Udacity Capstone Project Idea Proposal

Machine Learning Engineer Nanodegree

Muskaan Patel

23 December 2020

New Delhi

Credit Card Approval Prediction using Machine Learning

- ## Domain Background:

This project comes under the domain of Application of Machine Learning and Data Science in the Finance Industry. Data Science has inarguably beeen the trendiest domain of the decade, asserting its need in every single sphere of corporate life. It was not long ago when we discovered the massive potential of incorporating ML/AI in the financial world. Now, the very idea of two being disjointed sounds strange. Data Science has been incremental in providing powerful insights and helped massively increase the efficiency, helping everyone from a scalp trader to a long-term debt investor. Accurate predictions, unbiased analysis, powerful tools that run through millions of rows of data in the blink of an eye has transformed the industry in ways we could’ve never imagined.

- ## Problem Statement:

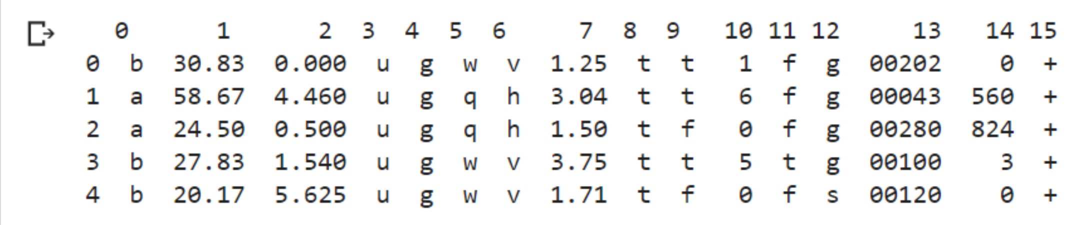
Nowadays, banks receive a lot of applications for issuance of credit cards. Many of them rejected for many reasons, like high-loan balances, low-income levels, or too many inquiries on an individual’s credit report. Manually analysing these applications is error-prone and a time-consuming process. Luckily, this task can be automated with the power of machine learning and pretty much every bank does so nowadays. In this project, I will build an automatic credit card approval predictor using machine learning techniques, just like the real banks do [1].

The accurate assessment of consumer credit risk is of uttermost importance for lending organizations. Credit scoring is a widely used technique that helps financial institutions evaluates the likelihood for a credit applicant to default on the financial obligation and decide whether to grant credit or not. The precise judgment of the creditworthiness of applicants allows financial institutions to increase the volume of granted credit while minimizing possible losses. The credit industry has experienced a tremendous growth in the past few decades. The increased number of potential applicants impelled the development of sophisticated techniques that automate the credit approval procedure and supervise the financial health of the borrower. The goal of a credit scoring model is to classify credit applicants into two classes: the “good credit” class that is liable to reimburse the financial obligation and the “bad credit” class that should be denied credit due to the high probability of defaulting on the financial obligation. The classification is contingent on sociodemographic characteristics of the borrower (such as age, education level, occupation and income), the repayment Performance on previous loans and the type of loan [2].

This paper aims to apply multiple machine learning algorithms to analyse the default payment of credit cards. By using the financial institutions client data provided by UCI MachineLearning Repository, we will evaluate and compare the performance of the model candidates in order to choose the most robust model. Moreover, we will also decide which are important features in our best predictive model [3].

- ## Datasets and Inputs:

In this project, I’ll be using Credit Card Approval Dataset from UCI Machine Learning Repository. A snippet of the dataset is displayed below.

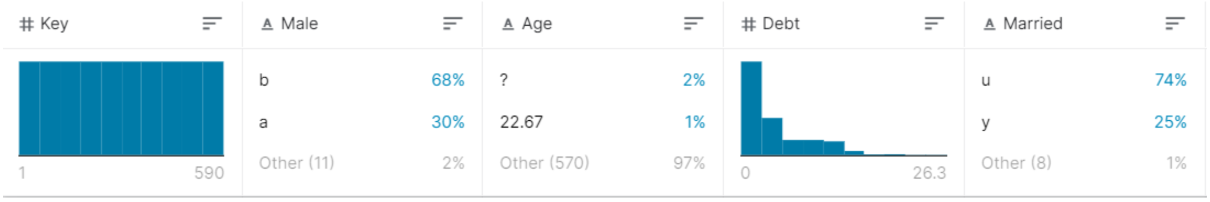


```
   0      1      2  3  4  5  6     7  8  9  10 11 12     13   14 15
0  b  30.83  0.000  u  g  w  v  1.25  t  t   1  f  g  00202    0  +
1  a  58.67  4.460  u  g  q  h  3.04  t  t   6  f  g  00043  560  +
2  a  24.50  0.500  u  g  q  h  1.50  t  f   0  f  g  00280  824  +
3  b  27.83  1.540  u  g  w  v  3.75  t  t   5  t  g  00100    3  +
4  b  20.17  5.625  u  g  w  v  1.71  t  f   0  f  s  00120    0  +
```

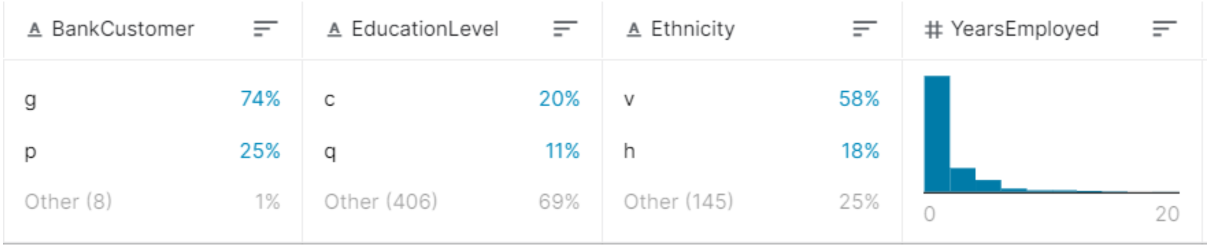
On analysing the dataframe, I find that since the data is confidential, the contributor of this dataset has anonymized the feature names. The features of this dataset have been anonymized to protect the privacy, but upon more research I found a blog that gives us a pretty good overview of the probable features. The probable features in a typical credit card application are Gender, Age, Debt, Married, BankCustomer, EducationLevel, Ethnicity, YearsEmployed, PriorDefault, Employed, CreditScore, DriversLicense, Citizen, ZipCode, Income and finally the ApprovalStatus [1,4].

The link to the dataset is: http://archive.ics.uci.edu/ml/datasets/credit+approval

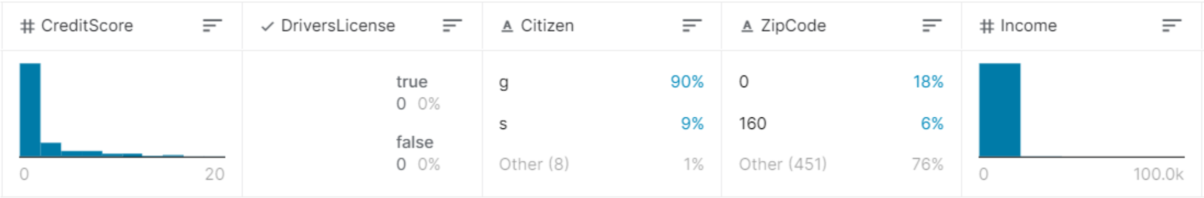
A descriptive distribution of my dataset is provided below. It describes the distribution of the various attributes of the data frame used [8].



A Male		A Age		A Married	
b	68%	?	2%	u	74%
a	30%	22.67	1%	y	25%
Other (11)	2%	Other (570)	97%	Other (8)	1%

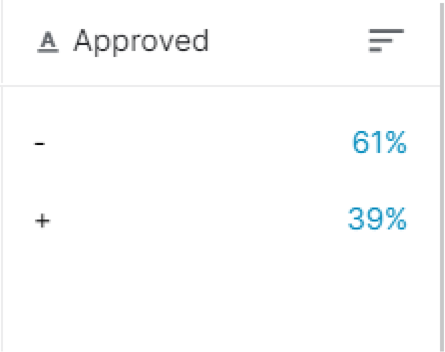


BankCustomer		EducationLevel		Ethnicity		YearsEmployed
g	74%	c	20%	v	58%	
p	25%	q	11%	h	18%	
Other (8)	1%	Other (406)	69%	Other (145)	25%	0 20



CreditScore	DriversLicense	Citizen		ZipCode		Income
	true 0 0%	g	90%	0	18%	
	false 0 0%	s	9%	160	6%	
0 20		Other (8)	1%	Other (451)	76%	0 100.0k

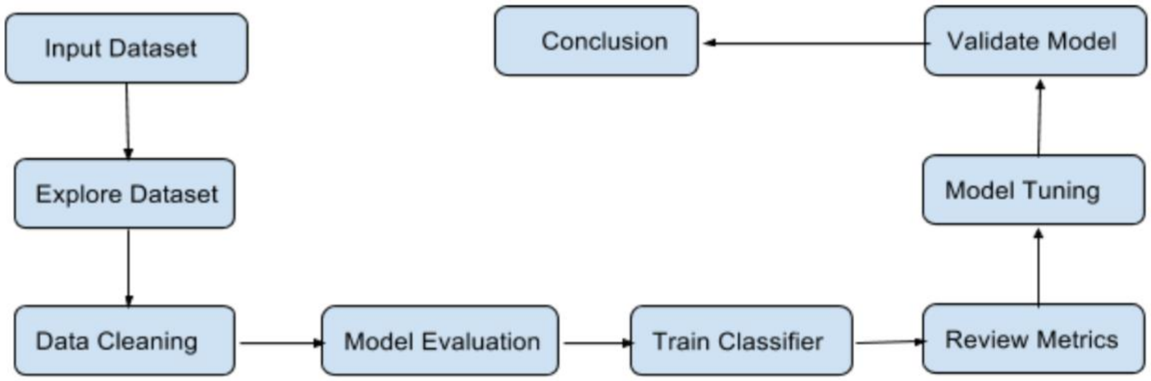
The key label attribute, i.e the "Approved" column has a fairly balanced distribution. Its description is also displayed below [8].



Approved	
-	61%
+	39%

Proposed Solution Statement:

- First, I will start off by loading and viewing the dataset.
- I will have to pre-process the dataset to ensure the machine learning model we choose can make good predictions.
- After our data is in good shape, we will do some exploratory data analysis to build our intuitions.
- After uploading the data to S3, I'll create an Estimator along with Hyperparameters for our preferred ML algorithm.
- I will build and deploy a machine learning model using endpoint configuration that can predict if an individual's application for a credit card will be accepted.
- I'll compare a few different ML algorithms along with their accuracy metrics and choose the best one.



- ## Benchmark Model:

Essentially, predicting if a credit card application will be approved or not is a classification task. According to UCI, our dataset contains more instances that correspond to “Denied” status than instances corresponding to “Approved” status. Specifically, out of 690 instances, there are 383 (55.5%) applications that got denied and 307 (44.5%) applications that got approved.

This gives us a benchmark. A good machine learning model should be able to accurately predict the status of the applications with respect to these statistics.

After doing a thorough research, I’ve found a number of papers on this topic. A number of algorithms have been used, more popular ones are Logistic Regression [2] Classification and Regression Trees (CART) [4]. But the maximum accuracy their model achieved was near to 76% [2]. I plan to exploit the AWS SageMaker’s interface and various Machine Learning features to achieve a good level of Explanatory Data Analysis, Feature Engineering, Data Pre-Processing, choosing an appropriate ML model and Hyperparameter Tuning to produce better and more accurate results.

The specific benchmark model that’ll be used for my project will be the Logistic Regression model along with 76% accuracy [2]. I’ll try to improve the other specific domains such as data pre-processing and exploratory data analysis to achieve a model with better accuracy and efficiency.

- ## Evaluation Metrics:

One of the best and most popular evaluation metrics is the F1 – score or the confusion matrix. We will now evaluate our model on the test set with respect to classification accuracy. But we will also take a look the model’s confusion matrix. In the case of predicting credit card applications, it is equally important to see if our machine learning model is able to predict the approval status of the applications as denied that originally got denied. If our model is not performing well in this aspect, then it might end up approving the application that should have been approved. The evaluation metric proposed is appropriate given the context of the data, the problem statement, and the intended solution. The confusion matrix helps us to view our model's performance from these aspects [1,2].

Project Design:

Regression models are useful for predicting continuous (numeric) variables. However, the target value in Approved is binary and can only be values of 1 or 0. The applicant can either be issued a credit card or denied- they cannot receive a partial credit card. We could use linear regression to predict the approval decision using threshold and anything below assigned to 0 and anything above is assigned to 1. Logistic regression could also be used because it will produce probability that the target value is 1. Probabilities are always between 0 and 1 so the output will more closely match the target value range than linear regression.

Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. We can also compare the accuracy produced by a simple Decision Tree model and analyse our results accordingly [5].

Other than that, there are a number of other algorithms like Support Vector Machines and XGBoost algorithms. I can exploit all these models and compare its efficiency and accuracy along with other models and choose the best model accordingly.

References:

1. https://medium.com/datadriveninvestor/predicting-credit-card-approvals-using-ml-techniques-9cd8eaeb5b8c
2. http://www.ijrar.org/papers/IJRAR190B030.pdf
3. https://escholarship.org/uc/item/9zg7157q
4. http://rstudio-pubs-static.s3.amazonaws.com/73039_9946de135c0a49daa7a0a9eda4a67a72.html
5. https://medium.com/@kennymiyasato/binary-classification-project-using-decision-tree-with-kaggle-dataset-3123398a1c70#:~:text=Decision%20Tree%20Supervised%20Learning&text=Tree%20models%20where%20the%20target,lead%20to%20those%20class%20labels.
6. https://christophm.github.io/interpretable-ml-book/simple.html
7. https://towardsdatascience.com/interpretability-in-machine-learning-70c30694a05f
8. https://www.kaggle.com/redwuie/credit-card-approval?select=train.csv