Capstone-Proposal

July 3, 2018

1 Capstone Proposal

1.1 Domain Background

Loans are very popular in US. People take loans for education, property, car, etc. The difinition of loan is:
A loan is money, property or other material goods that is given to another party in exchange for future repayment of the loan value amount along with interest or other finance charges [1].There is a risk that a lender may not receive given credit back, it's called 'Credit Risk' [2]. Lenders, of course, want to minimize credit risks, for that they calculate 'Credit Score' [3]. A credit score is a number representing the creditworthiness of a person, the likelihood that person will pay his or her debts. Credit score is composed from several components related to a person, one of major components is 'Credit History'. A credit history is a record of a borrower's responsible repayment of debts.
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. It's possible to use alternative data, including telco and transactional information, to predict their clients' repayment abilities, which can help people to get loans from trustworthy lenders.
My interest in the project is to take part in a real world problem, to work with real world not-prepared/not-cleaned data, take part in a competition (this project is from kaggle platform [4]).

1.2 Problem Statement

There is a company Home Credit which works with the category of unbanked population. They have a set of data: client's previous credits provided by other financial institutions that were reported to Credit Bureau; Monthly balances of previous credits in Credit Bureau; monthly snapshots of credit card balances, cash loans applicant has/had with Home Credit; previous applications for Home Credit. Also they have a historical target score for existing customers, they want to predict for new applicants. The goal of this project is to predict probability that a person will pay it's debts based on related to the person financial information, some kind of alternative to 'Credit Score'. It looks like a regression problem, when it's needed to predict a number based on existing historical data.

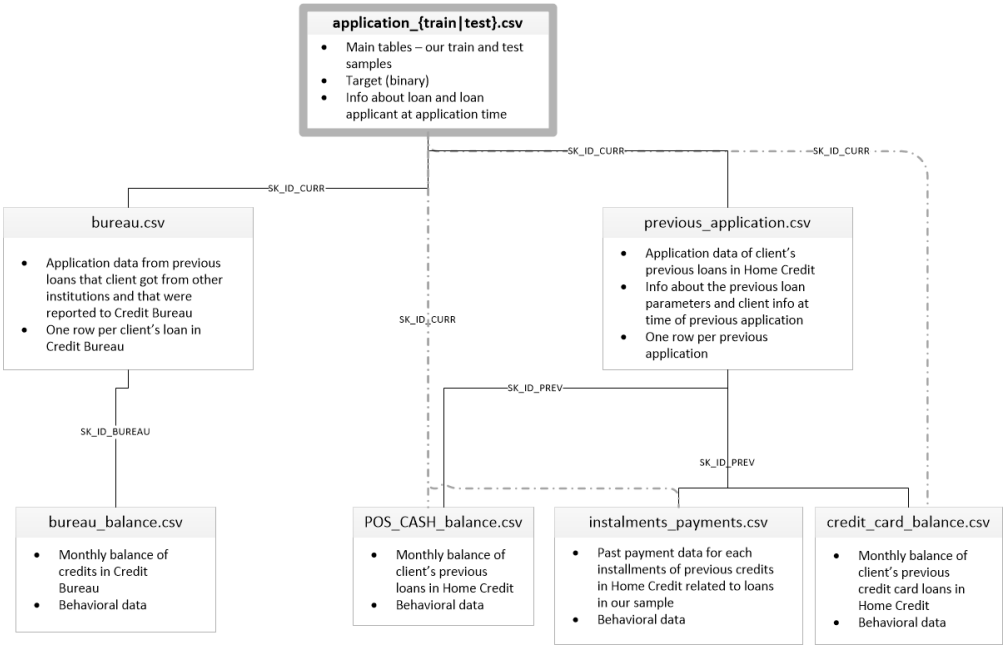
1.3 Datasets and Inputs

Dataset is provided by competition organizer - HomeCredit company, and it is hosted on kaggle platform. There are seven different sources of data:

- application_train/test: the main training and testing 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: the loan was repaid or 1: the loan was not repaid.
- bureau: data concerning client's previous credits from other financial institutions. Each previous credit has its own row in bureau, but one loan in the application data can have multiple previous credits.
- bureau_balance: monthly data about the previous credits in 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: 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 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: payment history for previous loans at Home Credit. There is one row for every made payment and one row for every missed payment.

HomeCredit_columns_description.csv file contains descriptions for the columns in the various data files. There are 221 row in the file with human-readable description for each column.

This diagram shows how all of the data is related:



Data Relation

application_train.csv file contains number of columns with categorical and continuous values, it also contains `target` column which should be predicted. It looks like classic Regression task [5], however `target` is either 1 or 0, so the complexity here is develop some sort transformation

of the target to make it in range [0, 1] and then apply regression algorithms. Overall it looks like solveable problem based on provided data.

1.4 Solution Statement

An existing regression algorithm of supervised learning [6] can be considered to solve the problem. E.g. SVM regressor or neural netwrok based solution which predicts probability. Set of provided inputs is possible to transform to a matrix of numbers, where each row will represent a loan related data and target probability, a row will be algorithm intput to train on. Of course, data should be prepared beforehand (cleaned up, categorical features transformed, continuous features scaled if needed). Once algorithm is trained it is possible to use it to make prediction of probability if applicant will pay debts back, based on data collected.

1.5 Benchmark Model

Test set is provided by HomeCredit company to test a model. It is in the application_test file, which is the same as application_train, except it does not have `target` column. It is supposed that the trained algorithm predictions will be stored in a certain format and submitted to kaggle for evaluation. HomeCredit makes comparision to their existing results which are histrically observed targets. Results from such comparision between existing data results and developed model will be as a benchmark for the developed model.

1.6 Evaluation Metrics

Submissions are evaluated on area under the ROC curve [7] between the predicted probability and the observed target. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The evaluation logic is the next: submitted results are actually two probability distributions of two classes - 0: the loan was repaid or 1: the loan was not repaid. ROC curve will show how we did that split, it visualizes all possible classification thresholds. AUC represents the probability that a classifier will rank a randomly chosen positive observation higher than a randomly chosen negative observation.
The result will help HomeCredit business to decide if they want e.g. to minimize False Positive Rate or maximize True Positive Rate.

1.7 Project Design

At first it is important to explore data, understand each column, what does it mean thinking about it in human common sense, think of it if I can do a decision which I expect model should do. Working with data is the next step and it is very important step because it is real, unprepared data. I expect there are missing values, which should be fulfilled with some values, categorical values should be encoded to numerical, continuous values should be normalized and analized if any transofrmation should be applied, if values skewed a lot. All these data preparation steps are very important because may have high impact on most of supervised learning algorithms perfromance.
The next thing is outlier analisis, if it's safe to remove them or not. Sometimes outliers are very important and actually are target points we want algorithm to take into account. Some algorithms are sencitive to outliers, some are not.
It is important to analyze correlation between features and engineer new features if needed.

There are several files on input data, it is important to use all available data, so it is important to make feature extraction and generate one final csv from all provided files with training and testing data.
The next step is to find appropriate algorithm to use. There are a lot of options, and it depends on data which one to use, how many continuous and categorical values are there, how many features will be at the end of preprocessing all the files into the one single file, how many datapoints are there, everything is important for algorithm selection. Of course, there will be few possible candidates and all of them should be tried out. It is possible several candidates will be potentially good. For each candidate it is important to give a chance and work on parameters tuning. In this project there are no any requirements how fast should be the final model, how much memory it should use, so only final result is important, which broadens spectre of possible candidates. Parameter tuning may take a lot of time, so only computational time is limiting here, e.g. it may not be possible to teach few nerual networks on the same machine because it may consume all GPU memory.

1.8 Resources

1. Investopedia - Loan
2. Wikipedia - Credit Risk
3. Wikipedia - Credit Score
4. Kaggle - Home Credit
5. Wikipedia - Linear Regression
6. sk-learn - supervised learning
7. Wikipedia - ROC curve