Capstone Project Proposal: Home Credit Default Risk

James Matheson

June 21, 2018

# Assessing the Problem

Trasitionally, borrowing costs have been tied to creditors’ assessment of credit default risk based on simple, but broad financial criteria. Unfortunately, this leaves many potential borrowers out of the market, or paying higher interest rates. Conversely, these same criteria may not always be sufficient to redudce default risk by clients who may meet the criteria, but may eventually default because of other, potentially forseeable, reasons not sufficiently accounted for in the criteria.

# Client Solutions

Using historical client credit data, I aim to create a model capable of predicting, as accurately as possible, an individual applicants’s liklihood of defaulting. This model could then be used by a client to make more refined decisions as to whom they offer loans, the types of loans offered and the interest terms.

# Data

The data is available on the [Kaggle Compeitions website](https://www.kaggle.com/c/home-credit-default-risk/data). It consists of:

1. application\_{train|test}.csv - the main data file, with test and train sets
2. bureau.csv - each client’s previous creidit as provided by a credit bureau
3. beareau\_balance.csv - each client’s monthly credit balances as provided by a credit bureau
4. POS\_CASH\_balance.csv - monthly balance snapshots of previous POS and cash loans that the applicant had with Home Credit.
5. credit\_card\_balance.csv - monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
6. previous\_application.csv - all previous applications for Home Credit loans of clients who have loans in the sample
7. installment\_payments.csv - repayment history for the previously dispersed credits in Home Credit as applicable to the sample
8. HomeCredit\_columns\_description.csv - column descriptions for the other files.

# Proposed solution

1. Clean and merge the “train” data to get a single data frame (if possible!) that contains all of the above files in simple
2. Examine the data to find strong collelations with debtors’ abilty to repay the loans.
3. Create an R program to test for default risk.
4. Apply this test to the “test” data and examine the results.
5. Use visualization tools to deomonstrate these correlations, the accuracy of the model as applied to the train data, and the results of the model as applied to the test data.

# Deliverables

# 1. The R-coded code for assessing and predicting credit risk

## Assessing the Problem

Trasitionally, borrowing costs have been tied to creditors’ assessment of credit default risk based on simple, but broad financial criteria. Unfortunately, this leaves many potential borrowers out of the market, or paying higher interest rates. Conversely, these same criteria may not always be sufficient to redudce default risk by clients who may meet the criteria, but may eventually default because of other, potentially forseeable, reasons not sufficiently accounted for in the criteria.

# Client Solutions

Using historical client credit data, I aim to create a model capable of predicting, as accurately as possible, an individual applicants’s liklihood of defaulting. This model could then be used by a client to make more refined decisions as to whom they offer loans, the types of loans offered and the interest terms.

# Data

The data is available on the [Kaggle Compeitions website](https://www.kaggle.com/c/home-credit-default-risk/data). It consists of:

1. application\_{train|test}.csv - the main data file, with test and train sets
2. bureau.csv - each client’s previous creidit as provided by a credit bureau
3. beareau\_balance.csv - each client’s monthly credit balances as provided by a credit bureau
4. POS\_CASH\_balance.csv - monthly balance snapshots of previous POS and cash loans that the applicant had with Home Credit.
5. credit\_card\_balance.csv - monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
6. previous\_application.csv - all previous applications for Home Credit loans of clients who have loans in the sample
7. installment\_payments.csv - repayment history for the previously dispersed credits in Home Credit as applicable to the sample
8. HomeCredit\_columns\_description.csv - column descriptions for the other files.

# Proposed solution

1. Clean and merge the “train” data to get a single data frame (if possible!) that contains all of the above files in simple
2. Examine the data to find strong collelations with debtors’ abilty to repay the loans.
3. Create an R program to test for default risk.
4. Apply this test to the “test” data and examine the results.
5. Use visualization tools to deomonstrate these correlations, the accuracy of the model as applied to the train data, and the results of the model as applied to the test data.

# Deliverables

1. The R-coded code for assessing and predicting credit risk
2. The of this code as applied to the “test” data set, presented via R Markdown.