Predicting Customer’s Loan Default Risk in Bank

# **Sathish Manthani Winter 2020** <https://github.com/sathishmanthani>

# Which Domain?

This data is going to come from the finance/banking domain. I listed the references below:

* 1. Loan default in Finance:

<https://en.wikipedia.org/wiki/Default_(finance)>

* 1. Identifying the factors responsible for loan defaults:

<https://www.lexjansen.com/mwsug/2016/AA/MWSUG-2016-AA02.pdf>

* 1. Predicting and preventing loan defaults with proactive credit risk management

<https://bigdataforbanking.com/solutions/credit-risk-monitoring/>

* 1. Predicting Chances of Loan Default by a New Customer | Insight Hub

<https://www.youtube.com/watch?v=jmfQeRZRn74>

* 1. The Consumer Loan’s Payment Default Predictive Model:

<https://ideas.repec.org/a/spr/jknowl/v9y2018i3d10.1007_s13132-016-0382-8.html>

* 1. Default Rate – How it works:

<https://corporatefinanceinstitute.com/resources/knowledge/credit/default-rate/>

* 1. Important Factors Determining Fintech Loan Default:

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3587657>

* 1. Predicting Bank Loan Default with Extreme Gradient Boosting:

<https://arxiv.org/ftp/arxiv/papers/2002/2002.02011.pdf>

* 1. Key Technological Advancements in Finance: <https://learning.oreilly.com/library/view/machine-learning-applications/9781484237878/html/464968_1_En_14_Chapter.xhtml>

# Which Data?

I have taken the datasets from a data science competition site:;

<https://zindi.africa/competitions/data-science-nigeria-challenge-1-loan-default-prediction/data>

I found 3 different datasets for each of the train and test data;

1. Demographics data
2. Performance (loan default) data
3. Previous loans data.

Demographics Data:

* customerid – Unique ID of the customer
* birthdate - Date of birth of the customer
* bank\_account\_type - Type of primary bank account
* longitude\_gps
* latitude\_gps
* bank\_name\_clients - Name of the bank
* bank\_branch\_clients - Location of the branch. It’s not a mandatory field
* employment\_status\_clients - Type of employment that customer has
* level\_of\_education\_clients - Highest level of education.

Performance data:

* customerid (Primary key used to merge to other data)
* systemloanid (The id associated with the particular loan. The same customerId can have multiple systemloanid’s for each loan he/she has taken out)
* loannumber (The number of the loan that you have to predict)
* approveddate (Date that loan was approved)
* creationdate (Date that loan application was created)
* loanamount (Loan value taken)
* totaldue (Total repayment required to settle the loan - this is the capital loan value disbursed +interest and fees)
* termdays (Term of loan)
* referredby (customerId of the customer that referred this person - is missing, then not referred)
* good\_bad\_flag (good = settled loan on time; bad = did not settled loan on time) - this is the target variable that we need to predict

Previous loans data:

* customerid (Primary key used to merge to other data)
* systemloanid (The id associated with the particular loan. The same customerId can have multiple systemloanid’s for each loan he/she has taken out)
* loannumber (The number of the loan that you have to predict)
* approveddate (Date that loan was approved)
* creationdate (Date that loan application was created)
* loanamount (Date that loan application was created)
* totaldue (Total repayment required to settle the loan - this is the capital loan value disbursed +interest and fees) termdays (Term of loan)
* closeddate (Date that the loan was settled)
* referredby (customerId of the customer that referred this person - is missing, then not refrerred)
* firstduedate (Date of first payment due in cases where the term is longer than 30 days. So in the case where the term is 60+ days - then there are multiple monthly payments due - and this dates reflects the date of the first payment)
* firstrepaiddate (Actual date that he/she paid the first payment as defined above)

# Research Questions? Benefits? Why analyze these data?

Predicting loan default risk is a critical part of money lending because lenders must know whether giving out a loan will result in profit or loss. Generally, loans are profitable and generate revenue for banks because of interest. But, sometimes a borrower may default which results in a loss of money for lending banks. So, it is important that the lender is able to estimate the risk of a borrower being defaulted before borrowing him/her the money.

Given the several factors that might affect borrower default rate, it may be infeasible to come up with good estimates manually. The objective of this project is to explore whether or not we can employ machine learning models to better predict the loan default risk of borrower. Using exploratory data analysis, we may be able to describe loans and the financial situations of their borrowers, we may also determine the key relationships between default rates and a few other features. Also, we will investigate key relationships between loan default risk and customer behavior.

Research Questions:

1. What are the attributes of the customer influencing the Loan default risk factor.
2. Understand the relationship among customer demographic and financial features
3. What are attributes available in bank accounts data. Would there be any optional attributes for which the data would be missing.
4. Are there any attributes negatively or positively moving with the Loan default risk status feature.
5. Can we really predict who would default in a loan thus avoid bad loans? If yes, with what accuracy.
6. What type of account holders can default the loan most? Is it savings account, checking account or others?
7. Is education level of the customer play a role in loan default risk?
8. What is the impact of employment status on the loan default risk?

# What Method?

What methods will you be using? What will those methods provide in terms of analysis? How is this useful?

Data import

To answer these questions, I would have to initially import the dataset to a dataframe in Python.

CSV library can be used for importing CSV files. Since I have multiple datasets I need to join them first based on the unique key (Customer ID) and do the data processing.

Data Quality

I would check the quality of the data in each variable. If there are any NA, NULL values then removing them from the dataset would give us cleaner data.

Descriptive statistics

Then explore the categorical variables like Education level, Employment status, Credit Score, Account type and check their distribution

I will calculate covariance/correlation. Descriptive statistics would help us understand the data range, median and how each variable is distributed.

Relationships and modeling

I need to understand the factors that are truly influencing the loan default risk and how strong is the relationship between such variables. Explore if the relationship between the target variable and the independent variables is linear or non-linear. Check if a transformation (e.g. logarithm) give us a linear relationship.

Predicting loan default risk status is a binary classification problem, so I will try with Logistic regression, Random forest, Gradient boosting models and test the accuracy of the model.

Data visualization

Relationships are best viewed graphically using Matplotlib package in Python. Boxplots, histograms would work best for exploring single variables. Scatter plot, bar charts would show the bi-variate relationships. Pie Chart can be used for illustrating the proportions.

# Potential Issues?

What challenges do you anticipate having? What could cause this project to go off schedule?

I would require good amount of Good and Bad loans data to build a robust predictive model. If the dataset is skewed towards one of the approval status types heavily then the model may not work accurately. So, that is one potential risk with the approach. At this time, I don’t see any reasons why this project can’t be completed on time.

If I think the given dataset is not good enough then I will look for other data sources for loans data and complete the project.

# Concluding Remarks

Advancements in technology are changing the way of performing financial transactions in a traditional way. During the last few months, credit card usage has increased rapidly all over the world. Customers started using the credit cards for a number of reasons such as paying regular bills, emergencies, online shopping, travel, buying extra protection(insurance) for the rentals etc. Covid-19 pandemic has changed the consumer habits. So, it is more important than ever for banks to predict the lending risk. Defaulting in loans can be restricted by ensuring that loans are made to only borrowers who are most likely to repay, and who is unlikely to become insolvent. Machine learning can help banks predicting the good versus bad customers as far as lending is concerned.

In this project, based on data from Zindi datasets, I will identify the factors that are important for a banking company to identify the lending risk for each customer by looking for the factors in customer demographics and financials. I will model multivariate categorical data using the logistic regression, random forest and gradient boosting models. I would have to cleanse the data, handle the missing data and join the demographics and loans data.

I will split the dataset into two parts, one is training dataset and the other one is testing dataset. I will use the training dataset to build the model. I will use the above said modeling techniques for modeling. I will use the test dataset to test the accuracy of the target variable (loan default risk status) and conclude the best model based on the accuracy.

---- End ---