**Data Science Project Report – Analysing Loan Approval Statistics**

By Ilia Rafi Samii

Project Overview

In this project, my main goal is to analyse a dataset containing loan information from both approved and declined loans, and determining the likelihood based on this data as to whether an individual is approved for or denied from receiving a loan. To achieve this main objective, I have analysed a variety of key statistics related to credit scores, risk scores, and accounting related data such as income, debt-to-income ratio, current ratio as well as the ratio of monthly expenses to monthly revenue. After analysing the results, I verified that credit and risk scores are a good indicator as to whether an individual is approved for a loan. In addition, approved loans were surprisingly associated with a slightly higher debt-to-income ratio as compared to declined loans. As expected, however, the declined loans had a much higher percentage of monthly expenses as compared to the approved loans.

Concept Overview

When taking a loan from a bank, it is not difficult to realise that credit and risk scores play a very important role in the bank’s decision to grant you a loan. Intuitively, we can understand that higher credit scores make individuals more desirable to grant loans to. That’s why it is important to take credit scores into consideration when doing our analysis. Other statistics, such as the debt-to-income ratio, also give us a good indication to the probability of a bank approving a loan. If an individual has a low debt-to-income ratio, it indicates that the individual has enough income to compensate for his or her liabilities and would therefore be a more ideal candidate for a loan approval. Another good indicator would be an individual’s equity. A person can sell his or her assets in order to generate enough income to pay off their debt, but if the person has too many liabilities, it isn’t a good sign. From my basic knowledge of accounting, I know that a person’s equity is equal to his or her assets minus his or her liabilities. Therefore, the higher an individual’s equity, the better it is for his or her credibility in terms of receiving a loan.

Data Overview and Preparation

This project uses SQL, python and PowerBI to explore and visualize data. SQL was utilized to create tables from the initial CSV file to better organize and manage the data. A total of three tables were created: approved\_loans (containing the loan application date, age, annual income, credit score and risk score of the applicant), declined\_loans (containing the same info as the approved\_loans), and loans\_accountingdata (containing information such as the assets, liabilities, debt, net worth and income of the applicant). By creating these tables, the data is organized based on the important key factors that determine whether the applicant will be approved for a loan.

After organizing and filtering the data using SQL, I used python to calculate key statistics including the mean and variance of credit and risk scores for both approved and declined loans. After converting my SQL tables into CSV files, I used the data from the “loans\_accountingdata” table to implement a linear regression algorithm that predicts whether an applicant will be approved or denied a loan.

Finally, I took advantage of PowerBI features to create data visualizations. Looking at these visualizations, we can see that there is a negative correlation between credit scores and risk scores, meaning that the higher an individual’s credit score, the lower his/her risk score. We can also see that the bank is more likely to grant a loan to an individual that has a low proportion of liabilities and debt as compared to his or her assets and income.

Here is a sample of the SQL and python code used to create the tables and calculate key statistics:

# IMPORT LIBRARIES AND CONNECT TO THE DATABASE FILE

import sqlite3

import pandas as pd

df = pd.read\_csv('Loan.csv')

conn = sqlite3.connect('Loan.db')

cursor= conn.cursor()

# CREATE TABLE FOR LOANS

table\_name = 'loans'

columns = ", ".join([f"{col.replace(' ', '\_')}" for col in df.columns])

create\_table\_query = f"CREATE TABLE IF NOT EXISTS {table\_name} ({columns});"

cursor.execute(create\_table\_query)

cursor.fetchall()

# UPDATE TABLE WITH THE DATA FROM Loan.csv FILE

for index, row in df.iterrows():

values = ", ".join([f'"{row\_item}"' for row\_item in row])

insert\_sql = f"INSERT INTO {table\_name} ({', '.join(df.columns.str.replace(' ', '\_'))}) VALUES ({values})"

cursor.execute(insert\_sql)

# SUBSET OF DATA CONTAINING CREDIT AND RISK SCORE FOR APPROVED LOANS

query = "SELECT ApplicationDate, Age, AnnualIncome, CreditScore, RiskScore, LoanApproved FROM loans WHERE LoanApproved = '1'"

cr1 = pd.read\_sql\_query(query,conn)

cr1.to\_csv("creditrisk\_loanapproved.csv",index = False)

# CREATE TABLE FOR APPROVED LOANS

table\_name = 'approved\_loans'

columns = ", ".join([f"{col.replace(' ', '\_')}" for col in cr1.columns]) # Eg: page\_id TEXT, name TEXT, urslug TEXT, ...

create\_table\_query = f"CREATE TABLE IF NOT EXISTS {table\_name} ({columns});"

cursor.execute(create\_table\_query)

cursor.fetchall()

# UPDATE TABLE WITH THE DATA FROM creditrisk\_loanapproved.csv FILE

for index, row in cr1.iterrows():

values = ", ".join([f'"{row\_item}"' for row\_item in row])

insert\_sql = f"INSERT INTO {table\_name} ({', '.join(cr1.columns.str.replace(' ', '\_'))}) VALUES ({values})"

cursor.execute(insert\_sql)

# SUBSET OF DATA CONTAINING CREDIT AND RISK SCORE FOR DECLINED LOANS

query = "SELECT ApplicationDate, Age, AnnualIncome, CreditScore, RiskScore, LoanApproved FROM loans WHERE LoanApproved = '0'"

cr0 = pd.read\_sql\_query(query,conn)

cr0.to\_csv("creditrisk\_loandeclined.csv",index = False)

# CREATE TABLE FOR APPROVED LOANS

cursor.execute('''

CREATE TABLE IF NOT EXISTS declined\_loans (

ApplicationDate TEXT,

Age INTEGER,

AnnualIncome REAL,

CreditScore INTEGER,

RiskScore REAL,

LoanApproved INTEGER

);

''')

# CALCULATING IMPORTANT STATISTICS FOR APPROVED LOANS

import numpy as np

import math

cr1 = pd.read\_csv("creditrisk\_loanapproved.csv")

# CREDIT SCORE

# Average credit score

mean\_cs\_approved = np.mean(cr1['CreditScore'])

#print(mean\_cs\_approved)

# Variance of credit scores

var\_cs\_approved = np.var(cr1['CreditScore'])

#print(var\_cs\_approved)

# 95% Confidence Interval for credit scores

lowerlimit\_cs\_approved = mean\_cs\_approved - 0.95 \* math.sqrt(var\_cs\_approved / 20000)

upperlimit\_cs\_approved = mean\_cs\_approved + 0.95 \* math.sqrt(var\_cs\_approved / 20000)

print(lowerlimit\_cs\_approved)

print(upperlimit\_cs\_approved)

print(f'We can say with 95% confidence that credit scores fall in the range between {lowerlimit\_cs\_approved} and {upperlimit\_cs\_approved}')

# RISK SCORE

# Average credit score

mean\_rs\_approved = np.mean(cr1['RiskScore'])

#print(mean\_rs\_approved)

# Variance of credit scores

var\_rs\_approved = np.var(cr1['RiskScore'])

#print(var\_rs\_approved)

# 95% Confidence Interval for risk scores

lowerlimit\_rs\_approved = mean\_rs\_approved - 0.95 \* math.sqrt(var\_rs\_approved / 20000)

upperlimit\_rs\_approved = mean\_rs\_approved + 0.95 \* math.sqrt(var\_rs\_approved / 20000)

print(lowerlimit\_rs\_approved)

print(upperlimit\_rs\_approved)

print(f'We can say with 95% confidence that credit scores fall in the range between {lowerlimit\_rs\_approved} and {upperlimit\_rs\_approved}')

# CALCULATING IMPORTANT STATISTICS FOR DECLINED LOANS

import numpy as np

import math

cr0 = pd.read\_csv("creditrisk\_loandeclined.csv")

# CREDIT SCORE

# Average credit score

mean\_cs\_declined = np.mean(cr0['CreditScore'])

#print(mean\_cs\_approved)

# Variance of credit scores

var\_cs\_declined = np.var(cr0['CreditScore'])

#print(var\_cs\_approved)

# 95% Confidence Interval for credit scores

lowerlimit\_cs\_declined = mean\_cs\_declined - 0.95 \* math.sqrt(var\_cs\_declined / 20000)

upperlimit\_cs\_declined = mean\_cs\_declined + 0.95 \* math.sqrt(var\_cs\_declined / 20000)

print(lowerlimit\_cs\_declined)

print(upperlimit\_cs\_declined)

print(f'We can say with 95% confidence that credit scores fall in the range between {lowerlimit\_cs\_declined} and {upperlimit\_cs\_declined}')

# RISK SCORE

# Average credit score

mean\_rs\_declined = np.mean(cr0['RiskScore'])

#print(mean\_rs\_approved)

# Variance of credit scores

var\_rs\_declined = np.var(cr0['RiskScore'])

#print(var\_rs\_approved)

# 95% Confidence Interval for risk scores

lowerlimit\_rs\_declined = mean\_rs\_declined - 0.95 \* math.sqrt(var\_rs\_declined / 20000)

upperlimit\_rs\_declined = mean\_rs\_declined + 0.95 \* math.sqrt(var\_rs\_declined / 20000)

print(lowerlimit\_rs\_declined)

print(upperlimit\_rs\_declined)

print(f'We can say with 95% confidence that credit scores fall in the range between {lowerlimit\_rs\_declined} and {upperlimit\_rs\_declined}')

Here are some of the most important visualizations generated using PowerBI:

A blue and red line

Description automatically generated

The above scatter plot shows the correlation between credit scores and risk scores.

A graph of blue and orange bars

Description automatically generated

The above chart shows the proportion of debt to income for an individual with an **approved loan**.

A graph of a graph showing a number of debt payments

Description automatically generated with medium confidence

The above chart shows the proportion of debt to income for an individual with a **declined loan**.

A pie chart with numbers and a few words

Description automatically generated with medium confidence

The pie chart above indicates applicants’ purpose for requesting a loan. As we can see from the chart, the highest percentage of applicants needed a loan for debt consolidation purposes.

Analysis of the Results

As I anticipated, the average credit score for approved loans (584.5341) turned out to be significantly higher than the average credit score for declined loans (567.5542). Based on this, we can predict whether an individual is approved for a loan depending on whether his or her credit score is closer to the average credit score for approved loans or the average credit score for declined loans. Our results also show that the average risk score is much higher for declined loans (54.1063) than it is for approved loans (40.1334). We can say with 95% confidence that credit scores for approved loans will fall between 584.20453 and 584.86367. Additionally, we can say that, with 95% confidence, risk scores for approved loans will roughly fall between 40.10733 and 40.15945. Looking at the total debt-to-income ratios, we can see that on average it is slightly higher for approved loans (0.285744) as compared to declined loans (0.285732). While this is a little surprising, it tells us that the total debt on average makes up about 28.5744% of total income for individuals with approved loans and it makes up about 28.5732% of total income, on average, for individuals who were denied loans. On the other hand, the ratio of monthly expenses to monthly revenue for approved loans turned out to be a lot smaller for approved loans (0.13189) than it was for declined loans (0.38273). This is a good sign, as it indicates that monthly expenses only make up approximately 13.2% of monthly income for approved loan holders, whereas it makes up about 38.3% of monthly income for those who were denied a loan.