|  |  |
| --- | --- |
| A drawing of a face  Description automatically generated  BDA 105-Spark Project Report | **Analysis of customers who are capable of repaying the loan**  **Domain-Finance** |

By - 14th December 2019

|  |  |
| --- | --- |
| Harshida Jegadeesh Chandar | jegadeeh@mcmaster.ca |
|  |  |

**Table of Contents**

* Introduction
* Body of knowledge
* Procedure
* Analysis of data
* Result
* Conclusion

# **Introduction**

# The dataset given is provided by a financial bank which is trying to improvise the service to its clients. This dataset includes telco and transactional information of the customers. As a first step towards this the bank wants to extend its service to the deserving population of customers and to filter clients who are capable of repaying the loan.

# The data contains 37 features and 307,497 rows which are customers records. The feature target is the one that has to be focussed on. It contains two categories 0 and 1. Here 1 represents the clients who have had difficulties in repayment of loans which means they have made late payments.0 represents the clients other than the clients in 1.Out the this population customers who are actually capable of repaying the loan have to be filtered. Out of 37 features 9 features are categorical and rest of them are numerical. The loan dataset contains 3 major dimensions –

# **amt\_credit** is the loan amount taken from bank by the customer

* **amt\_annuity** is the amount which is repaid by the customer every year
* **amt\_income\_total** is the income of the income of the customer annually

These columns are the most important ones that will be used for analysis.

# **Body of knowledge**

# The dataset was checked for duplicate values and null values. There were neither duplicate nor null values. Except 5 features all other features have different categories or levels. The 5 features that do not have categories are amt\_annuity, amt\_credit, amt\_income\_total, days\_birth and days\_employed. There is one more level in code\_gender ‘XNA’ but this does not seem to have any impact on the analysis as there are only 4 such customers.

* flag\_own\_car, flag\_own\_realty have 2 levels (Yes, No)
* All flag\_documents have 2 levels (Given, Not given)
* Cnt\_children (1-19)
* Name\_income\_type 8 levels
* Name\_education\_type, name\_family\_status 5 levels
* Name\_housing\_type 6 levels
* Organization\_type 58 levels
* Flag\_phone features 2 levels
* Region\_client features 3 levels
* Reg\_region features 2 levels
* Cnt\_fanily\_members (1-20)

# **Procedure**

# The dataset was imported as such into Scala IDE after cleansing for analysis and the dataset was imported again after mapping the categorical columns to numerical. This is done just to check the correctness of the dataset by building a model and checking its accuracy

# Once after importing the data with the help of Spark core techniques it was analysed using SQL queries.

* As it is difficult to read and understand thousands of rows of data and since any kind of analysis involves visualisation the output of the SQL queries was visualized using Pyspark as it is handy for visualizations
* Spark configurations in Scala IDE were given in jupyter notebook for configuration purposes
* The data was pulled into the notebook using Sparksession and it was registered as a table so that it is easier to query from the table
* The SQL queries were written in jupyter notebook and the resulting output dataset was converted to data frame and then visualised

**Analysis**

# Created data class for the dataset to be imported

# A screenshot of a social media post Description automatically generated

# A screenshot of a social media post Description automatically generated

Imported the data from local system

# A screenshot of a social media post Description automatically generatedA screenshot of a social media post Description automatically generated

The count of customers in category 1 is 8% of the total population. Since these customers are the ones incapable of repaying the loans, based on the pattern of customers in 1 we will be looking for the same pattern in 0 so that we can filter them out as the customers incapable of repaying and rest of the population as capable of repaying. This is visualized using pyspark.

# A screenshot of a cell phone Description automatically generatedA screenshot of a cell phone Description automatically generated

A screenshot of a cell phone

Description automatically generatedSeeing the count of each feature in target 1 and 0. We clearly see that female count is higher. Count of customers having education qualification secondary/secondary special and married customers are higher.

# A screenshot of a social media post Description automatically generatedA screenshot of a cell phone Description automatically generated**A screenshot of a cell phone Description automatically generatedDiscovering the pattern of customers in target 0 and 1**

# We see the relationship between the credit average vs amt\_annuity in first image and income\_avg vs days\_employed in second image. Seems like a linear relationship which means as the loan amount increases the amount to return also increases. But the pattern of customers in 1 and 0 are same in both the images which means this cannot be leveraged to differentiate the types of customers.

So the ratios between amt\_credit, amt\_income\_total and amt\_annuity is taken as a base to determine the customer who can or who cannot repay the loan

# **amt\_credit/amt\_annuity** gives the number of years to return, which means the if the number of years is higher, they will not be favourable customers

* **amt\_income\_total/amt\_credit** - If the ratio is greater than 1 it means the income of the customer is higher than the credit, he/she has taken which means it is easier for him to pay
* A screenshot of a social media post

  Description automatically generated**amt\_income\_total/amt\_annuity** - If the ratio is greater than 1 it means the income of the customer is higher than the annuity it is easier for him/her to return

Looks like customers having academic degree take a greater number of years to return while customers having **Secondary/secondary** **special** take lesser number of years to return who are favourable for the bank.

# A screenshot of a social media post Description automatically generated

This represents that customers having income type maternity leave and unemployed take a greater number of years to return whereas customers who are **working** can repay the loans as they have a regular source of income ultimately who are favourable customers

# A screenshot of a social media post Description automatically generated

The ratio of annuity/income for customers who are **pensioners and working** are greater than 1 which means they can easily repay their loans as their incomes are more than the amount that they have to return

# A screenshot of a social media post Description automatically generated

# All the flag\_document columns were added to check customers who have given 1 ,2 and 0 documents. This shows that customers who have given at least **2 documents** re more reliable as their income to credit ratio is higher.

All other Flag\_ phone and region columns were checked in the same manner and they did not show any commendable impact in the analysis.

A screenshot of a social media post

Description automatically generated

# Again, from any dimension it shows that people having education background of secondary/secondary special have higher chanced of repaying the loans as their income is more than credit. Higher secondary comes next.

A screenshot of a social media post

Description automatically generated

# People from organization type **Business entity** have higher chances of repaying the loan as their count is highest in the population for having a greater ratio

A screenshot of a cell phone

Description automatically generated

Yet again **working**-class people are higher in number who have greater income to credit ratio

# A screenshot of a cell phone Description automatically generated

A screenshot of a social media post

Description automatically generatedCustomers who **own a house** are more in number who have higher income t credit ration rather the ones who don’t.

# A screenshot of a social media post Description automatically generated

# Similarly, customers who are **married and living in their own House/apartment** are better off

A screenshot of a social media post

Description automatically generated

Customers who have been employed from 0 to 5000 days from the day of application have higher Income to credit ratio which means they are favourable. It slowly drifts off for customers above 5000 and stoops down after 7500. People who have retired have 365243 days as working days which is not the major population and do not have any effect on the analysis

A screenshot of a cell phone

Description automatically generated

People who have days\_birth from **12500 to 20000 days** are higher in count who are having Income more than credit. Which means people of age range **34 – 54 years** are better off at repaying loans as their income is more than credit.

# On comparing the output based on amt\_income\_total/amt\_annuity ratio and amt\_income\_total/amt\_credit ratio, output remained the same. For e.g.

A screenshot of a social media post

Description automatically generatedA screenshot of a social media post

Description automatically generated

# Again, working and secondary/secondary special customers have performed better.

# **Conclusion**

# Based on the analysis above the data set was filtered by categories of data that performed better.

This gave an output of 79,406 rows which means 79,406 customers are capable of repaying the loan based on their educational qualification, income type, age, days employed.

# A screenshot of a cell phone Description automatically generated

# All the other criteria got sorted to the best performing categories once they were filtered by giving the above condition.

Hence 79,406 is **25.8%** of the total population of customers is certainly capable of repaying loans