Machine Learning Engineer Nanodegree

Capstone Proposal

Ralph Pinotti Leite

December 26, 2018

1. Domain Background

For my capstone project, I will use 'Home Credit Default Risk' that is one challenge from Kaggle. Kaggle is an online community of data scientists that allows users to solve real problems as a challenge. The challenge in question is to unlock the full potential of the data that are possible to use to 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.

As a basic scenario, we have many people struggle to get loans due to insufficient or non-existent credit histories. Unfortunately, this population is often taken advantage of by untrustworthy lenders. 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. (https://www.kaggle.com/c/home-credit-default-risk)

Traditionally credit is extended to people based on a credit score. Most people in the developed world went through a "credit building" phase in their lives, but 4.5 billion people around the world have little or no credit history (source: LTP – Let's Talk Payments, March 16, 2017). Brazil currently has about 60 million people unbanked. This number represents almost half of the economically active population. The group moves R $ 665 billion a year, more than the GDP of countries like Chile and Singapore and is spread by economic classes, reports the State of São Paulo. This is a potential market for new companies, but currently, less than 10% of the 350 fintechs in Brazil are looking for those who do not

have access to basic financial services. http://www.convergenciadigital.com.br/cgi/cgilua.exe/sys/start.htm?infoid=48184&sid=161

I think these numbers and statistics very interesting because it can show a huge opportunity to financial market. I expect to have good progress and result in this project, it is an opportunity to apply machine learning techniques that are not usual in this field (usually professionals solve this problem using logistic regression). With the project result, I intend to have real-world problem results to improve the technique used in my company.

2. Problem Statement

The objective of this competition is to use historical loan application data and bureau information to predict whether or not an applicant will be able to repay a loan. Predicting whether or not a client will repay a loan or have difficulty is a critical business need. For each client information, we should classifier with the probability of repaying and compare with the original result (the observed result of client repays or not) based on an algorithm of our choice aiming the best result. We will measure the result with the Area Under the Curve (AUC). The metric is between 0 and 1 with a better model scoring higher. A model that simply guesses at random will have a ROC AUC of 0.5. After finding our best classifier, we can apply the created rule in several clients that have the same information that the original model.

3. Datasets and inputs

The data is provided by Home Credit, a service dedicated to provide lines of credit (loans) to the unbanked population. The information is used for a challenge in Kaggle. The name of the challenge is “Home credit default risk”. The data set contains information from 356,255 individuals who had previously been recipients of loans from Home Credit and each individual represented by their loan ID. The full data are composed by the following data source and can be found in (https://www.kaggle.com/c/home-credit-default-risk/data)

A. Application_train/application_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.
B. 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.
C. 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.
D. 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.
E. 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.
F. 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.
G. 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.

4. Solution Statement

In this project, I will assume a general machine learning roadmap that has the following steps:

1. Understand the problem and the data
2. Data cleaning and formatting.

3. Exploratory Data Analysis
4. Baseline model
5. Improved mode
6. Chosen model and features interpretation

We will first understand the data, our task, and what our target variable is. Along the way, we performed necessary preprocessing steps such as encoding categorical variables, imputing missing values, and scaling features to a range. Then, we performed a fairly simple EDA to try and identify relationships, trends, or anomalies that may help our modeling. After that, we will implement a baseline model upon which we hope to improve. Then we built a second slightly more complicated model to beat our first score. We also carried out an experiment to determine the effect of adding the engineering variables to search the best AUC.

5. Benchmark model

Each kaggle competition has a Leaderboard rank. The leaderboard is composed of teams that are ordered by results. Each result is the metric that was chosen to evaluate the challenge.

As a benchmark model, we have the Kaggle competition own result. We have 4 benchmarks based on AUC metrics from users. (https://www.kaggle.com/c/home-credit-default-risk/leaderboard)

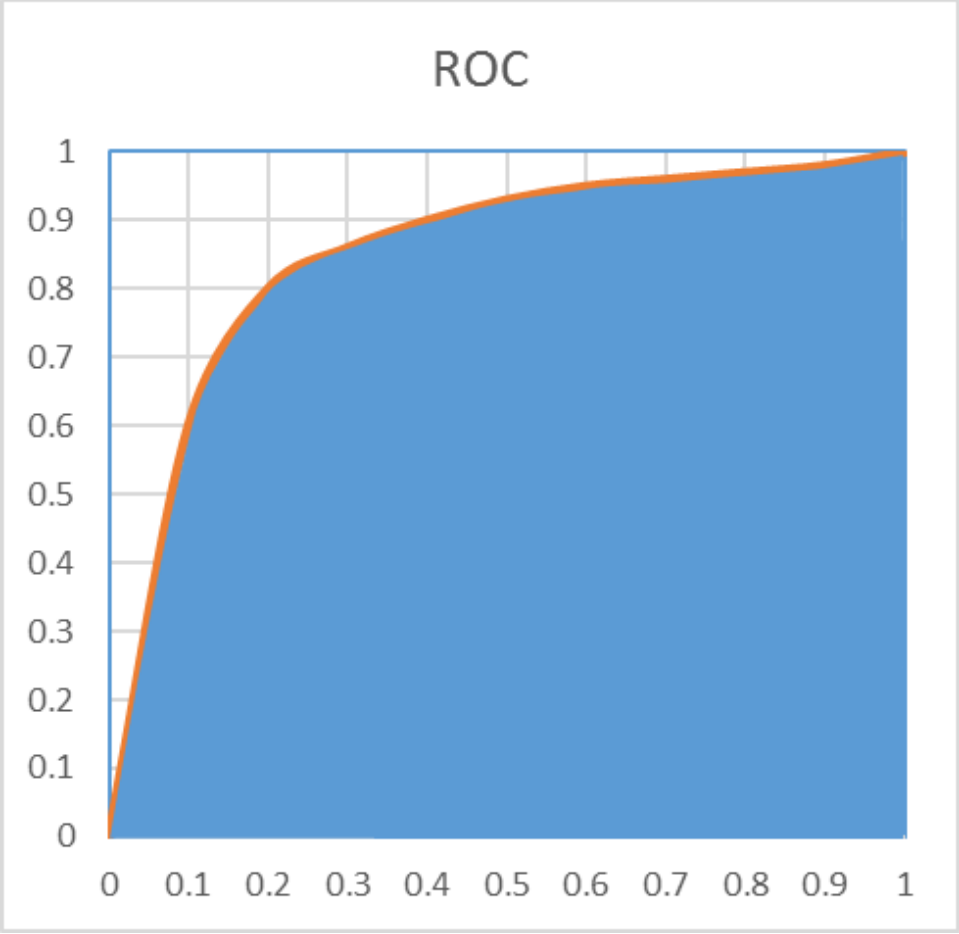
1. In the money: Fist 3 high AUC
2. Gold: From 4 to 24 high AUC
3. Silver: 25 to 559 high AUC
4. Bronze: 600 to 719 high AUC

Usually, the leaderboard is composed of experienced users and in this case, the difference between the first AUC and the last Bronze (719th) are in the decimal case. I expect to do a good job and get an AUC around the bronze rank.

6. Evaluation metrics

We will measure the result with the Area Under the Curve (AUC), the metric is between 0 and 1 with a better model scoring higher. A model that simply guesses at random will have a ROC AUC of 0.5. After finding our best classifier we can apply the created rule in several clients that have the same information where the model was built

ROC AUC (https://medium.com/@andygon/eli5-roc-curve-auc-metrics-ac4fe482f018**)**



The area shaded blue is your model's performance. The closer it gets to the top left corner, the better your model is doing at distinguishing your two classes apart. This area/graph is specific to your model. Every model you build will have its own ROC.

The vertical axis graphs the True Positive Rate (TPR), aka the relation between True Positives and all recorded Positives, versus;

The horizontal axis graphs the False Positive Rate (FPR), which is the complementary/parallel metric of the True Negative Rate, which, just like TPR, is the True Negatives over all recorded Negatives.

7. Project Design

I will follow a simple step that can be applied to several problems in the real world.

1. Understand the problem and the data
2. Data cleaning and formatting
3. Exploratory Data Analysis
4. Baseline model
5. Improved model
6. Chosen model and features interpretation

1. Understand the problem and the data

a) Understand the problems and the data availed to use
b) Count how many classes we have. Balanced or no balanced source
c) Find the target variable
d) Make a description of how many descriptive variables we have
e) Look for a good metric to validate the result: AUC for this problem
f) Understand what kind of problem is this: Supervised, no supervised etc...
g) Selected some algorithm to use
h) Define how to join all tables

2. Data cleaning and formatting:

a) Exists or not blank values in numeric and a string variable that will need to be handled
b) Wich variable is numeric or categorical

c) In numeric variable, adjust the format if necessary

d) Determine what variable has a lot of missing value and decided to remove from the dataset

e) Define if in a categorical variable are necessary to group our not

3. Exploratory Data Analysis

a) Compile lists and short descriptions of each feature

b) Generate and display statistical descriptions of the numerical features

c) Find if the numeric features have a normal distribution

d) Group categories to build new variables if necessaire

e) Create categorization based on odds/relative risk ore others metrics

4. Baseline model

a) Create a train data set and test dataset

b) Selected features with an algorithm to use in the first model(SelectKbest)

c) Create a simple model with default parameters as a baseline model ursin Logistic regression

d) Measure the result with AUC in the test data set

5. Improved model

a) Improve the parameters of the baseline model with grid

b) SelectKBest for feature selection.

c) Run another kind of classifiers as:

   i. Logistic Regression Classifier
   ii. Gaussian Naive Bayes Classifier
   iii. AdaBoost Classifier
   iv. Bagging Classifier
   v. Extra Trees Classifier
   vi. Gradient Boosting Classifier
   vii. SVM
   viii. RandomFloret classifiers

6. Chosen model and features interpretation

a) Chose the best result in AUC from previous steps

b) If possible, try to explain the selected features in the model selected

c) Answer the problems with the result of the model