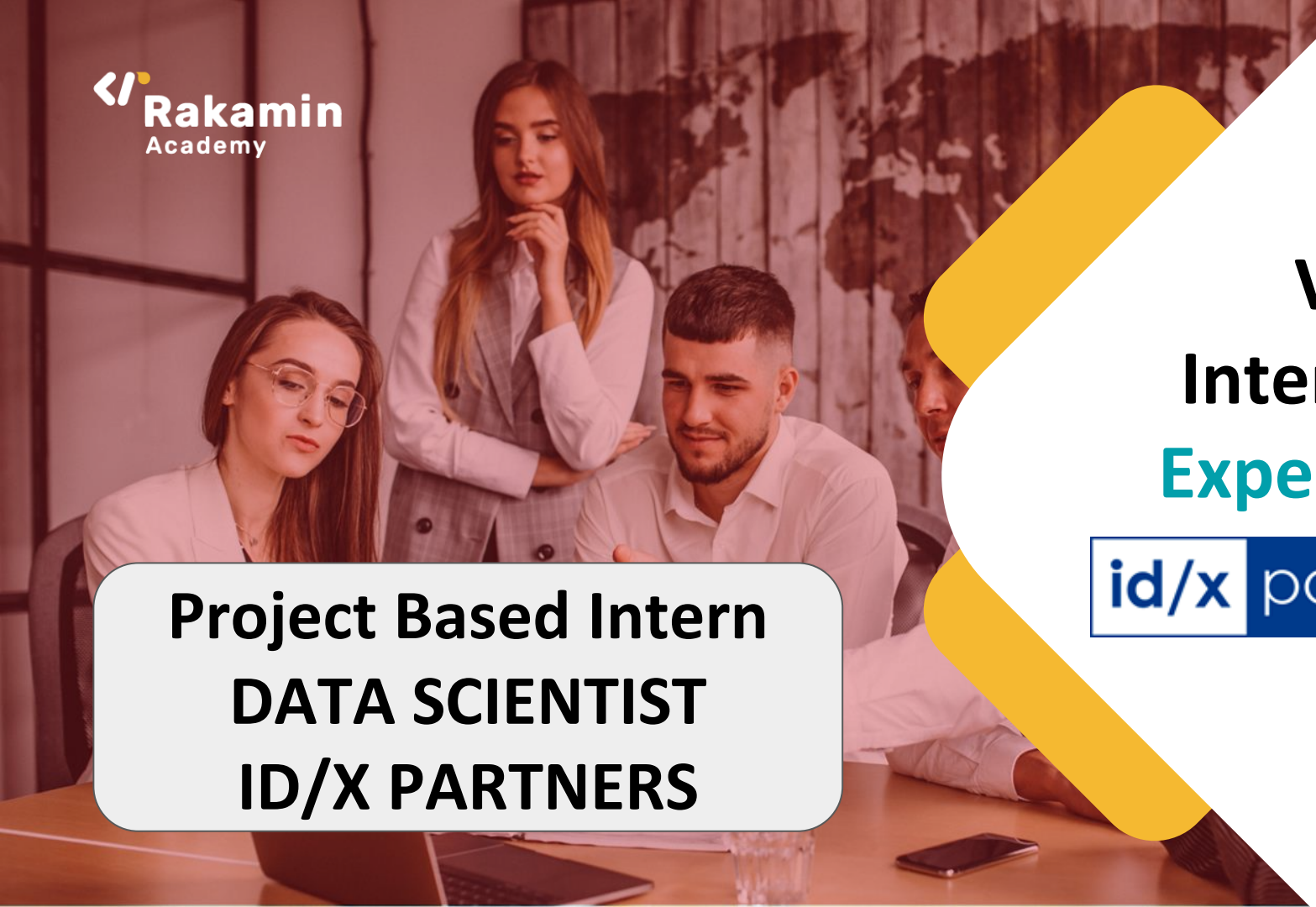


Rakamin Academy

# Virtual Internship Experience

id/x partners

## Project Based Intern
## DATA SCIENTIST
## ID/X PARTNERS

Rakamin Academy

# Credit Risk Prediction

Antonius Andi Syah Putra Halawa

# Outline

- Business Understanding
- Analytic Approach
- Data Requirements
- Data Collection
- Data Understanding
- Data Preparation
- Exploratory Data Analysis
- Model Building
- Model Evaluation
- Model Deployment

# Business Understanding

Credit risk is the risk of loss due to a borrower's failure to make payments on a loan or meet other financial obligations. In other words, credit risk is the possibility that a borrower will default on their debt obligations, causing the lender or investor to lose some or all of their investment.

Credit risk is a significant concern for lenders, as it can have a significant impact on their financial performance. To manage credit risk, lenders use a variety of tools and techniques, including credit scoring, risk assessment, and loan monitoring. They may also require collateral or other forms of security to reduce their risk exposure.

# Analytic Approach

Predictive Model; Supervised Learning

Classification, namely the type of supervised learning where the data type of the variable label/target is discrete.

Regression, which is a type of supervised learning where the data type of the variable label/target is continuous numeric.

# Data Understanding

1. **Libraries**
2. **Importing Data**

   *df = pd.read_csv('loan_data_2007_2014.csv')*

3. **Exploring Data**

   *- df.head()*

   *- df.tail()*

   *- df.shape*

   *- df.id.nunique()*

   *- df.member_id.nunique()*

   *- df.info()*

   *- df.describe().T*

# Data Preparation

1. **Remove unnecessary columns**
2. **Missing values (NaNs)**
3. **Define Target Variable/Labeling**

   - Target Variable =`loan_status’

   - Define of target variabel `bad loan` & `good loan`

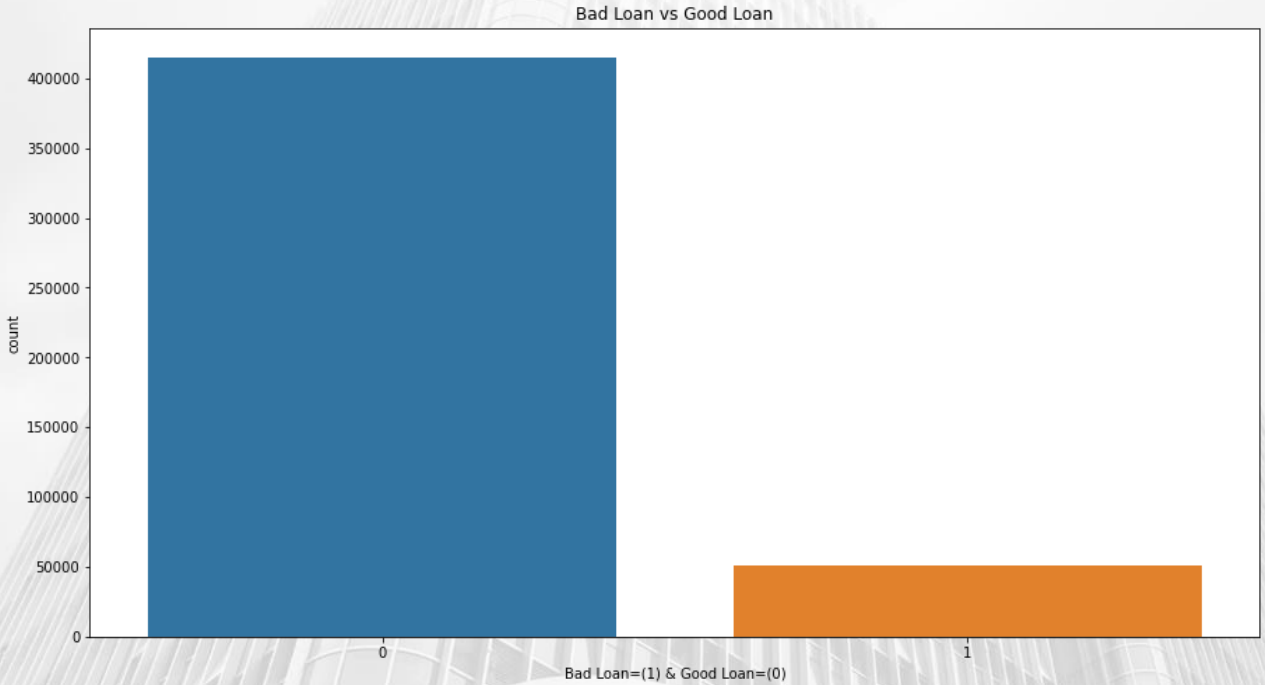
**4. CLEANING, PREPROCESSING, FEATURE ENGINEERING**

Cleaning/modification

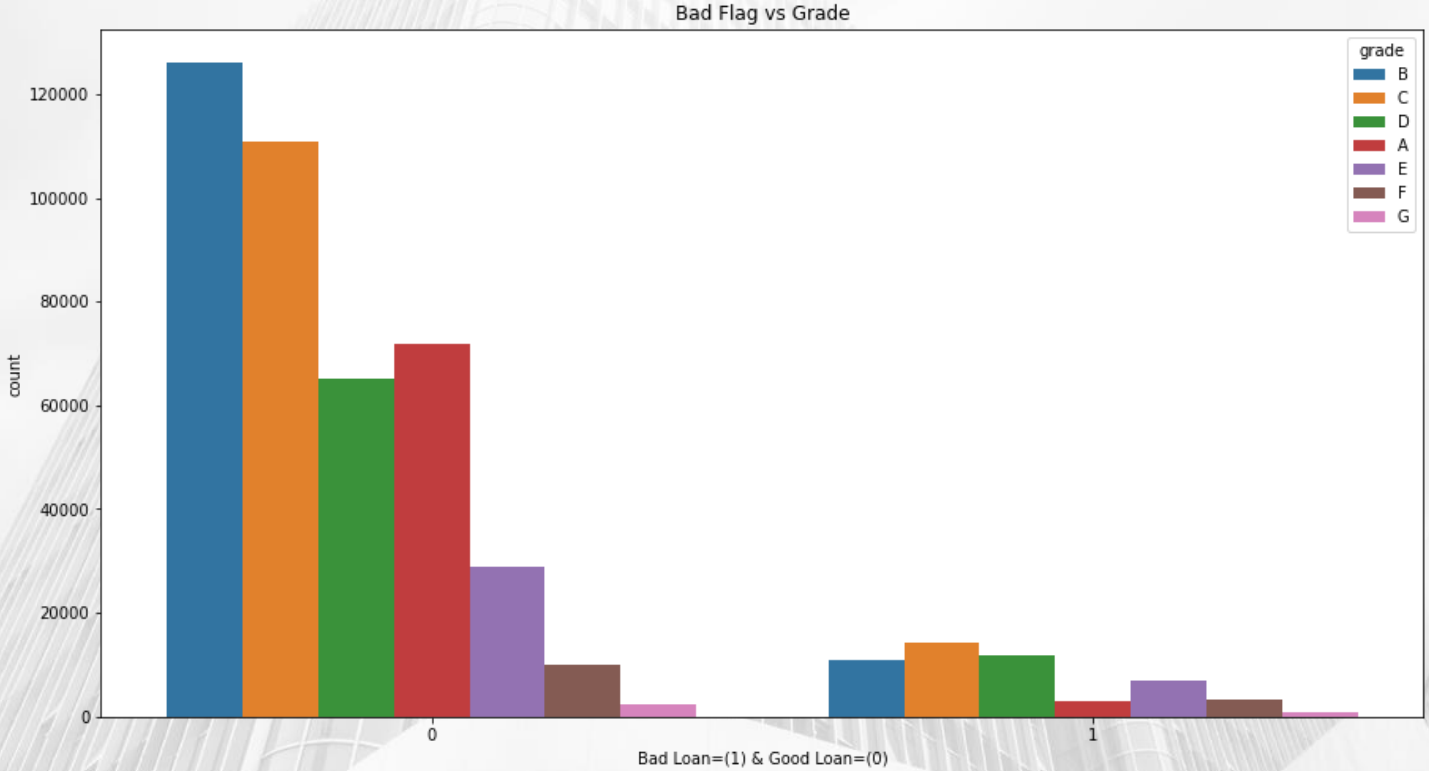
Column : **emp_length, term, earliest_cr_line, issue_d, last_pymnt_d, next_pymnt_d, last_credit_pull_d**

# Exploratory Data Analysis

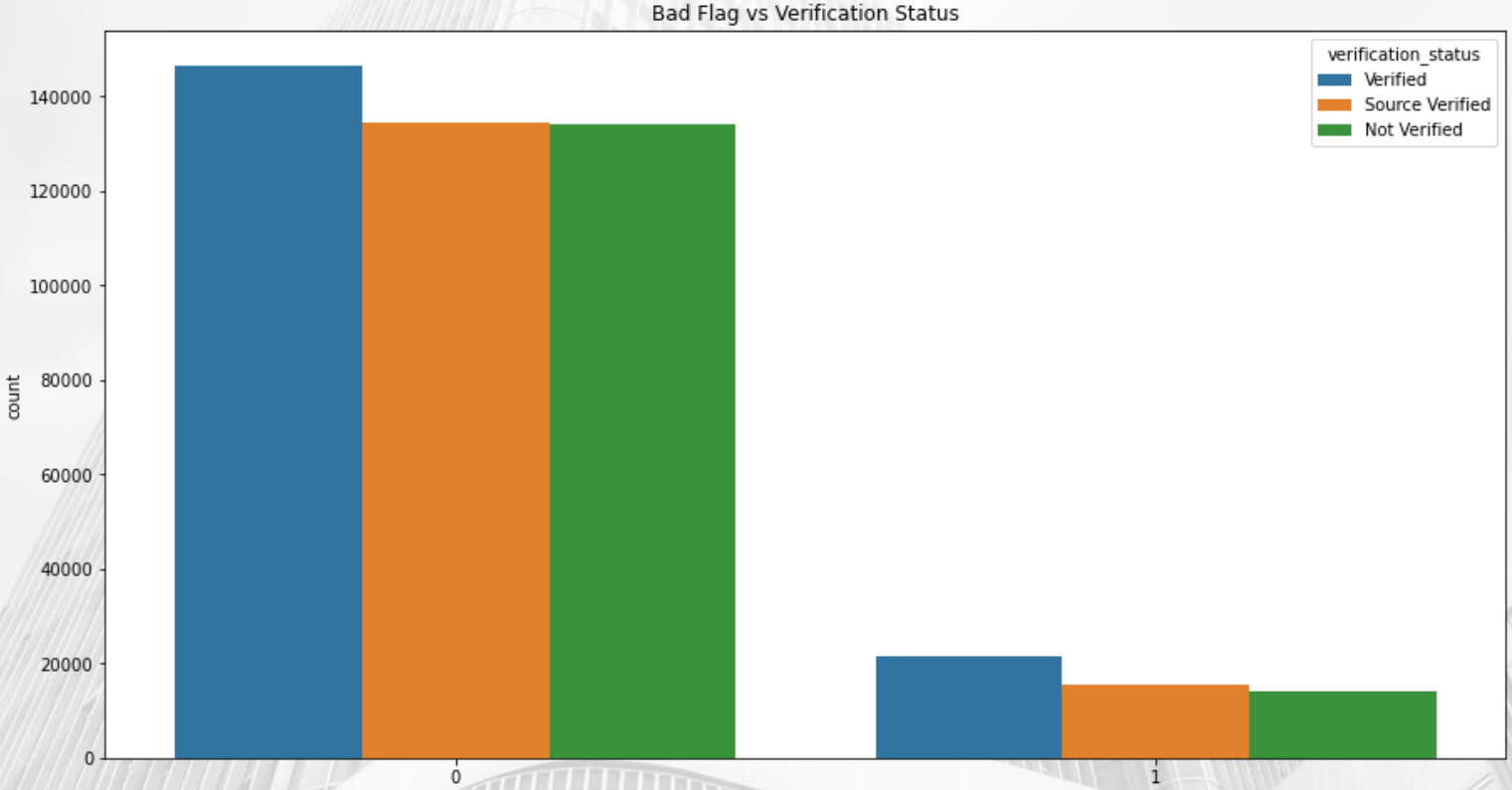
- **Visualization of Data**



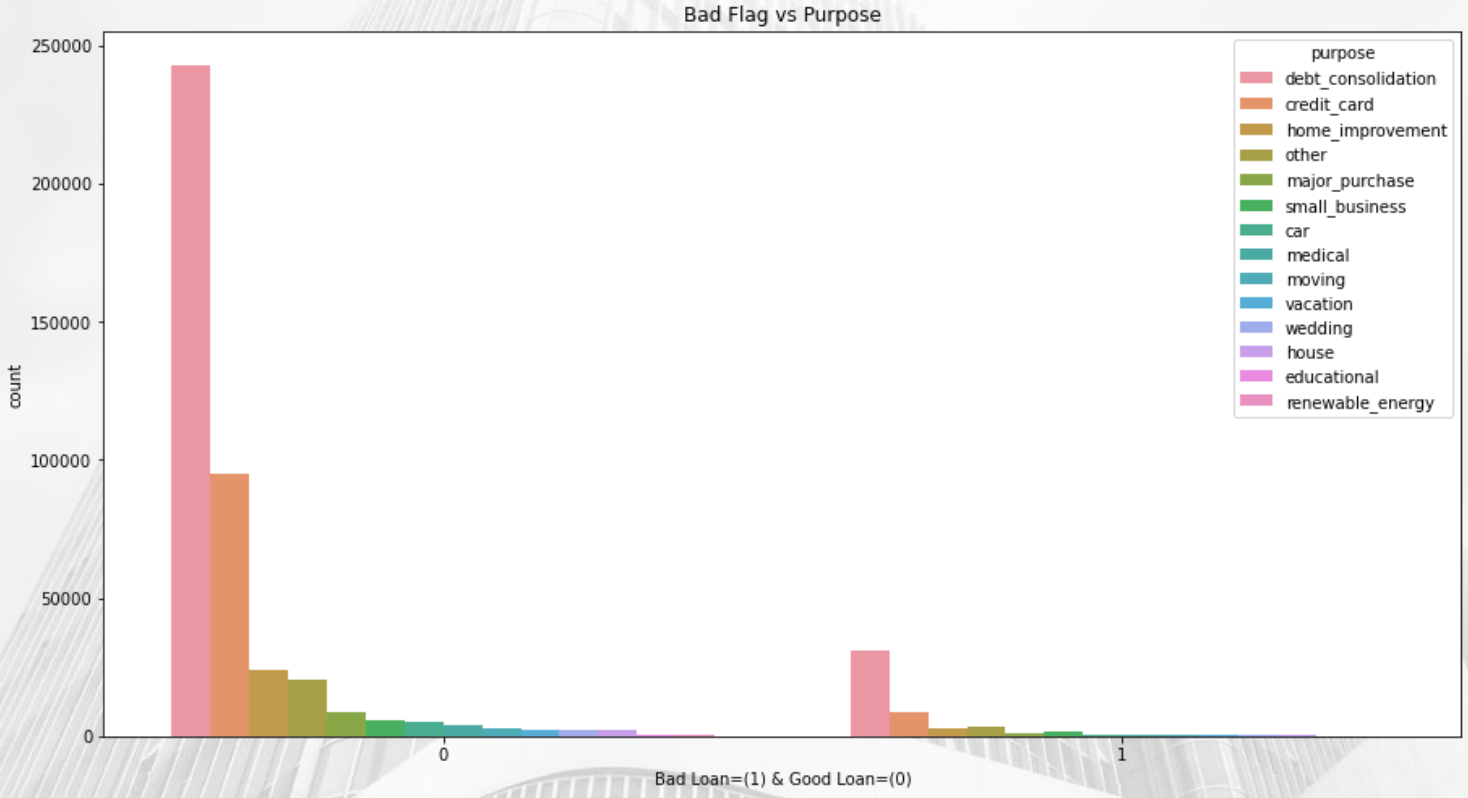
- **Visualization of Data**



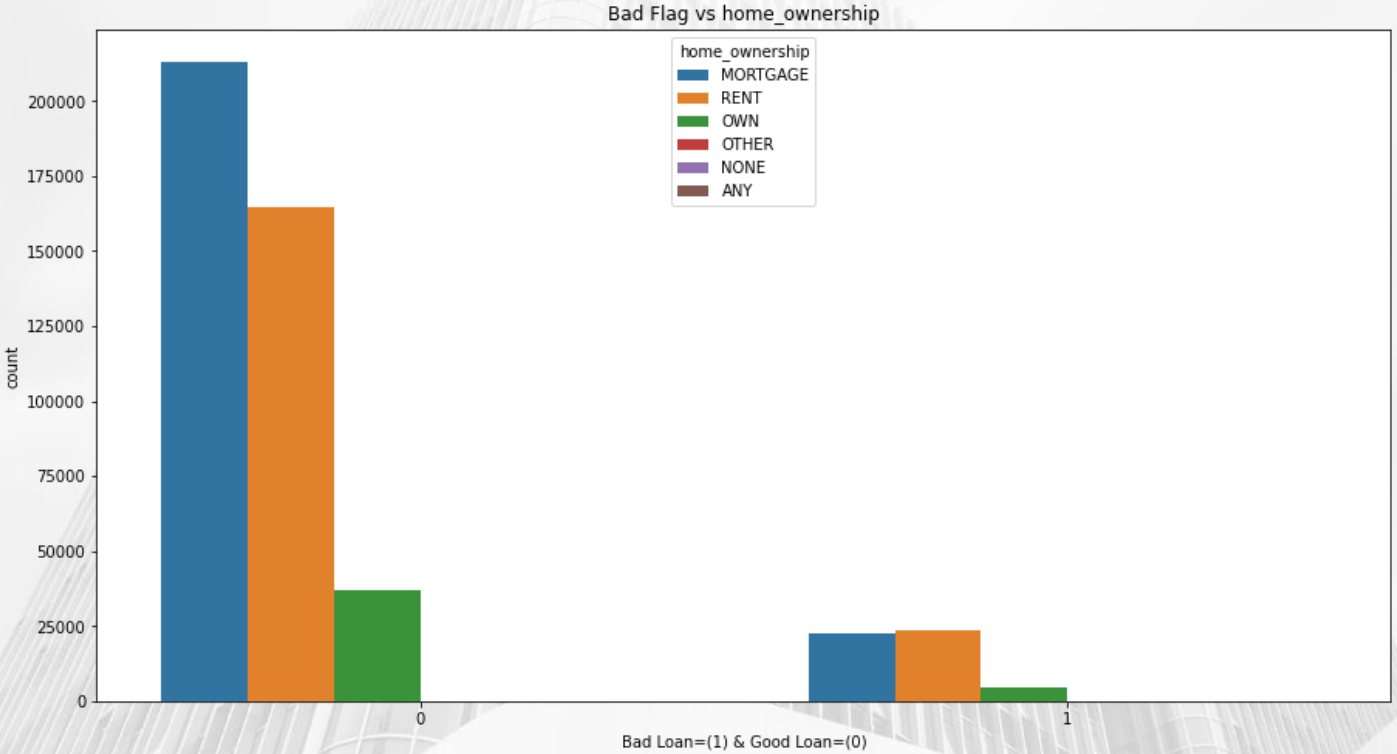
- **Visualization of Data**



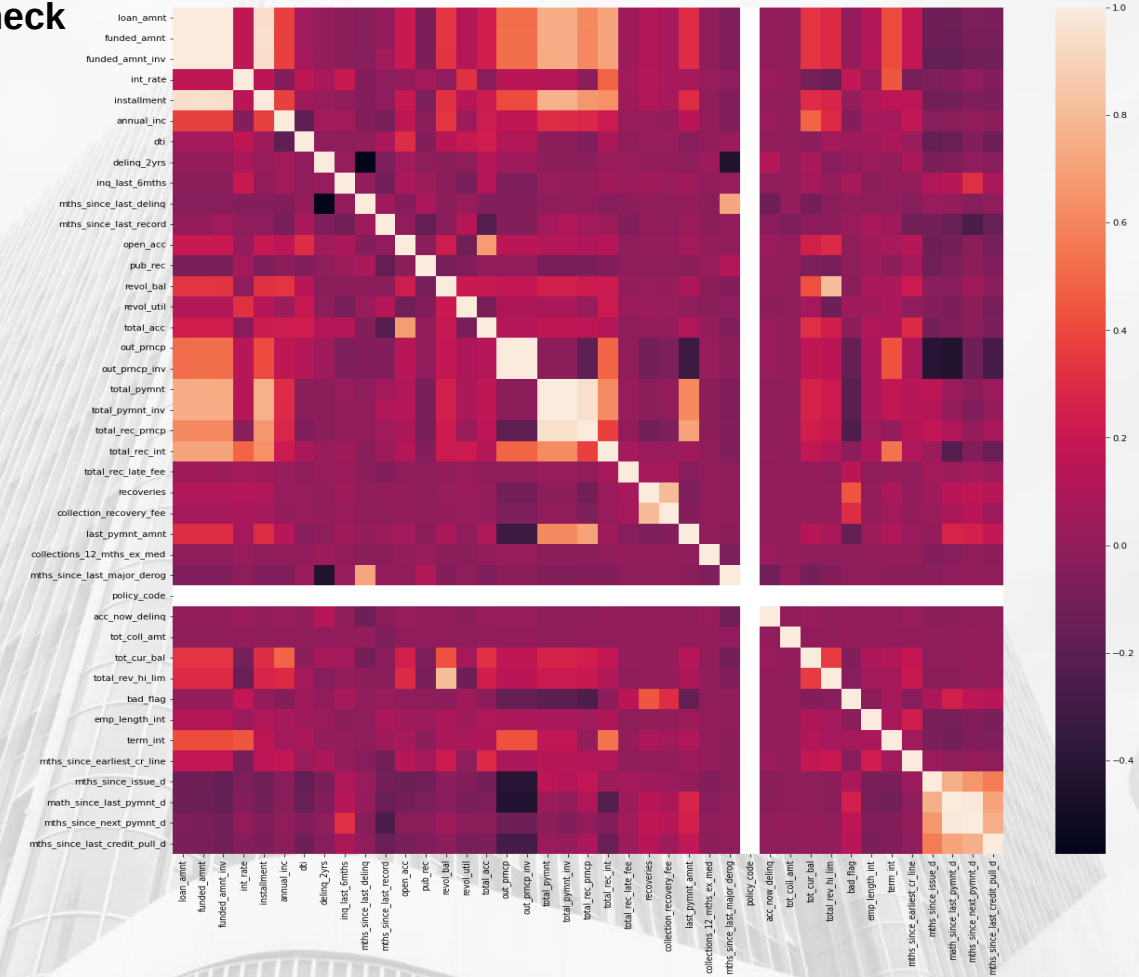
- **Visualization of Data**



- **Visualization of Data**



- **Correlation Check**



- **Check Categorical Features**

  Disposal of features that have a very high unique value (high cardinality) and features that have only one unique value.

- **Missing Values**

- **Feature Scaling And Transformation**

  - One Hot Encoding
  - Standardization
  - Transformed DataFrame

# Model Building

## 1. Algoritma Supervised Learning

- Decision Tree
- K-Nearest Neighbors (KNN)
- Logistic Regression (For Classification)
- Linear Regression (For Regression)
- Random Forest
- Support Vector Machine (SVM)
- XGBoost

## 2. Train-Test Split

- X_train.shape, X_test.shape

```
((373028, 99), (93257, 99))
```

## 3. Training

- **Random Forest**

*rfc = RandomForestClassifier(max_depth=4)*

*rfc.fit(X_train,y_train)*

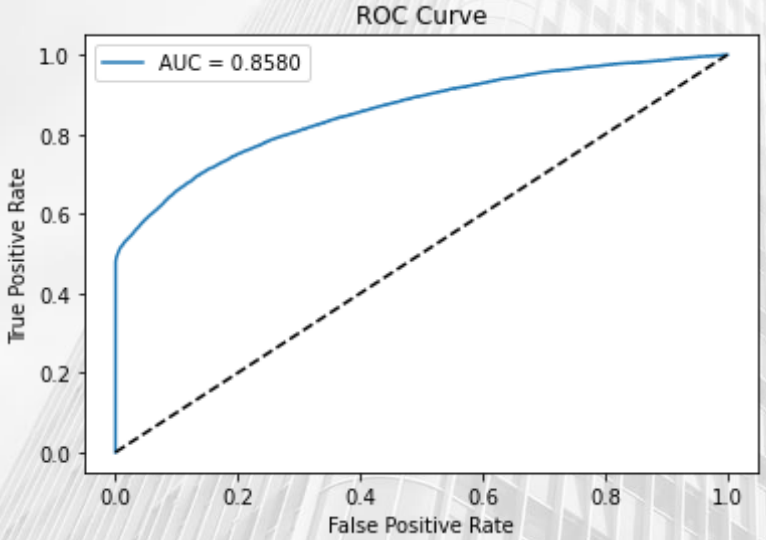
- **Feature Importance**

  Feature Importance is a measure used to evaluate the importance of each feature in a machine learning model to the results of prediction of target variables.

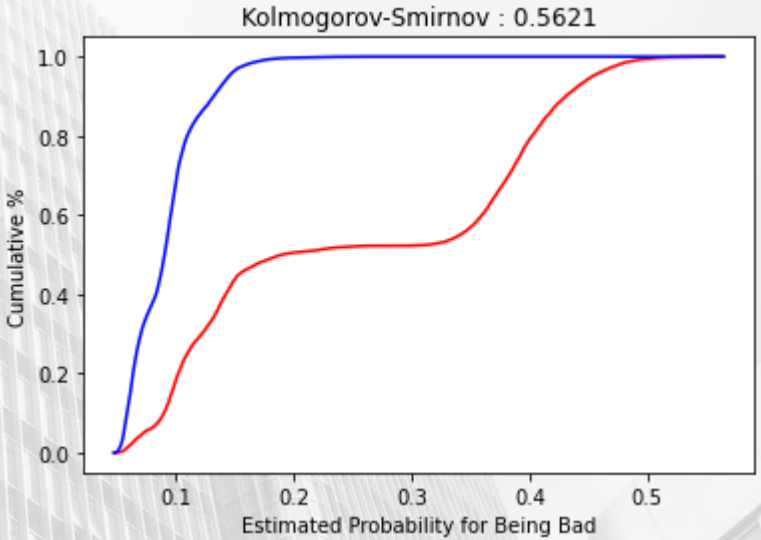
- **Validation**

  To measure model performance, AUC (Area Under the Curve) and KS (Kolmogorov-Smirnov) are evaluation metrics commonly used in risk prediction modeling and credit scoring.

- **AUC**


ROC Curve
AUC = 0.8580
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

- **KS**


Kolmogorov-Smirnov : 0.5621
Cumulative %
Estimated Probability for Being Bad
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5

# Model Evaluation

The built model produces AUC = 0.857 and KS = 0.56 performance. In the world of credit risk modeling, generally AUC above 0.7 and KS above 0.3 is considered good performance.