Data Science Portofolio

Satria Fajar Bagaskara
Dibimbing Data Scientist Batch 18

linkedin.com/in/satriabagaskara
github.com/dataaga
medium.com/@dataaga

#Tentang Saya

Aga

In previous role, dedicated to implementing policy of infrastructure financing for public-private partnership projects in Indonesia for over 3 years of working experience. Policy Analyst transitioning into Data Scientist role after completed Dibimbing Data Science Bootcamp with hands-on skills in SQL queries, data analytics, data storytelling, data wrangling, modeling and deployment. Passionate about provide a deeper understanding problem and translating to insights for decision making process in e-commerce and fintech industry. Have an excellent interpersonal skill, agile, eager to learn new things and grow more.

Education

Diponegoro University
Bachelor of Engineering / Civil Engineering
July 2013 - Des 2018

Dibimbing Bootcamp (Batch 18)
Data Science
Des 2022 - July 2023

Certificate

Data Science
Dibimbing.id (July 2023)

Associate Data Scientist
Data Camp (January 2023)

Data Scientist Professional
IBM (April 2023)

WORK EXPERIENCE

Public Policy Analyst

Feb 2020 - Present

Ministry of Public Works and Housing
Directorate General of Infrastructure Financing for Public Works

Site Engineer

Feb - April 2019

PT Pilaren
Construction Project of CPO Jetty

Engineer Intern

Nov 2015 - Jan 2016

PT PP (Persero) Tbk.
Construction Project of Manhattan Mall & Condominium

Skills

Software

Jupyter · Google Colab · Tableau · Looker Studio · MySQL · Docker · Streamlit

Library

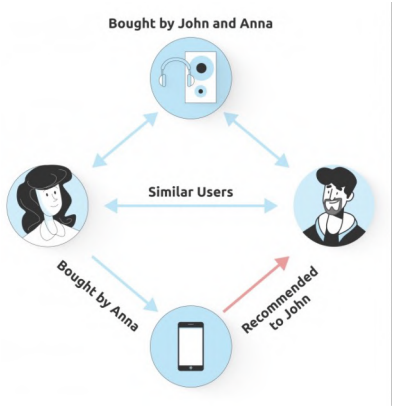
Plotly · Folium · Seaborn · Surprise · Scikit-learn · Pandas · Numpy · Matplotlib · BeautifulSoup · LightGBM · XGBoost

Programming Language

Python · SQL

Hypothesis Testing · Data Cleaning · Data Manipulation · Web Dashboard Development · Supervised Machine Learning · Unsupervised Machine Learning · Hyperparameter Tuning · Gradient Boosted Trees · Ensemble Models · Recommendation System · Regularized Regression · Data Wrangling · Logistic Regression · Data Visualization · Exploratory Data Analysis · Webscraping · Analytical Skills · Predictive Analytics

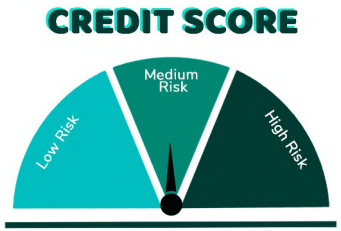
Data Science Projects



Recommendation System (Collaborative Filtering) of Brazilian E-commerce (Olist)

using preferences of *similar users* in *the past* to *get your* preference in *the future.*

github.com/dataaga/Olist_CollaborativeFiltering



Credit Score Prediction

using manual data and alternate data of the borrowers to help lenders decide whether to approve/extend or deny the borrowers

github.com/dataaga/Credit_Scoring_MultiClass

Predicting

CREDIT SCORING

Multi-Class Prediction

Poor, Standard, Good

Table of contents

01 Business Background

02 Objective

03 Data

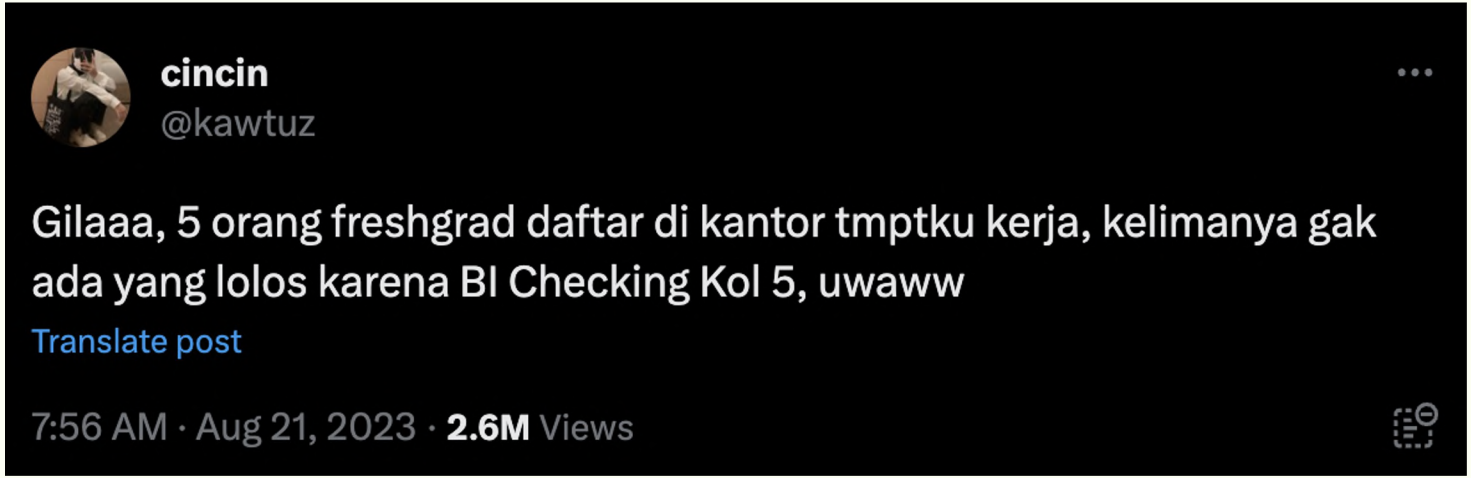
04 Exploratory Data Analysis

05 Modelling

06 Conclusion & Recommendation

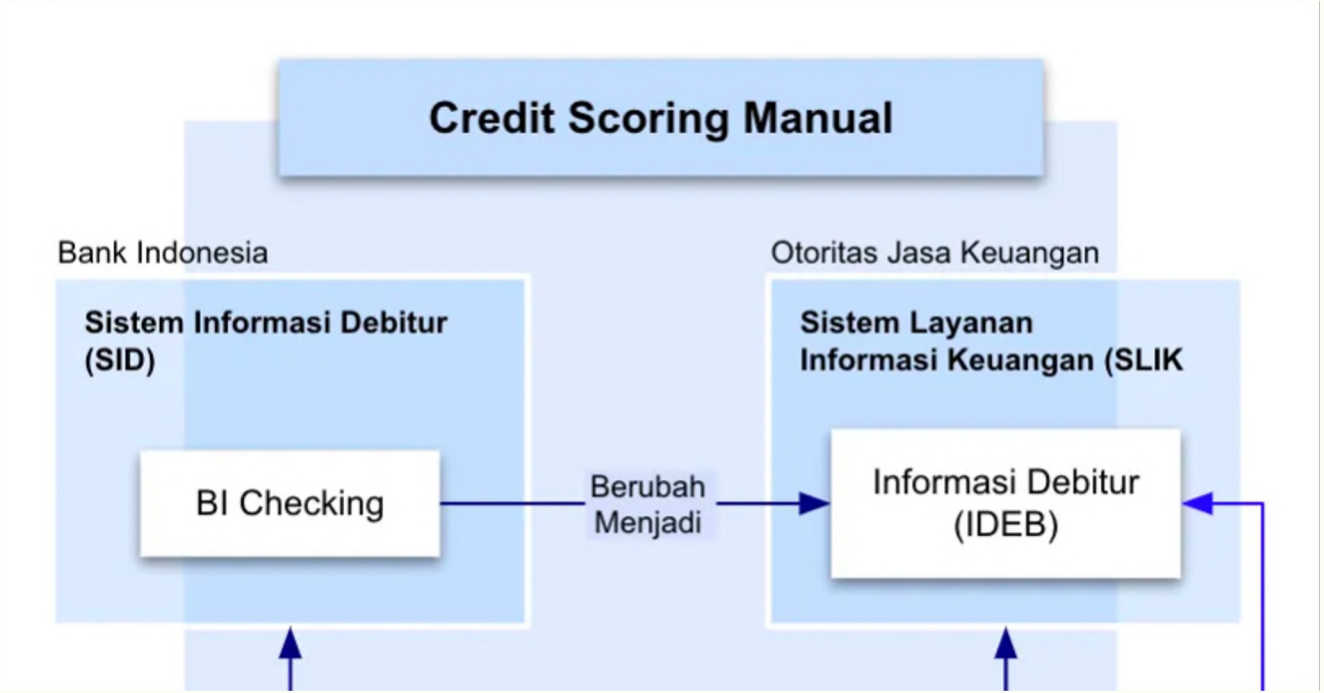
01.
BUSINESS BACKGROUND

Gabisa **kerja** karena **Ga Lolos SLIK (BI Checking)**



Source: akun twitter @kawtuz

SLIK adalah **pengganti** BI Checking (SID) untuk **Credit Scoring** (Traditional)



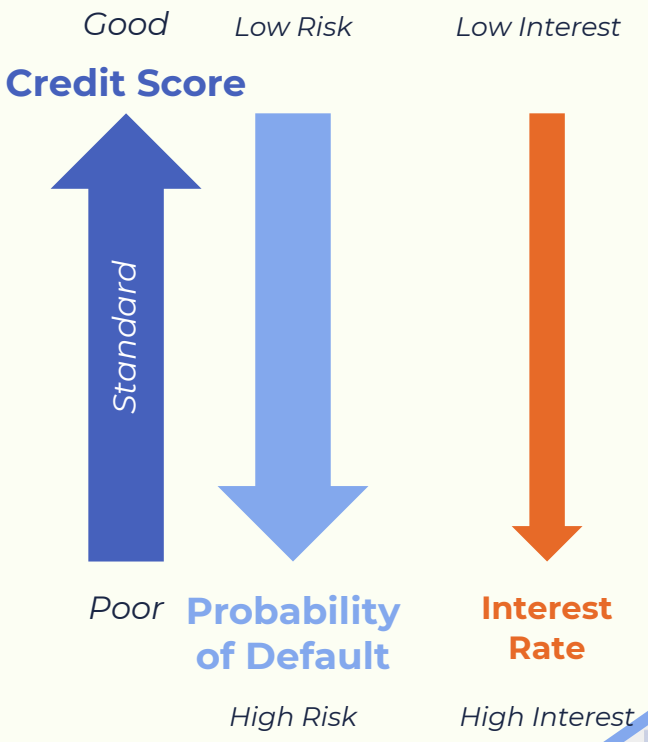
Apa itu Credit Scoring?

- Credit scoring is **assessment** of an individual's or business's **creditworthiness**, used by **lenders** to help **decide** whether to **extend** or **deny credit**.
(Investopedia)

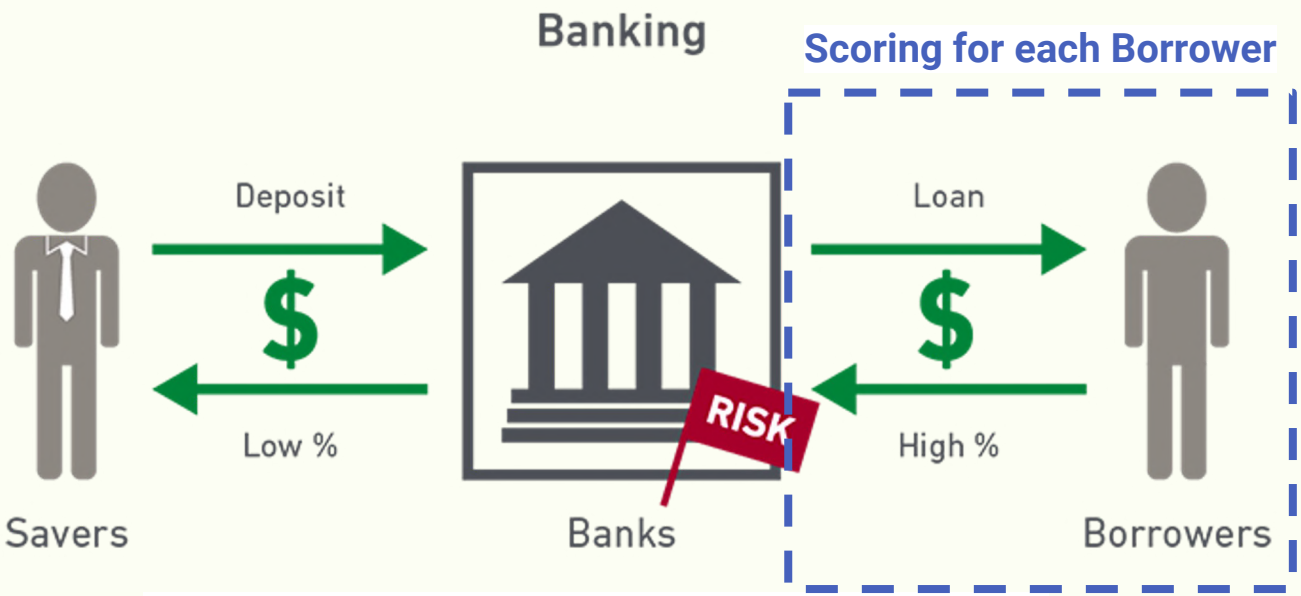
Apa itu Credit Scoring?

- Probability of Default (PD), is the likelihood that a borrower will **fail to pay back** a certain **debt**. For businesses, probability of default is reflected in the company's credit ratings. **For individuals**, a **credit score** is **one gauge** of **default risk**.

  (Investopedia)



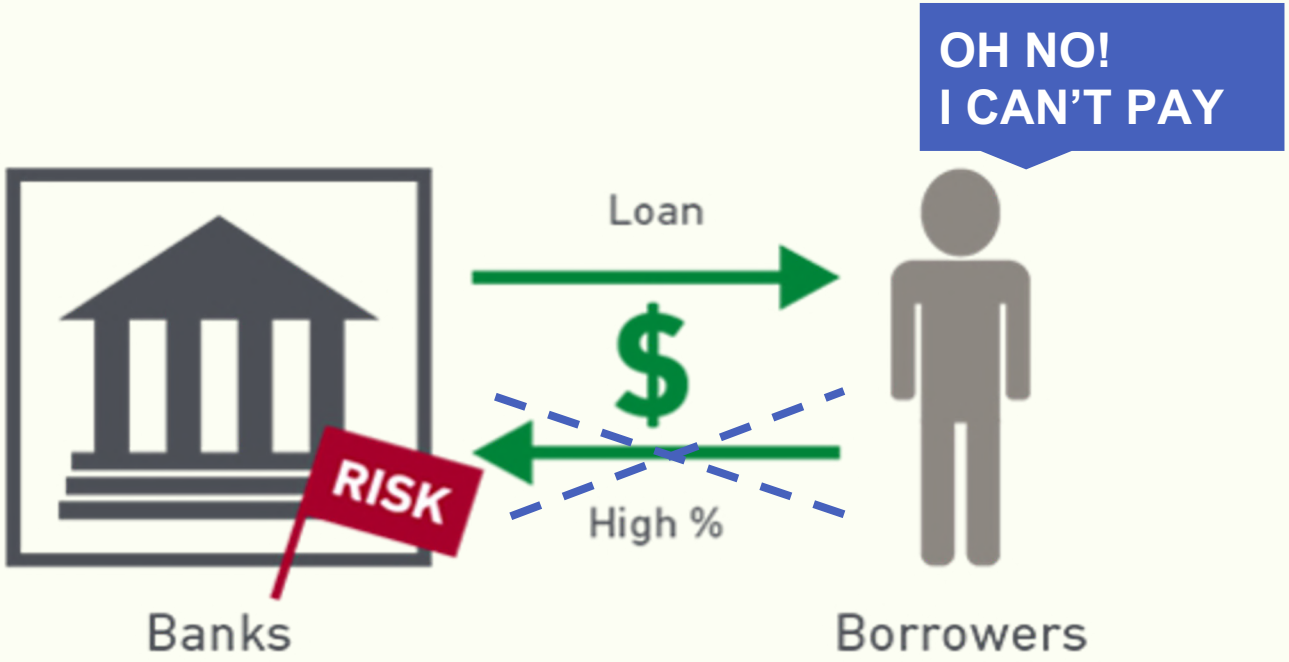
Sepenting apa Credit Scoring?



Bank Profit = Interest from Borrowers - Interest for Savers

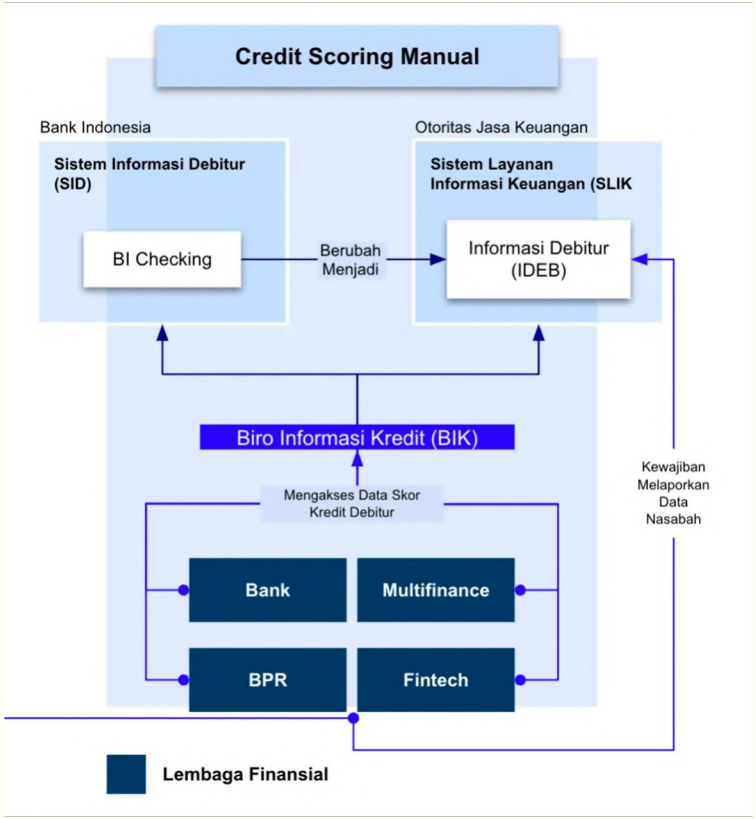
The Lender gather Deposit and to ***give Loan***

What if Borrowers can’t pay?



In order to reducing the risk,
the lender **need credit scoring**

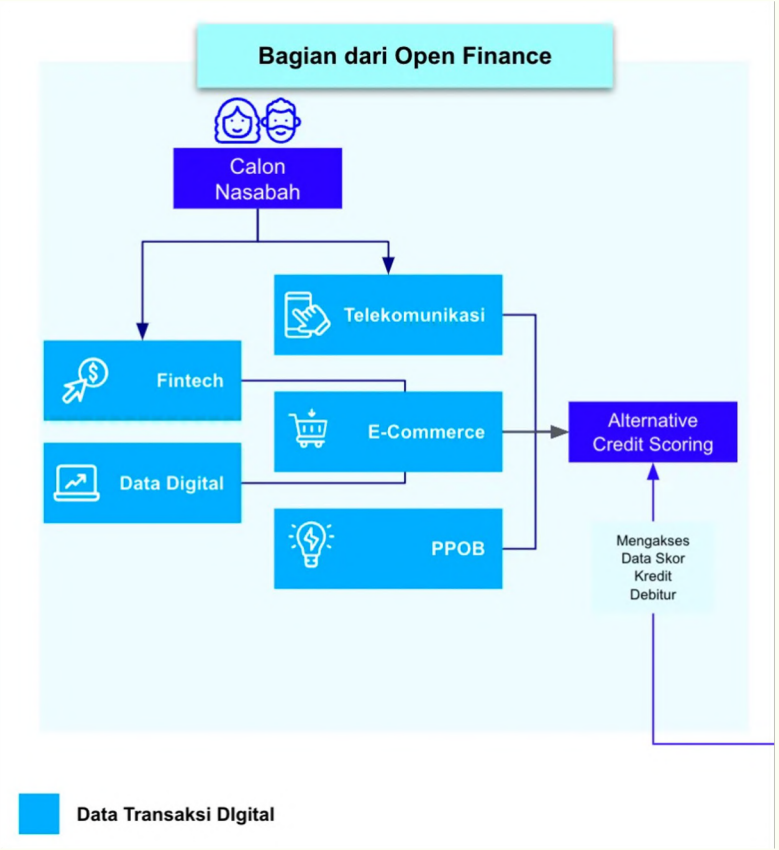
Credit Scoring Mechanism



Before ML come, credit scoring have traditional way that relies on **manual credit data** itself, which might not capture a complete picture of an individual's creditworthiness.

Source: https://id.blog.finantier.co/perkembangan-credit-scoring-indonesia-bdae9b66984

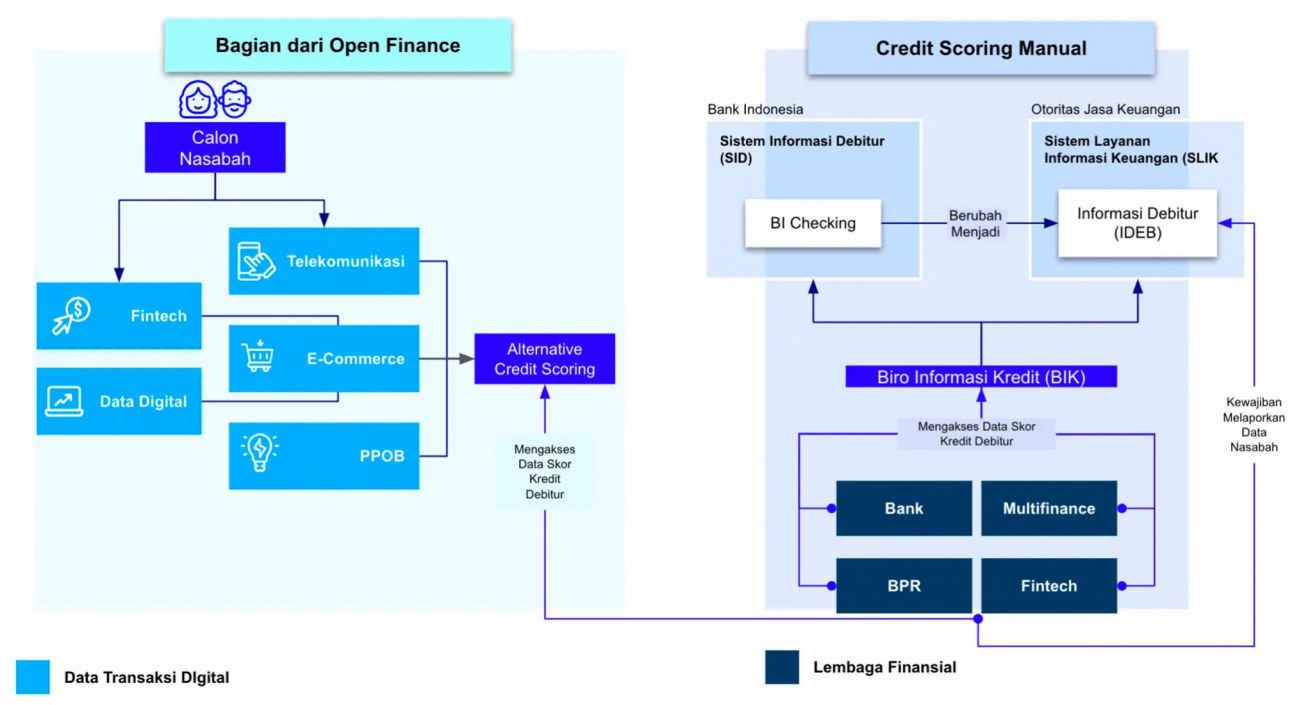
Credit Scoring Mechanism



Hence, Lender need an efficient and accurate way to decide if they gonna give loans/credit or not by add some **extra data** called **alternate data**

Source: https://id.blog.finantier.co/perkembangan-credit-scoring-indonesia-bdae9b66984

Credit Scoring Mechanism



Lenders using **Machine Learning** (ML) based on

Alternate Data + Manual (Traditional) **Credit Data**

Source: https://id.blog.finantier.co/perkembangan-credit-scoring-indonesia-bdae9b66984

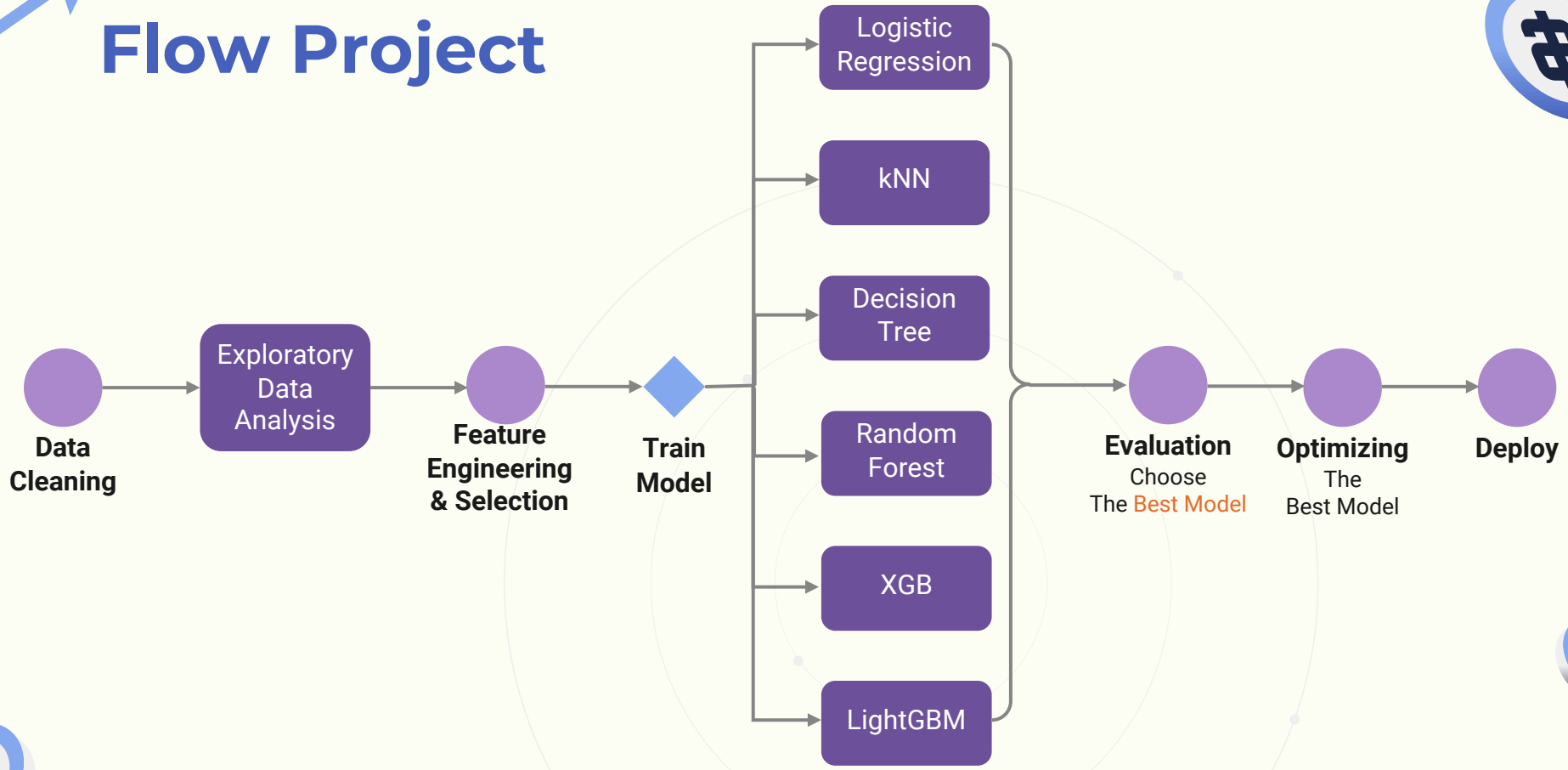
02.

OBJECTIVE

Goals
KreditDulu

"KreditDulu requires **Credit Scoring** through the utilization of **Machine Learning**, as well as **manual data** and **alternate data**, to **assist** the **lender** in **determining** whether to **extend** or **deny credit** to borrowers"

Flow Project



03. DATA

Dataset

100K rows

x

28 columns

- **17 Numerical Features**
- **6 Categorical Features**
- **4 Index Features**
- **1 Multi-Class Feature (Target)**

Dataset Source: https://www.kaggle.com/datasets/parisrohan/credit-score-classification

Manual (Credit-related) Data

Payment Behaviour/History
Amounts Owed/Credit Utilization
Length of Credit History
New Credit/Credit Inquiries
Type of Loan/Credit Mix

Alternate Data

Investing Info
Demographic
Credit Information

Features

A. Manual Data (FICO model):

1. Payment Behaviour/History:

* Payment_of_Min_Amount, Total_EMI_per_month, Payment_Behaviour: These factors can provide insights into the individual's payment behavior and financial responsibility.

2. Amounts Owed/Credit Utilization:

* Outstanding_Debt, Changed_Credit_Limit, Credit_Utilization_Ratio: These factors are important in understanding how the individual manages their credit accounts and balances.

3. Length of Credit History:

* Credit_History_months: The length of the individual's credit history can be a key factor in credit scoring.
* Delay_from_due_date, Num_of_Delayed_Payment: The delay in payment from the due date for loans or credit payments.

4. New Credit/Credit Inquiries:

* Num_Credit_Inquiries: The inquiries of credit whenever new credit account opened

5. Type of Loan/Account Mix:

* Num_of_Loan, Type_of_Loan, : These factors related type of loan are relevant for assessing credit

Features

B. Alternate Data

1. Investing History:

* Amount_invested_monthly: While investing information is not typically included in FICO scores, it could potentially provide additional insights into financial stability.

2. Demographic Information:

* Name, Age, SSN, Occupation, Annual_Income, Monthly_Inhand_Salary: FICO scores do not typically include demographic and income-related information, as they focus primarily on credit behavior.

3. Credit Information:

* Num_Bank_Accounts, Num_Credit_Card: These factors represent the number of bank accounts and credit cards held by the individual, which could be indicative of their credit behavior.
* Monthly_Balance: This factor indicates the balance maintained by the individual, which can reflect their financial stability.
* Interest_Rate: Interest rate on financial products can reflect the individual's credit risk. Also, after did labeling credit scoring to borrower, lenders will determine interest rate based on the score
* Month: The month data is queried

C. Target Variable:

* Credit_Score: Simplified into Poor, Standard and Good

04.

EXPLORATORY DATA ANALYSIS

Manual Data (Credit-related)

1. How is **Payment Behaviour** of borrowers?
2. How deep are the **Amounts Owed** (Credit Utilization)?
3. How long is the **Credit History**?
4. How many **New Credit Card** (Credit Inquiries) are there?
5. How many **Types of Loans** (Credit Mix) are there?

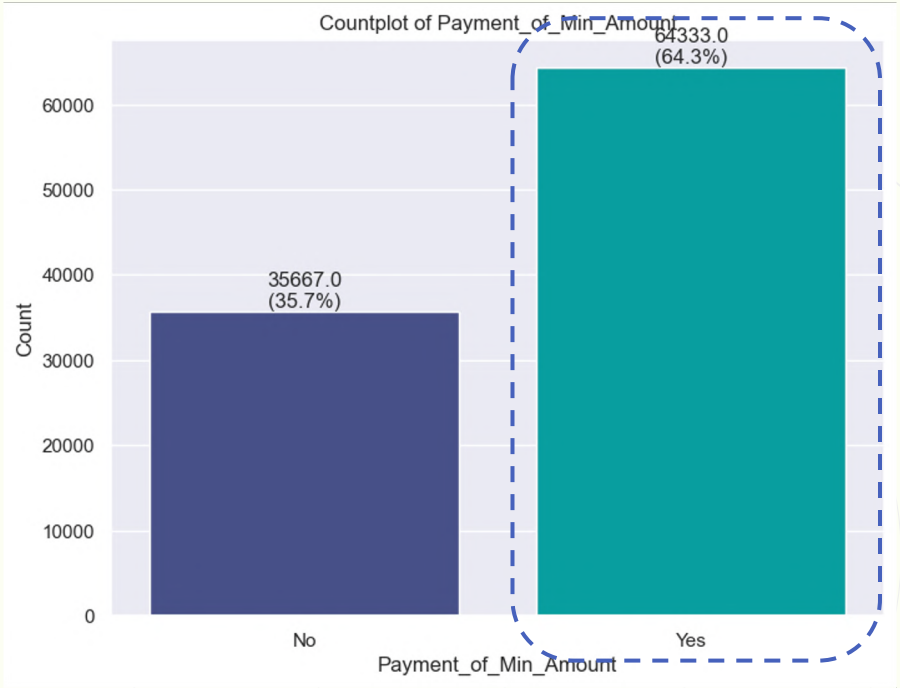
Credit Scoring

1. How to **achieve** the good score?
2. Who's the **Top 5 Good Score** (lowest interest) of borrowers ?
3. Who's the **Top 5 Poor Score** (highest interest) of borrowers ?

Manual Data
(Credit-related)

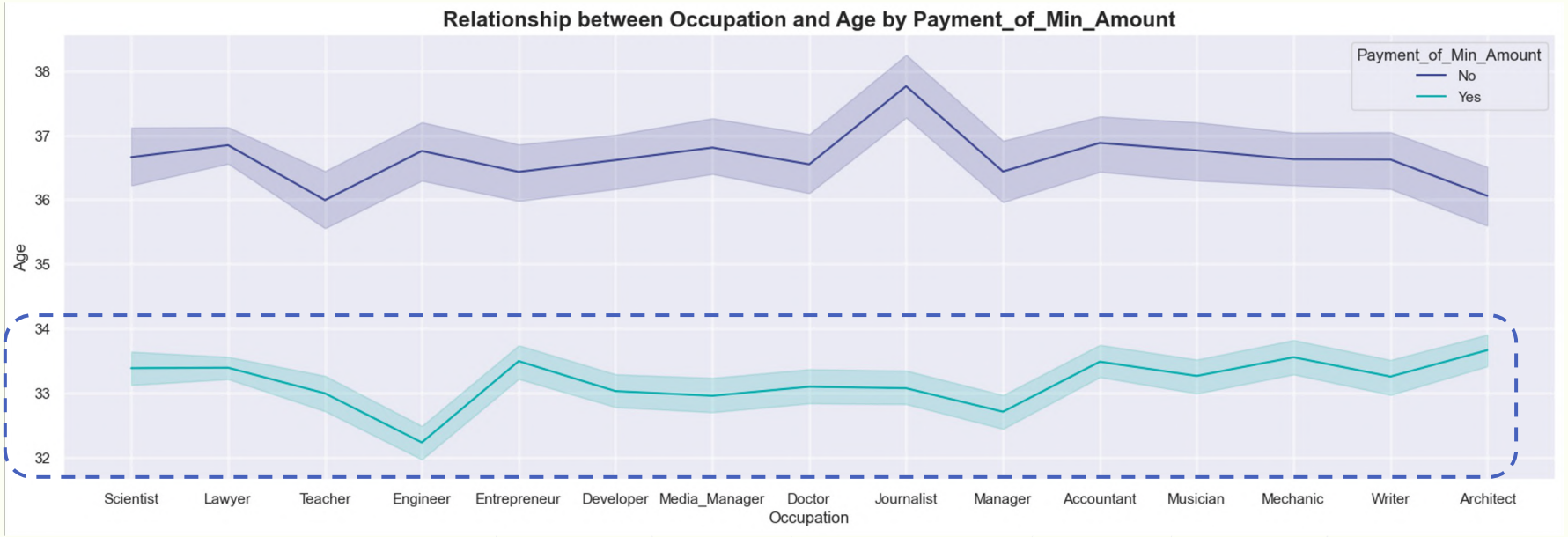
1. How good the **Payment Behaviour**?

1. Payment Behaviour



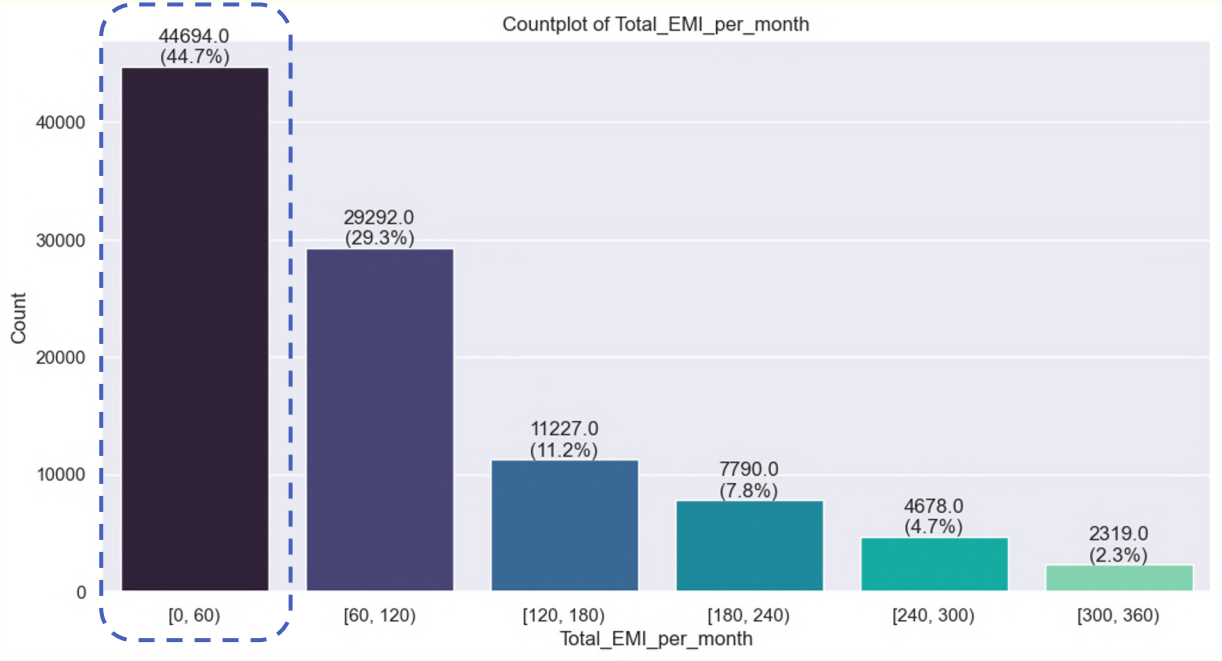
Majority of borrower did payment of minimum amount which indicates they **can't do full payment but still want to maintain a good credit score.**

1. Payment Behaviour



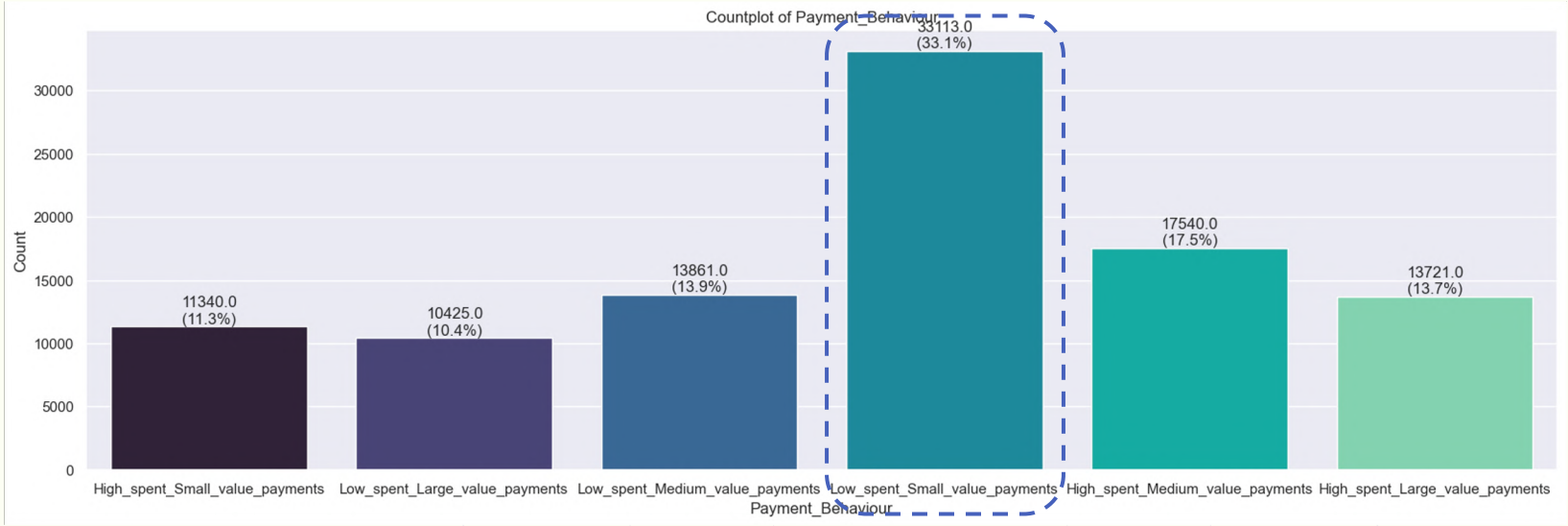
It appears that **younger borrower tend to be more willing to make payments with minimal amounts**, which indicate they are have difficulty to pay, and vice versa.

1. Payment Behaviour



Almost **half of the borrower did payment** with **small installment** (0-60 total EMI per month).

1. Payment Behaviour

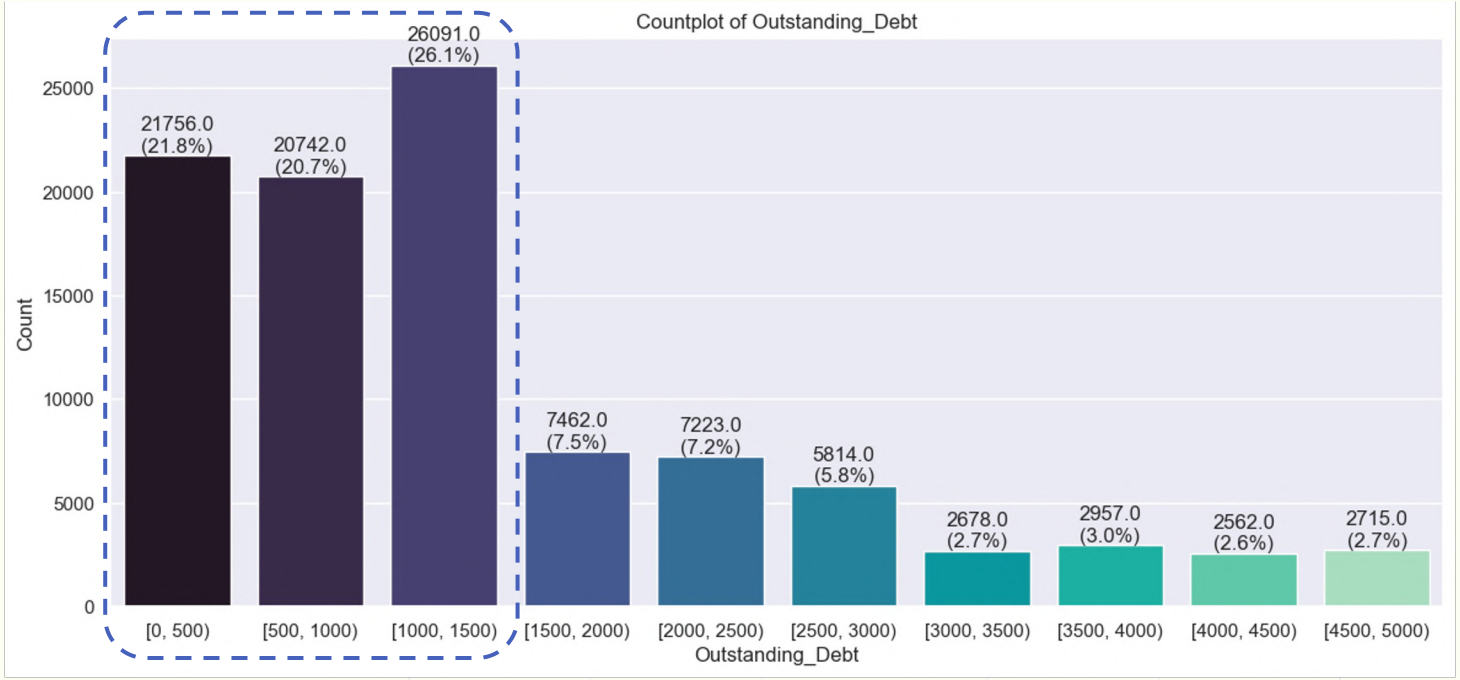


Almost ⅓ of **borrower** did **low spent** and **small payment**

Manual Data
(Credit-related)

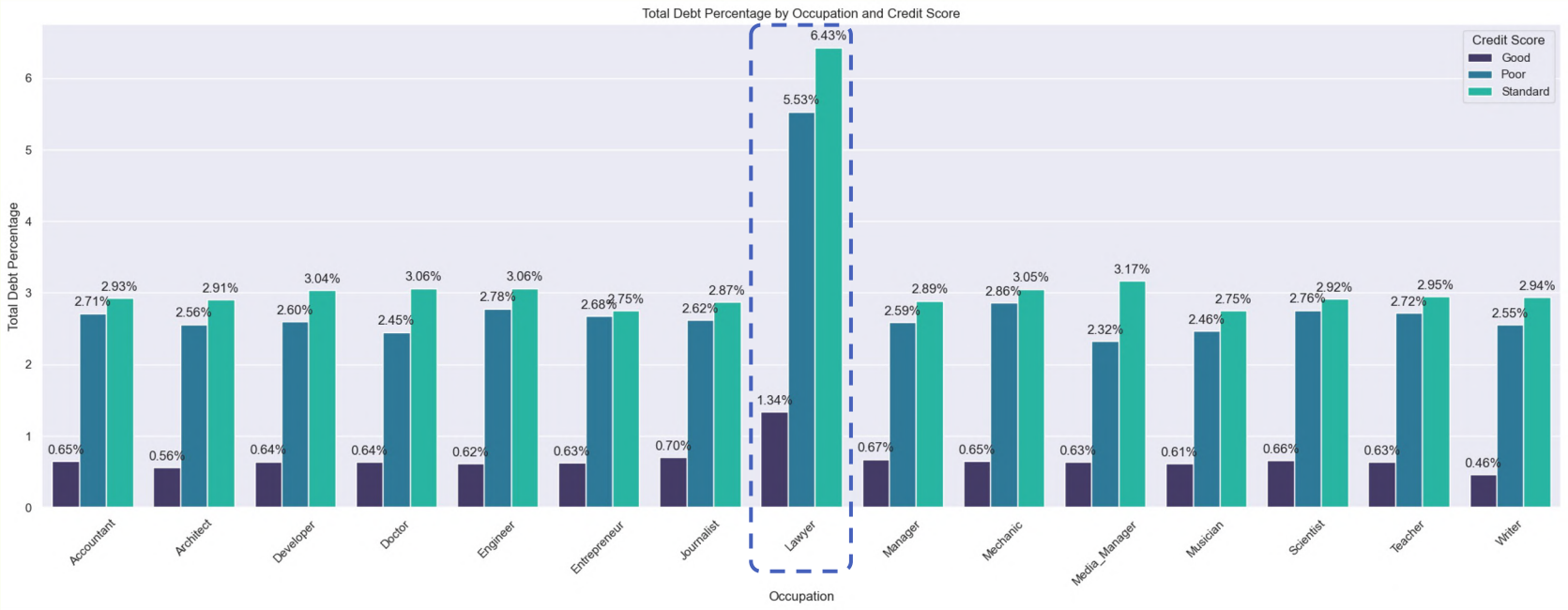
2. How deep are the **Amounts Owed** (Credit Utilization)?

2. Amounts Owed (Credit Utilization)



Almost 70% of borrower have low outstanding debt in range 0-1500 (30%), this indicate **majority have low outstanding debt.**

2. Amounts Owed (Credit Utilization)

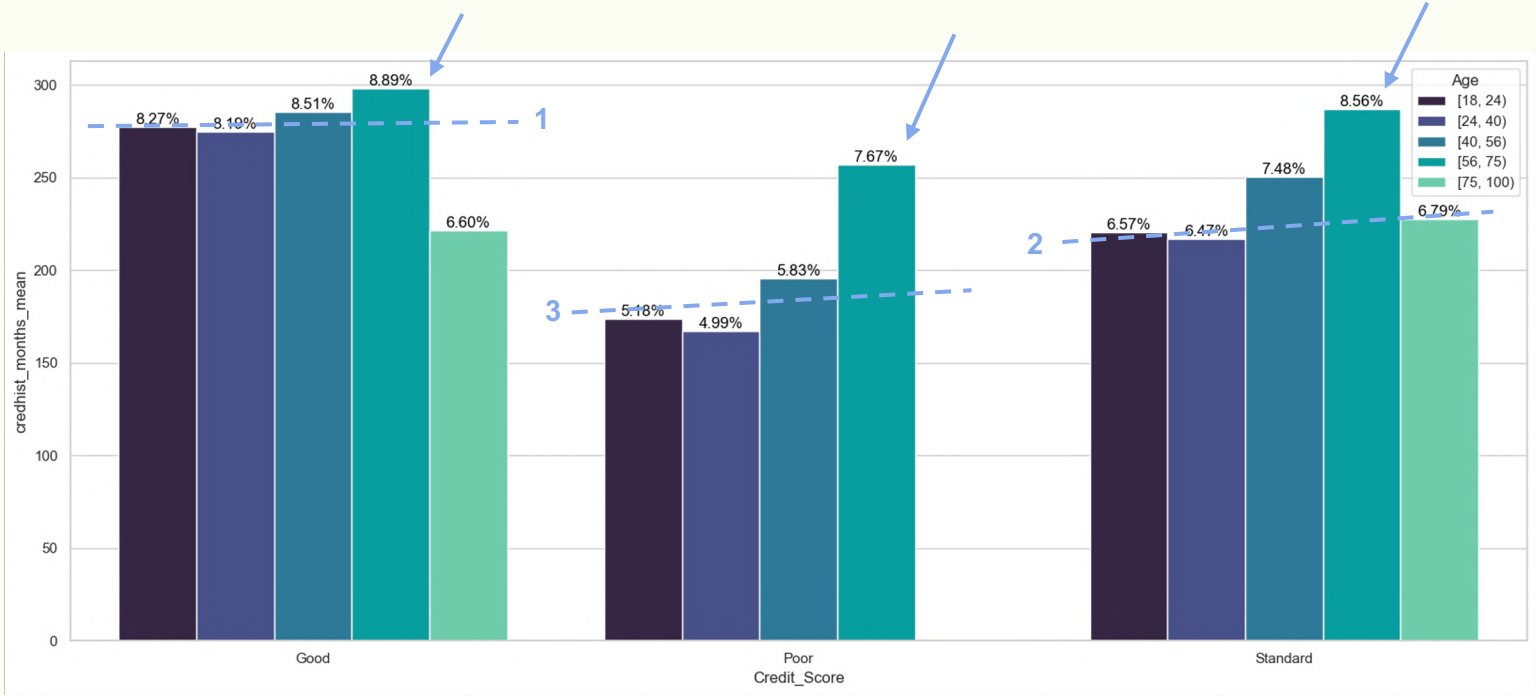


From the figure among the poor, standard, and good, **the lawyer has highest total debt** among others which **lead** them into **poor and standard score.**

Manual Data
(Credit-related)

3. How long is the **Credit History**?

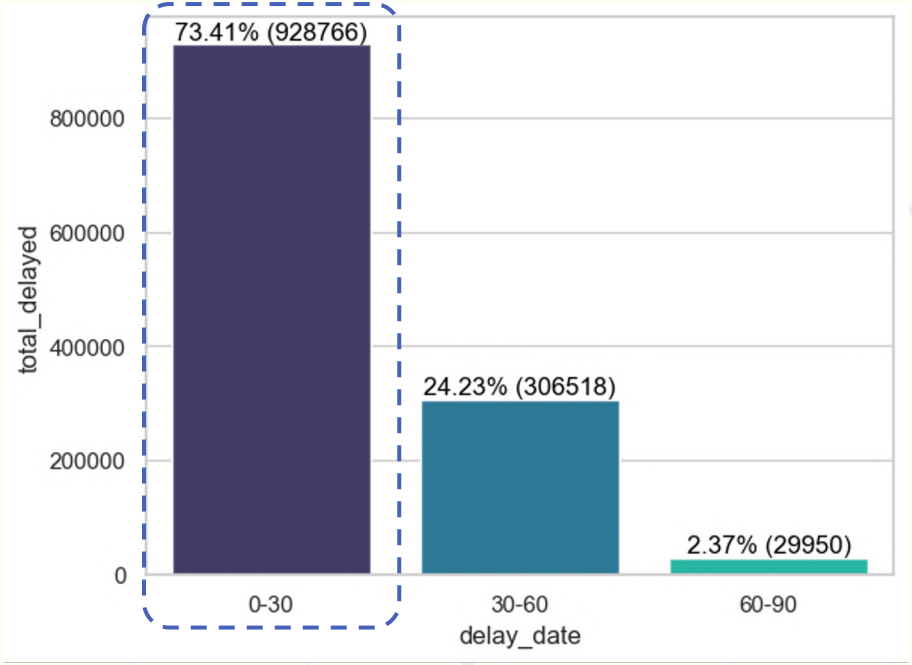
3. Credit History



Unfortunately **the age is don't have correlation to credit history** because the elder has majority among other ages.

However is it true that have **long credit history tend to have good credit score.**

3. Credit History

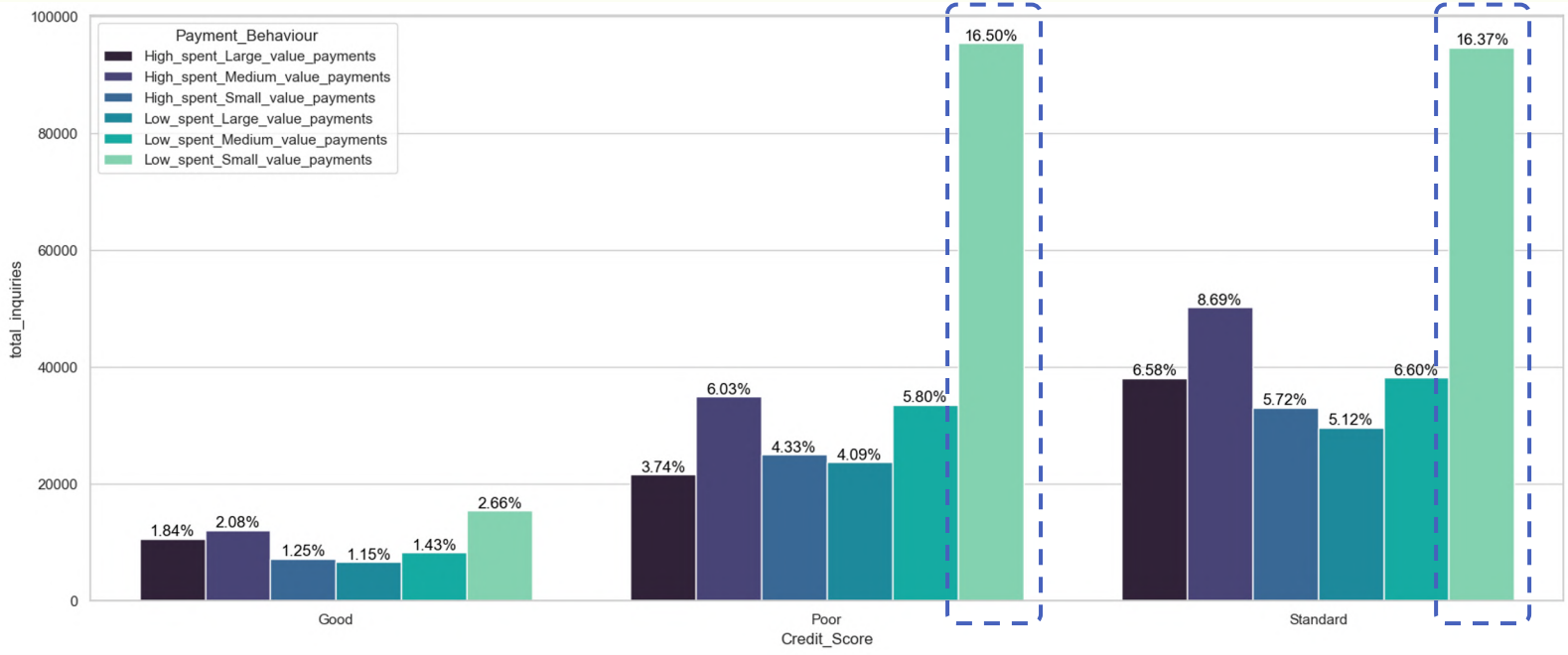


Mostly approximately 72% of borrowers **continue** to **keep** their **overdue payments within one month**, indicating that the **majority follow payment rules** in order to **maintain a good credit score.**

Manual Data
(Credit-related)

4. How many **New Credit Card** (Credit Inquiries) are there?

4. New Credit Card (Credit Inquiries)

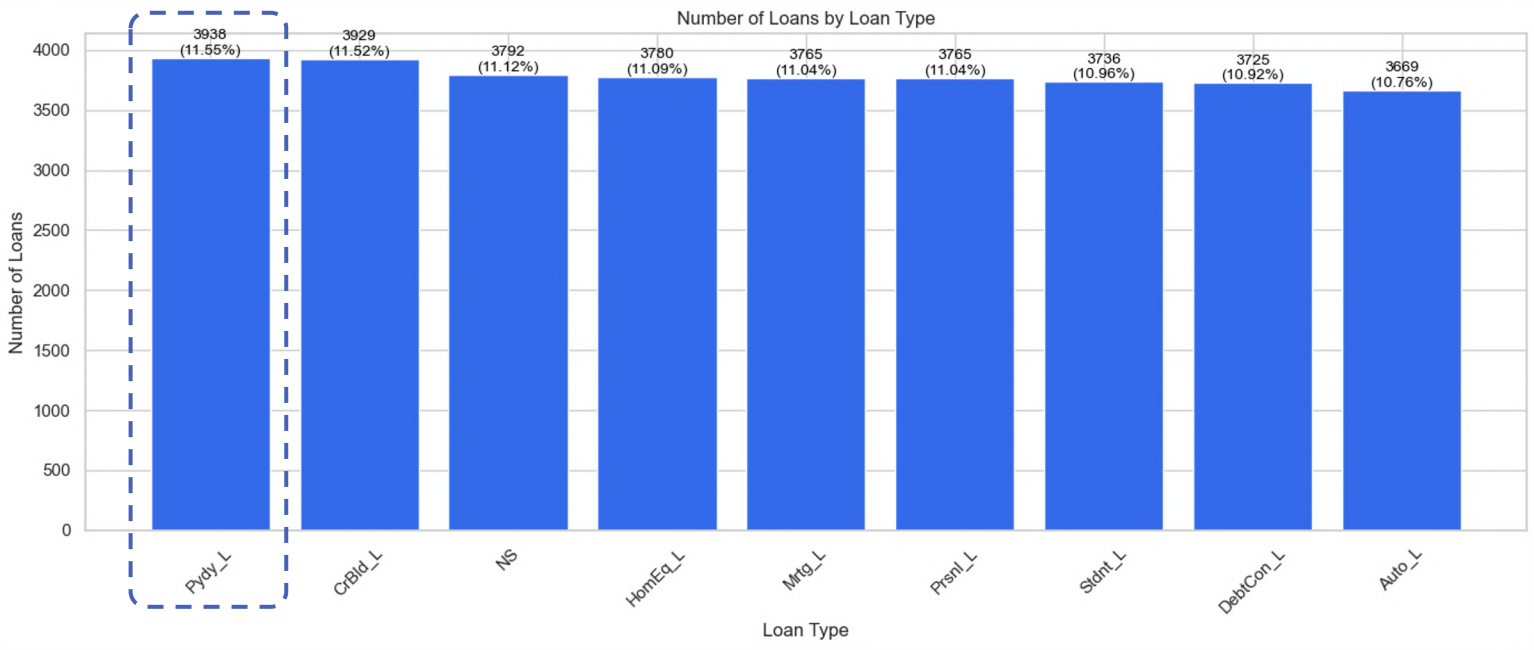


Many new credit cards are released to borrowers
who have **low spending** habits and make **small payments**, **resulting** in **poor** or **standard** credit scores.
This indicates that having **a lot of new credit card** could **lead** to **lower credit score**.

Manual Data
(Credit-related)

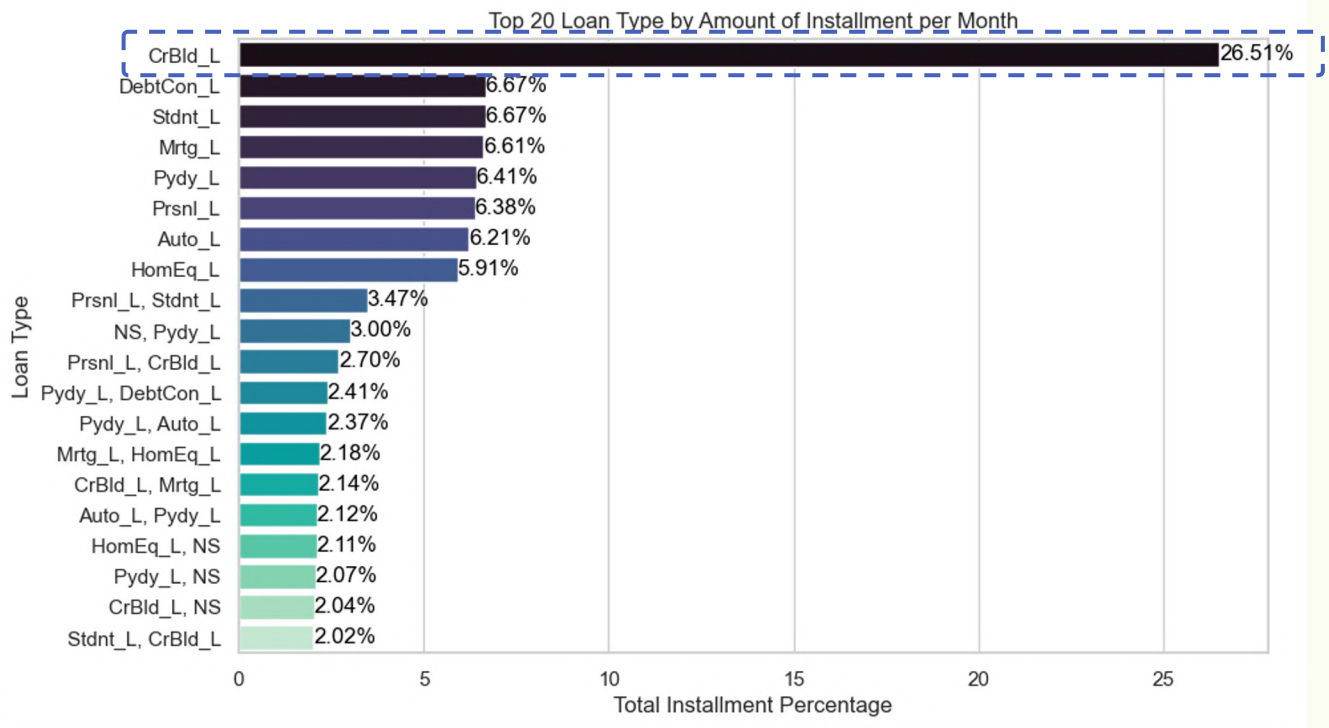
5. How many **Types of Loan** (Credit Mix) are there?

5. Type of Loan (Credit Mix)



The **largest loan type** that has taken by borrower is **Payday Loan**, however it has **tight difference** among others.

5. Type of Loan (Credit Mix)

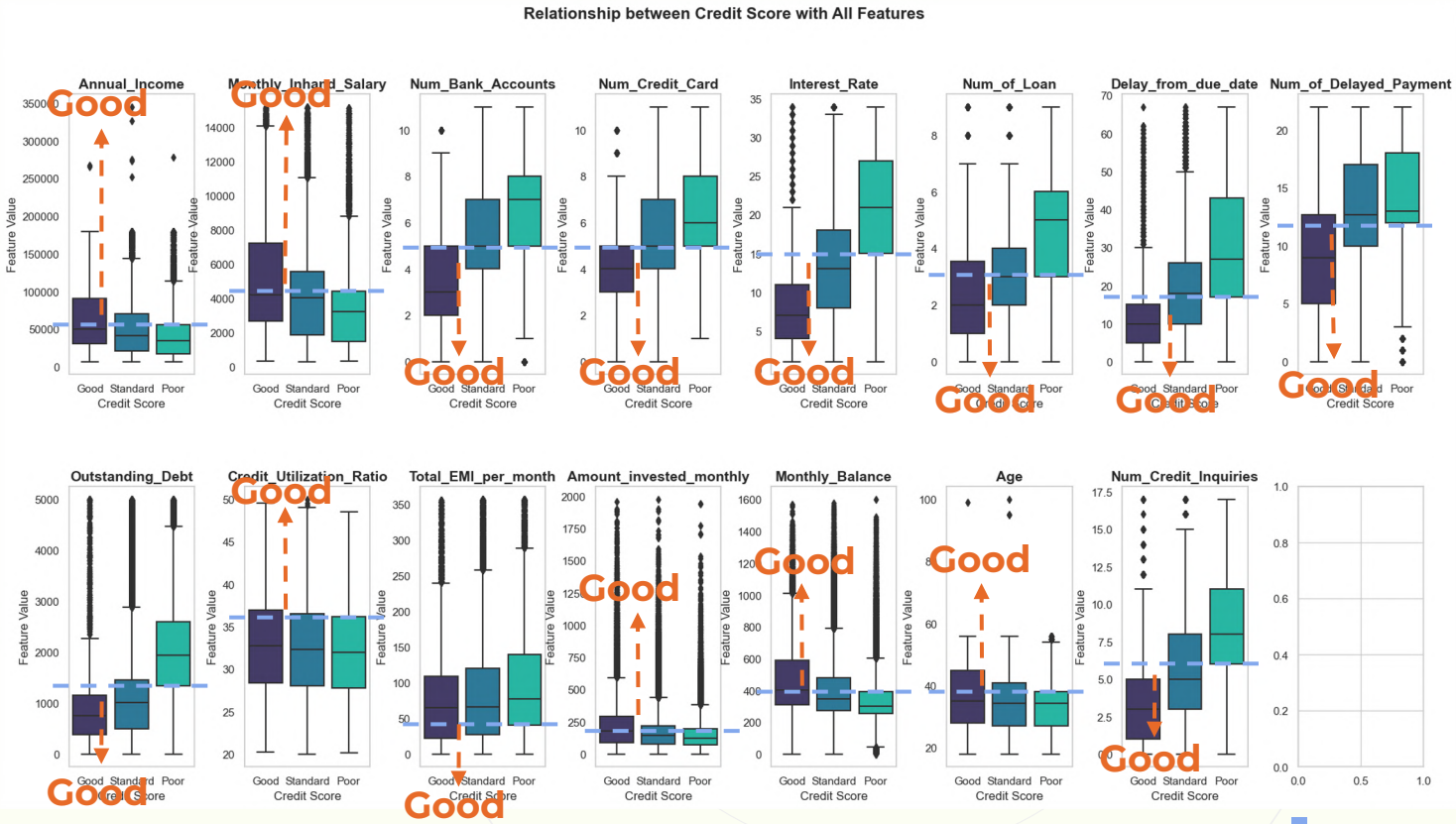


Top 1 type of loan that have largest installment per month is Credit Builder Loan. This indicates two possibilities: first, there are **many new borrowers**, or second, there are **many borrowers attempting** to **repair** their **damaged credit scores**.

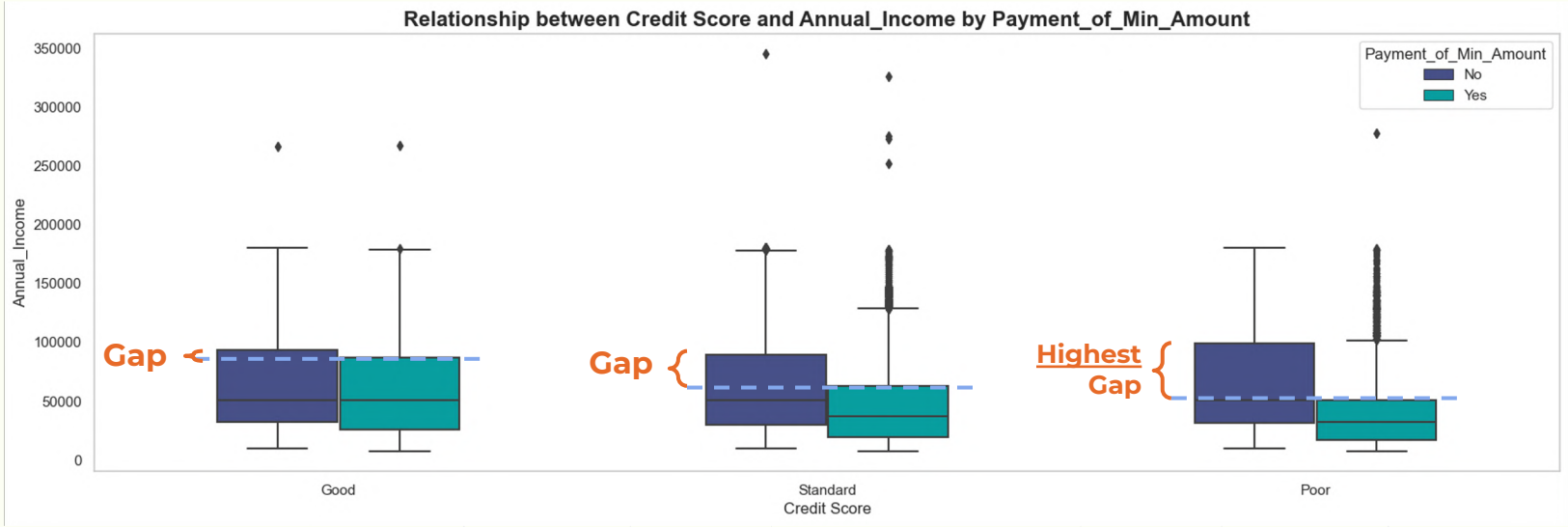
Credit Scoring

1. How to **achieve** the **Good Score**?

Achieving Good Score



Achieving Good Score



Some borrower that **have poor credit score** & **lower annual income** tend to **pay** with **minimum amount** in order to **maintaining credit score**,

meanwhile some borrower that have good score not affected to decide whether pay with minimum amount or not.

Achieving Good Score

To **maintain a good credit score** based on the data, borrowers should aim for the following financial behaviors:

1. Maintain an **annual income higher** than approximately 55,939 USD.
2. **Ensure a monthly take-home salary greater** than about 4,418 USD.
3. Have **fewer than around 5 bank accounts**.
4. Possess **fewer than around 5 credit cards**.
5. Keep **interest rates** on loans **below approximately 15%**.
6. **Limit** the number of **active loans** to **fewer** than **around 3.**
7. **Avoid delays in payments** beyond approximately **17 days from the due date.**
8. **Minimize** the occurrences of **delayed payments** to **fewer** than **around 12.**
9. Keep **outstanding** debt **lower** than **around 1,343 USD.**
10. **Maintain** a **credit utilization ratio higher** than around **36.21%.**
11. Ensure total **monthly EMI payments** are **less than around 40 USD.**
12. **Invest a monthly** amount of **at least or more** for around **196 USD**.
13. Maintain a **monthly balance higher** than around **393 USD.**
14. Instead of **not paying at all**, try to **do** at least **minimum amount**.

Gain **higher** income, **manage** debt,
make **on-time** payments, and use credit **wisely**

Credit Scoring

2. Who's the
Top 5 Good Score
(lowest interest) of borrowers?

Top 5 Borrowers with Good Score

(Lowest Interest)

Gained **higher** income

managed debt

on-time payments

use credit **wisely**

Name	Interest_Rate	Age	Occupation	Annual_Income	Credit_Utilization_Ratio	Delay_from_due_date	Num_Credit_Inquiries
Langep	1.0	20.000000	Entrepreneur	106733.13	39.243640	7.0	1.0
Ransdell Piersone	1.0	34.426366	Scientist	70175.78	38.021882	0.0	3.0
Shirboni	1.0	18.000000	Media_Manager	113781.39	37.998760	14.0	1.0
Raymondf	1.0	24.000000	Lawyer	65234.24	37.633227	11.0	1.0
Ellen Freilichl	1.0	52.000000	Scientist	86372.43	37.603381	10.0	3.0

Credit Scoring

3. Who's the
Top 5 Poor Score
(highest interest) of borrowers?

Bottom 5 Borrowers with Poor Score

(Highest Interest)

Should Gain **higher** income

Should **manage** debt

Should **on-time** payments

Should use credit **wisely**

Name	Interest_Rate	Age	Occupation	Annual_Income	Credit_Utilization_Ratio	Delay_from_due_date	Num_Credit_Inquiries
Andyo	16.0	34.426366	Architect	19041.48000	31.973401	50.0	11.0
Reddallr	16.0	22.000000	Mechanic	50485.81103	28.813102	22.0	10.0
Emelia Sithole-Matarisek	16.0	22.000000	Entrepreneur	18014.30000	33.989474	58.0	9.0
Lewis Krauskopft	16.0	37.000000	Mechanic	14061.55000	32.562537	58.0	13.0
Miyoung Kima	16.0	34.426366	Lawyer	50485.81103	33.617003	25.0	12.0

05.

MODELLING

Flow Project

Payment Behaviour/History
Amounts Owed/Credit Utilization
Length of Credit History
New Credit/Credit Inquiries
Type of Loan/Credit Mix

Manual Data

+

Alternate Data

Investing Info
Demographic
Credit Information

Train Model

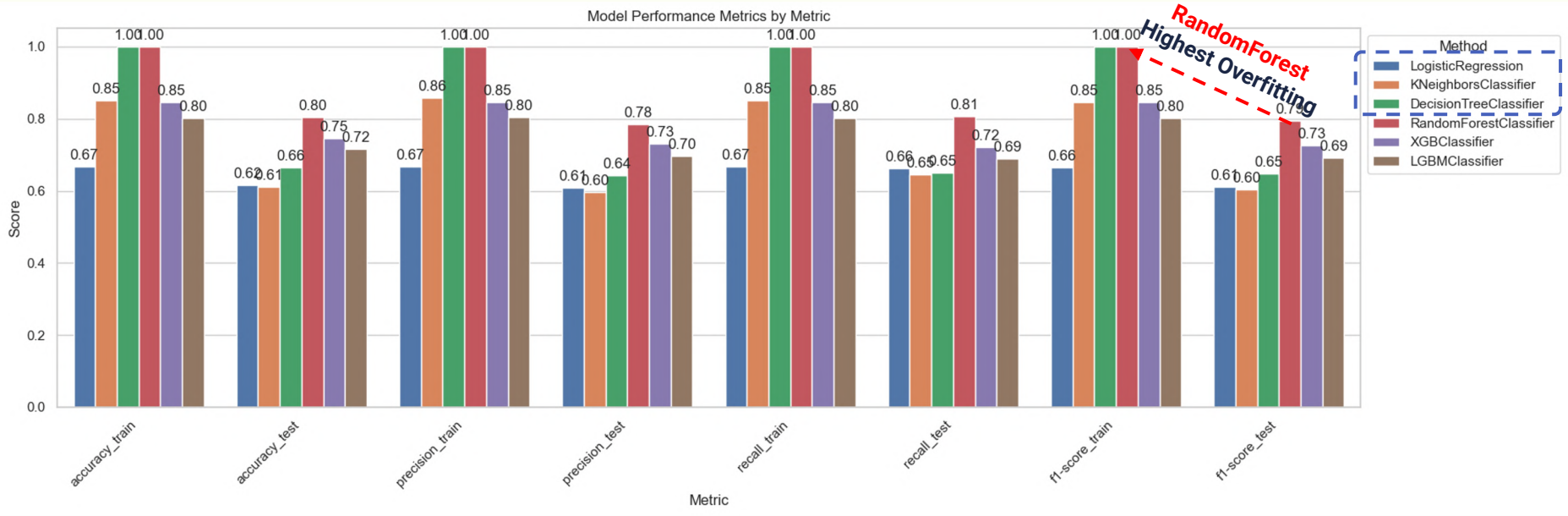
- Logistic Regression
- kNN
- Decision Tree
- Random Forest
- XGB
- LightGBM

Evaluation
Choose
The Best Model

Optimizing
The
Best Model

Deploy

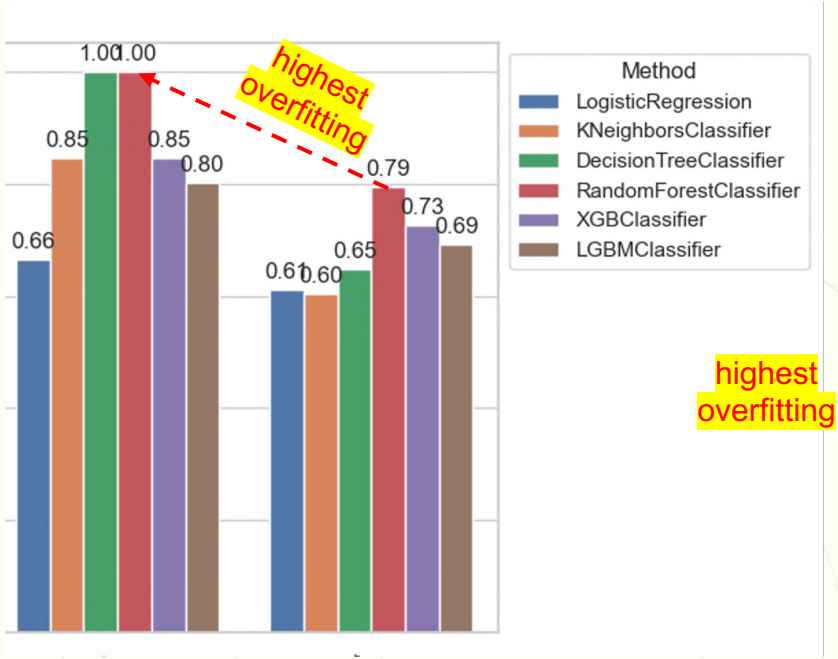
Base Model Evaluation Metrics



Best **three models** are **RandomForestClassifier**, **XGBClassifier**, **LGBMClassifier**, but it's clear that **RandomForestClassifier** has **highest overfitting** issue.

Furthermore, **I need to look into their F1-score of train and test** to get the best model.

Base Model Evaluation Metrics



highest overfitting

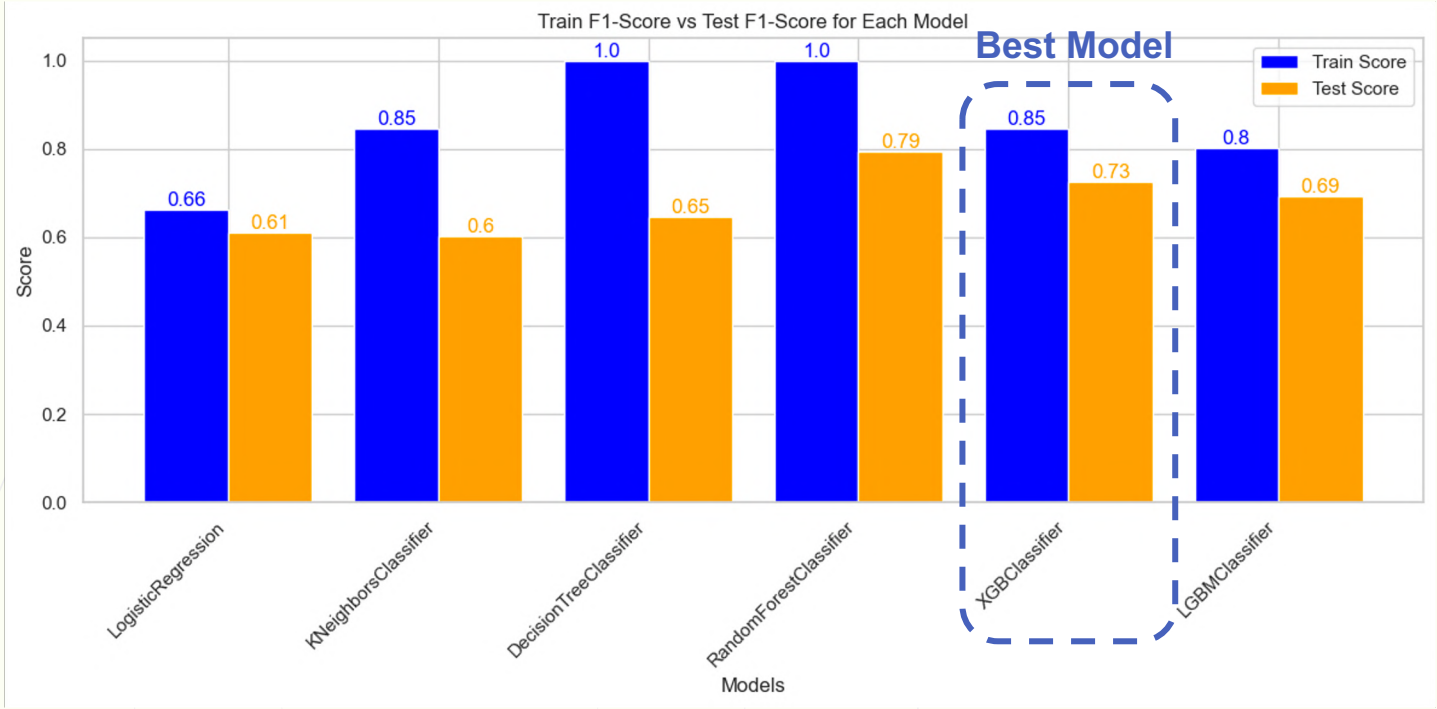
	Method	f1-score_train	f1-score_test
0	LogisticRegression	0.663737	0.610723
1	KNeighborsClassifier	0.845305	0.603372
2	DecisionTreeClassifier	1.000000	0.646936
3	RandomForestClassifier	0.999992	0.794408
4	XGBClassifier	0.846007	0.725748
5	LGBMClassifier	0.801222	0.691570

Top Three best model

Best **three models** are **RandomForestClassifier**, **XGBClassifier**, **LGBMClassifier**, but it's clear that **RandomForestClassifier** has **highest overfitting** issue.

Furthermore, **I need to look into their F1-score of train and test** to get the best model.

Base Model Evaluation Metrics



Based on the **highest test of F1-score** and a slightly **acceptable overfitting** difference, the **XGBClassifier** achieved the **best results**. Therefore, the model **needs to be tuned.**

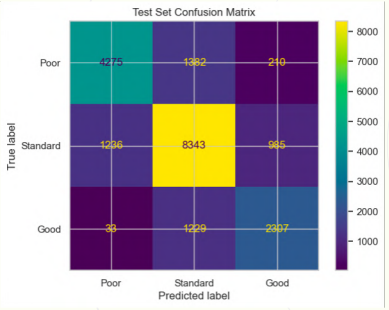
Tuning Model Evaluation Metrics

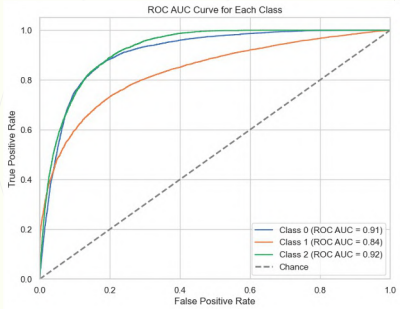
4%

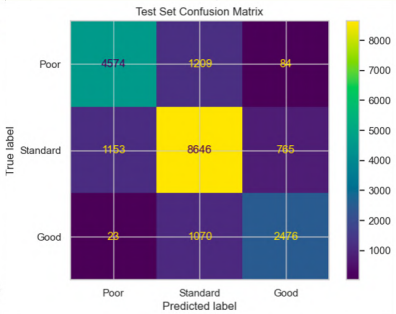
74% 78%

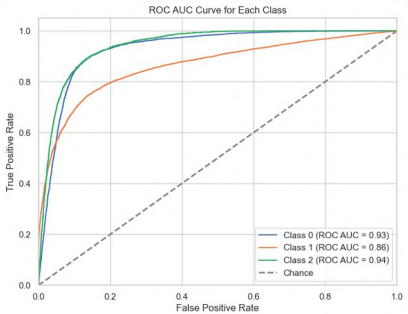
A slight accuracy increased
also improved F1-scores as follow below:

1. Poor Class, 0.75 -> 0.79
2. Standard Class, 0.78 -> 0.80
3. Good Class, 0.65 -> 0.72

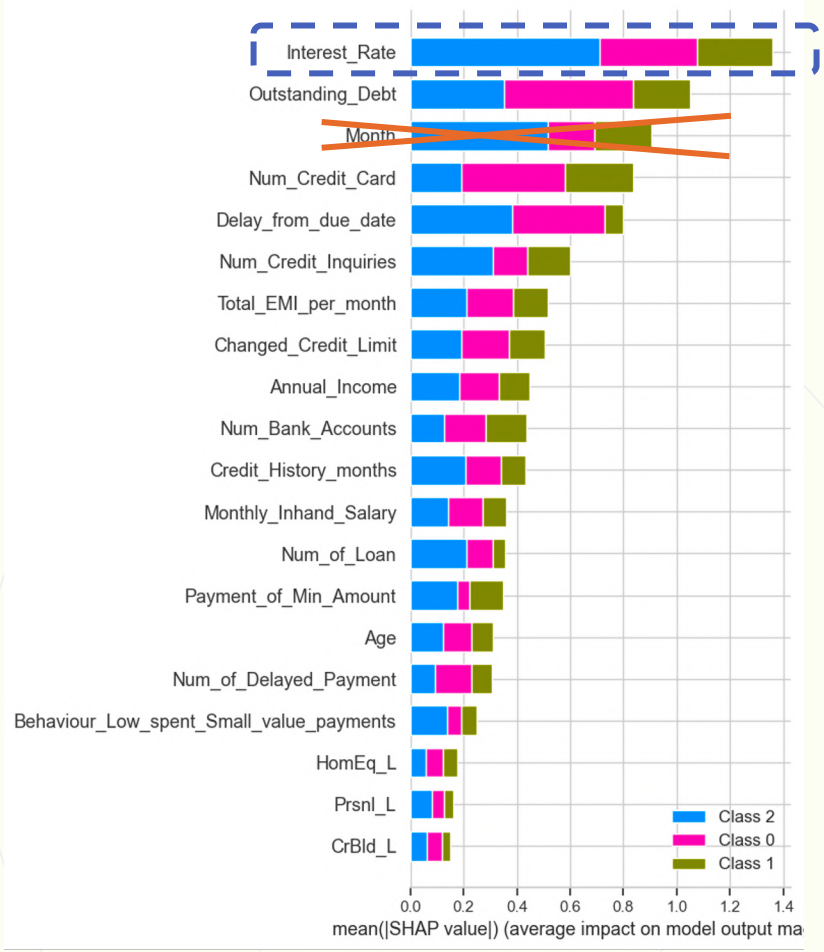








SHAP Summary Plot

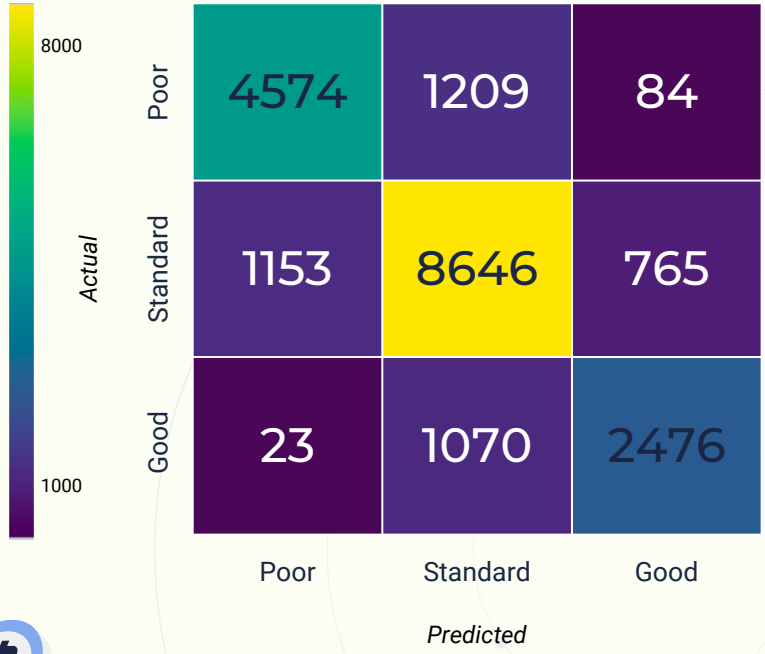


Interest Rate is most influential feature in the model for predicting credit scores across different classes. It makes sense because the higher credit score that we have then the lower interest that we got (vice versa)

Meanwhile for **other features** except interest rate such as **Outstanding Debt, Number of Credit Card, etc** suggest that borrower with high outstanding debt, many credit card, late payment, high installment per month, and so on, are more likely to **fall into "Poor" credit** category which will **cost with higher interest.**

Potential Impact on Business

Misclassification Cost Matrix
(Confusion Matrix)



KreditDulu have 17,900 clients and classifies into these categories as follows:

For the sake of simplify, assuming that you have Profit and Loss values for each classification scenario based on risk probability of default (PD) and higher PD will have higher interest, therefore let's say:

- Profit per True Positive (Correct classification): $100*
- Misclassifying as a better category: -$50*
- Misclassifying as a worse category: -$200*
- Misclassifying as a worst category: -$400*
- Misclassifying as a worst category (could default) : -$1500*

You can now calculate the Profit and Loss:
Profit = ((4574 + 8646 + 2476) * $100) + ((1153+1070) * -$50) + ((765 + 1209) * -$200) + (23 * -$400) + (84 * -$1000)
Profit = $1,569,600 - $111,150 - $394,800 - $9,200 - $84,000
Profit = $970,450

*Note: *the cost is assumption*

06.

CONCLUSION & RECOMMENDATION

Conclusion:

The model with **combination** of **manual** and **alternate data** could have **business impact** of **$970,450** due to its **78%** performance in predicting credit scoring for each borrower, which enables provision of fair interest rate and reduce risk of potential default. As the result, KreditDulu can use the model as a more effective tool to do credit scoring.

Recommendation:

1. Educating borrower to maintain their credit score which good for borrower (they can lend again and again), and good for lender (lower risk for default payment, faster repayment)
2. Enrich the model with more alternate data such as telecommunication transaction, e-commerce transaction, utility data, travel data, insurance data.
3. Exploring alternative method using machine learning/deep learning such as adaboost, SVM, neural network
4. Develop more to distinguish new borrower and existing borrower

THANKS!

Let's Connect!

saga.satriabagaskara@gmail.com

linkedin.com/in/satriabagaskara

github.com/dataaga

medium.com/@dataaga

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon** and infographics & images by **Freepik**

Source:

1. https://www.listendata.com/2019/08/credit-risk-modelling.html
2. https://www.investopedia.com/terms/d/defaultprobability.asp#:~:text=Default%20probability%2C%20or%20probability%20of,one%20gauge%20of%20default%20risk.
3. https://flinks.com/blog/loan-default-risk-prevention-and-management-in-economic-uncertainty/
4. https://www.investopedia.com/articles/pf/10/credit-score-factors.asp
5. https://www.creditkarma.com/advice/i/what-affects-your-credit-scores
6. https://www.sciencedirect.com/science/article/pii/S1877050922019287
7. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00768-7
8. https://www.bis.org/publ/work834.pdf
9. https://www.researchgate.net/publication/330724392_An_Empirical_Comparison_of_Machine-Learning_Methods_on_Bank_Client_Credit_Assessments
10. https://altair.com/newsroom/articles/credit-scoring-series-part-eight-credit-risk-strategies
11. https://altair.com/newsroom/articles/credit-scoring-series-part-six-segmentation-and-reject-inference
12. https://id.blog.finantier.co/mengenal-credit-scoring-dan-teknologi-pendukungnya-8360edcf5c42
13. https://blog.finantier.co/the-importance-of-credit-scoring-to-support-fintech-business-e478e52891b4