Deploying a Credit Risk Application Capstone Project

Build an algorithm for credit scoring

*David Ricardo Mora Guerrero*

July 29, 2021

# Definition

## Project Overview

## Problem Statement

When banks and financial institutions carry out a credit operation with a particular client face the problem of uncertainty. Certainly, is impossible for them to know, with full confidence, if the client will comply with the future payments, this fact is commonly called credit risk. In other words, credit risk, is the possible loss assumed by the institution as a consequence of a default of payments.

To reduce this cost, institutions commonly design different ways or strategies (business rules or filters) to classify / discriminate their clients. They try to reduce the cost of credit risk by trying to allocate their resources (loanable funds) in low-risk credit operations and/or avoid placing them into high-risk credit operations.

This project aims to provide a documentation, of how this problem can be addressed using a machine learning algorithm. The algorithm will learned to discriminate from “good” and “bad” clients, from a set of training data, which contains relevant features of the clients and their credit operations. Then, accordingly to what it learned, it will predict a credit score, which might be used by an entity to effectively discriminate between clients. And later, by integrating this algorithm to the credit approval flow, it could potentially reduce credit risk costs.

Additionally, I will show how to use Amazon SageMaker, AWS Lambda and API Gateway, to create a REST API, which you can establish communication with the model, and thus be able to receive requests and send responses in real time.

# Data

The data will be obtained from the following Kaggle source

https://www.kaggle.com/ethon0426/lending-club-20072020q1

Data coverage

Annual Income, Age, Employment Length, Home Ownership.

**Characteristics of the Loan**

Amount of the Loan, Interest Rate,

The monthly payment owed by the borrower if the loan originate

The number of months since the borrower's last delinquency.

# Setup

Python 3.6

Inbalanced-Learn, NumPy, Pandas, Scikit-Learn

# Tasks

To finish this project I should complete the following tasks.

1. Define a target variable

2. Perform an exploratory data analysis

3. Select the algorithm that best fits the validation data

4. Report the performance of the selected algorithm on test data.

5. Create a scale or credit score by using the probability estimates which can be interpreted by your users.

6. Deploy the model on an Amazon SageMaker endpoint.

7. Create a function in AWS lambda, which will pre-process the data that will be sent to the endpoint.

8. Create an API in Amazon API Gateway, which can be used to communicate with the model.

## Blog Post in Medium

All the process will be documented in the following medium account:

https://medium.com/@davidricardocr

## Test and measure performance

Two metrics will be used to evaluate and select an algorithm. One of them is the Kolmogorov–Smirnov statistic (or just KS test) which quantifies the distance between two empirical distributions functions. The second one, is the Area Under the Receiver Operating Characteristic Curve (ROC AUC) score.