## Predicting Loan Default Risk

**Author**: Sathish Manthani

**Date**: 2/27/2021

Contents

[Predicting Loan Default Risk 1](#_Toc65400203)

[Introduction 2](#_Toc65400204)

[Research Questions 3](#_Toc65400205)

[Data Sources 4](#_Toc65400206)

[Exploratory Analysis 5](#_Toc65400207)

[Correlation matrix 8](#_Toc65400208)

[Exploratory Data Analysis 9](#_Toc65400209)

[Test/Train data split 11](#_Toc65400210)

[Modeling and evaluation 11](#_Toc65400211)

[References 11](#_Toc65400212)

## Introduction



Image source: [site](https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.dnaindia.com%2Fbusiness%2Freport-top-20-defaulters-account-for-20-of-total-bad-loans-2688038&psig=AOvVaw1pGAysXbhFxysqYmJa_xhd&ust=1614531110428000&source=images&cd=vfe&ved=0CA0QjhxqFwoTCIDpk6DDiu8CFQAAAAAdAAAAABAJ)

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

Machine learning is key in determining the loan default risk. Data collection and analysis also plays vital role. I’m hoping this data can provide us more insights and visualizations into loans data to find meaningful patterns.

Below are my research questions for this project:

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?

## Data Sources

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:

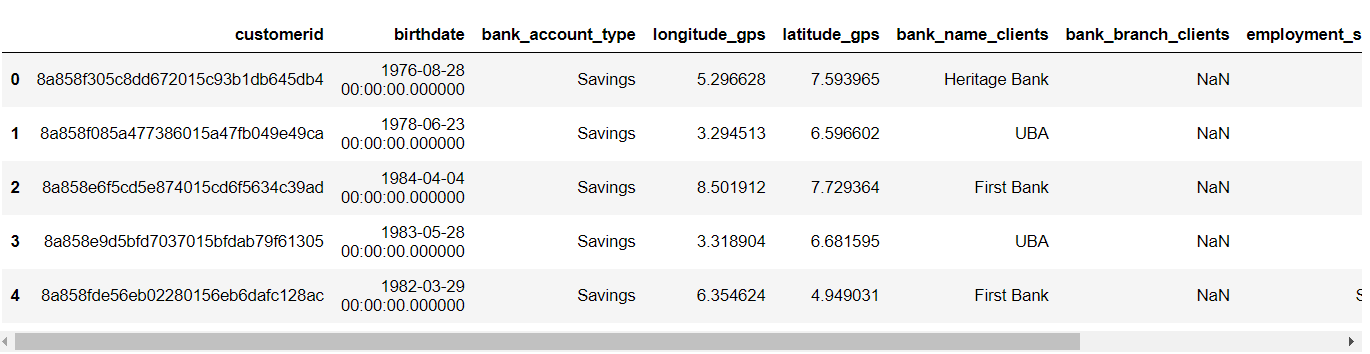
* customerid (Primary key used to merge to other data)
* systemloanid (The id associated with the particular loan. The same customerId can have multiple systemloanid’s for each loan he/she has taken out)
* loannumber (The number of the loan that you have to predict)
* approveddate (Date that loan was 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)

## Exploratory Analysis

The first step is to analyze the dataset by looking at few sample records.

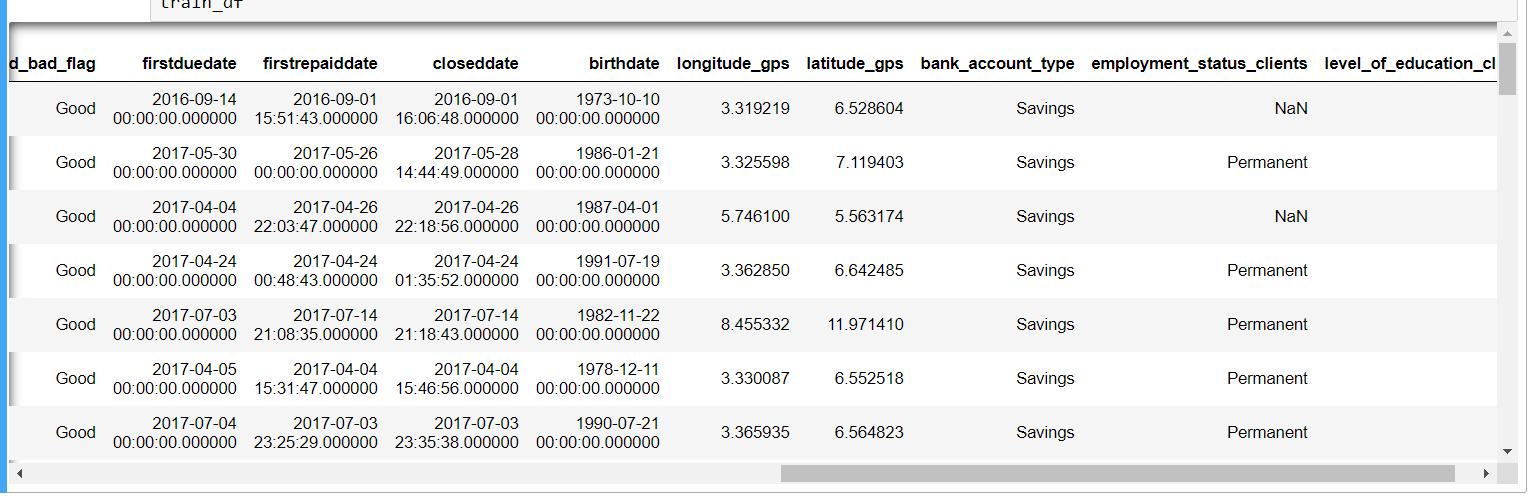
I got training and test datasets separately. I have demographics, performance and previous loans data from Zindi site.

Performance - Training dataset:

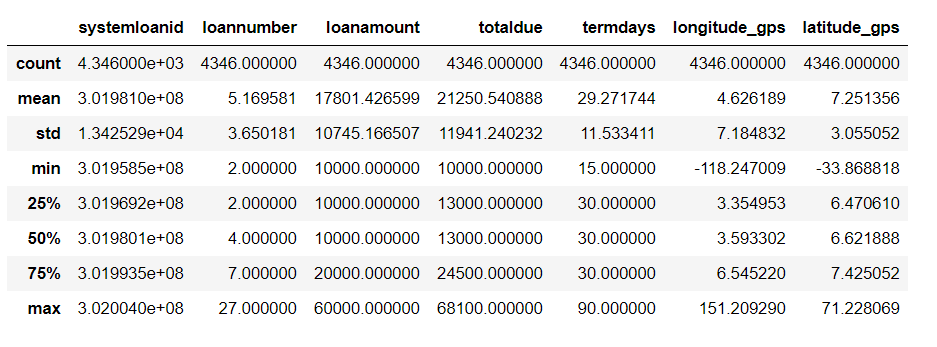


I concatenated the datasets using pandas merge and created a single dataframe.

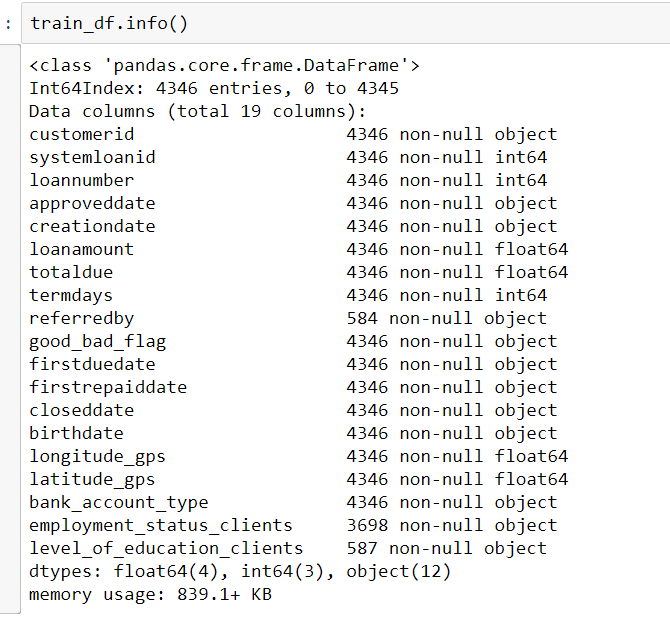
Merged dataset:



I looked at the descriptive statistics of the dataset and is shown below



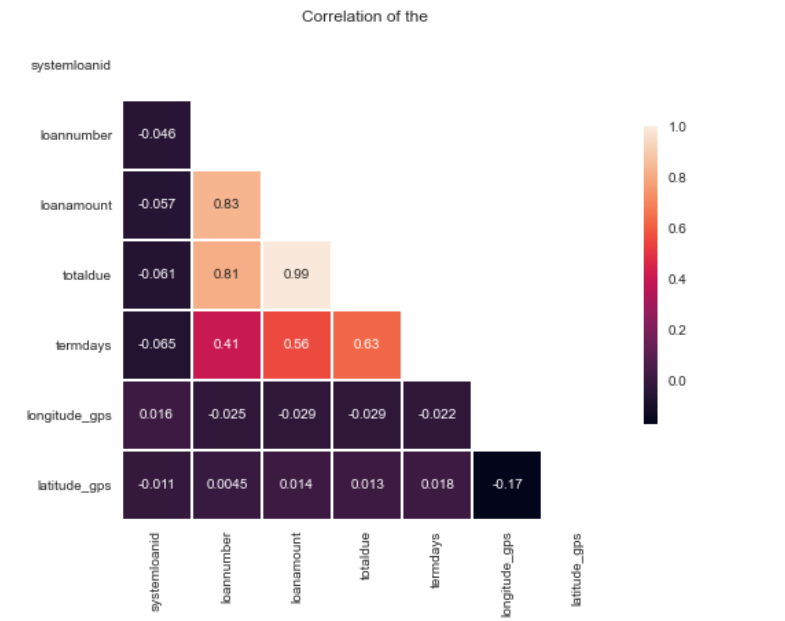
Training merged dataframe info:



I also converted datetime fields to Date datatype from Object datatype.

### Correlation matrix

I drew the correlation matrix among the variables to see what features are correlated.

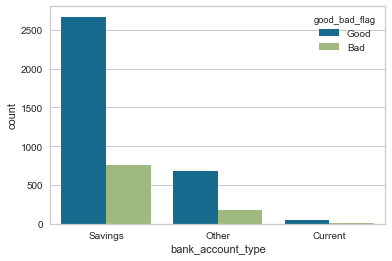


This is not telling us the correlation between features and target variable (Good\_bad\_flag). Because target variable is non-numeric. Hence this correlation matrix is not helping us much. We need to dig further to derive more insights.

### Exploratory Data Analysis

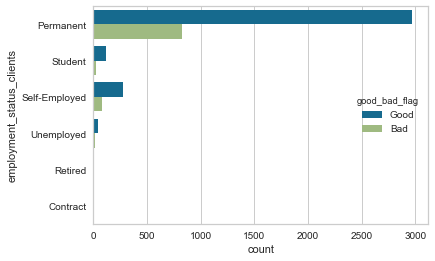
I did exploratory analysis on the data and transformed some of the features along the way.

As part of my research questions, I wanted to see the split of Good/Bad loans flag ration by the type of accounts. Below is the bar chart for Good\_Bad\_Flag by Bank account type.



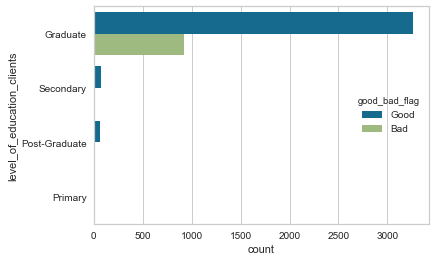
As you can see, there are more Savings account holders in the dataset and most of them are in good standing. More defaults are occurred on Savings accounts due to their high population.

Next, I wanted to see the good/bad accounts by employment status.



Parmanent employees have more bank accounts than other type and obviously they have more bad loans compared to all others.

Similarly, I also got the good/bad loans by Education type.



After the exploratory data analysis and data preparatory steps, I transformed the data to be able to build predictive model for the question.

### Test/Train data split

I already got training and testing datasets separately from the data source. So, I didn’t have to split the datasets.

## Modeling and evaluation

The objective is to predict the new customers who may likely to default if the loan is approved.

I used Random forest, Gradient boosting models to determine the loan default risk. I chose Random forest because of its better accuracy, robustness and easy of use. Gradient boosting being a greedy algorithm can fit the training dataset quickly. Since I got a good enough testing dataset, I want to see how Gradient boosting performs in predicting the target variable values

I will talk more about modeling and evaluation in my next week’s write up and presentation.

## References

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