credit record; 2, 3, 4, 5, on the other hand, represents a customer has a bad credit record). After we merge the datasets, we will be able to compare the relationships between all the variables.

**Predictions**

* Do applicants older than 25 have better chances to be ‘good’ clients?
* Do applicants with high education and high income have better chances to be ‘good’ clients?
* Does ‘Housing Type’ have any impact on predicting an applicant is ‘good’ or ‘bad’?
* Does ‘the number of children’ and ‘Income’ together have any impact on predicting an applicant is ‘good’ or ‘bad’?
* Does ‘Income Type’ have any impact on predicting an applicant is ‘good’ or ‘bad’?

**Inference**

We plan to build some models with the capability to determine whether applicants are trustworthy enough to lend money by analyzing their personal information. In this dataset, we are going to do some feature engineering among numerical columns and one-hot encoding to the rest of the categorical data. Then, we are going to explore the correlation between these variables so that we can identify the few important predictors among a large set of possible variables. Two datasets will be merged on customers ID, and the merged dataset will be separated into two parts with one for training purposes and one for testing purposes. Then, we will be building models with optimized accuracy on training data, we'll apply the model to the testing data and see what accuracy will be to consolidate our conclusion further. To overcome overfitting problems, we will use the k-fold cross-validation method to estimate how well our models perform on new data.

Methods and Models will be attempted:

* Summary statistics
* Logistic Regression
* Decision Tree
* Random Forest
* XGBoost
* LightGBM