

# Modelling for Credit Risk Management

In collaboration with id/x partners



Created by:


**Muhammad Ariq Arfina**

[ariqarfina05@gmail.com](mailto:ariqarfina05@gmail.com)

LinkedIn : [Muhammad Ariq Arfina](#)

Github : [ariqarfina](#)

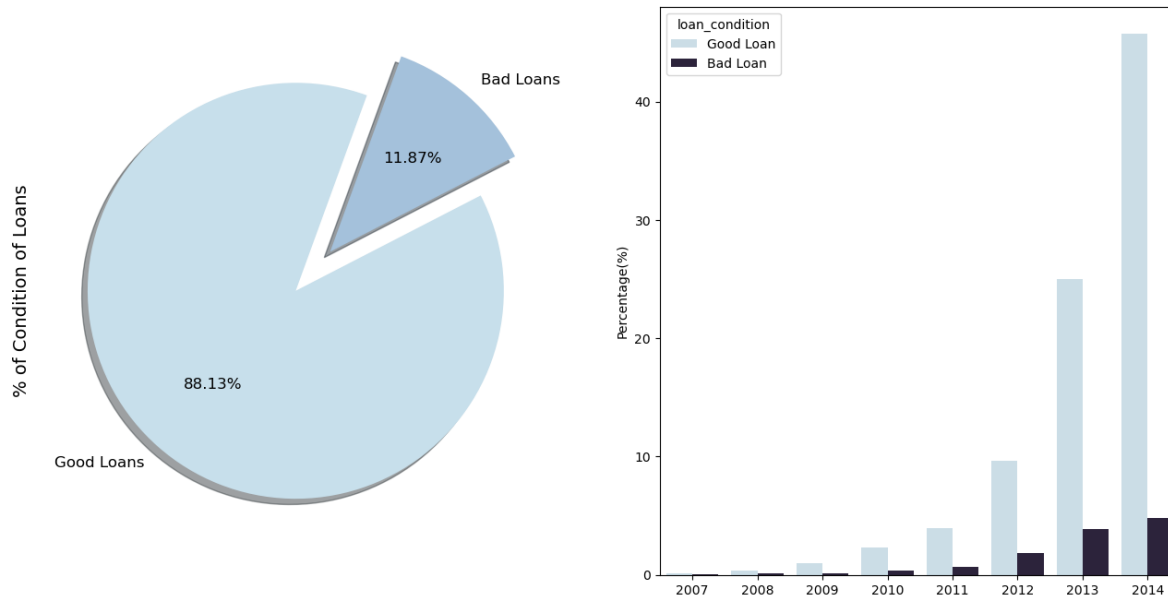
Supported by:  
**Rakamin Academy**  
Career Acceleration School  
[www.rakamin.com](http://www.rakamin.com)

“We are  consultants. We have a client, Credifo Bank, an Indonesian Fintech Startup that is developing a Credit Risk Management system. For Credifo Bank, we were tasked with developing a model that predicts good credit risk. We were given a dataset by Credifo Bank, namely Loan Dataset from 2007 - 2014.”



## What's the Problem?

Information on Loan Conditions



\*Bad Loan are loans that are: a) charged off, b) Late more than two weeks, c) in grace period, d) Defaulted

**Customers who take loans at Credifo Bank are increasing every year. However, every year, customers who are detected by bad loans\* are increasingly following as well.**

**It can be seen in this dataset that 11.87% of Bad Loaners\* were detected.**

For more details, you can see dataset [here](#)

## Goal & Objective

### Goal:

Reducing the risk of default, which can harm the Company

### Objective:

- Create a model that can classify credit risk management
- Making credit risk management

We were given a dataset by Bank Credifo, namely Loan Dataset from 2007 - 2014.

Loan Dataset

id	emp_title	purpose	pub_rec,revol_bal	last_pymnt_d	acc_now_delinq	open_rv_24m
member_id	emp_length	title	revol_util	last_pymnt_amnt	tot_coll_amt,tot_cur_bal	max_bal_bc
loan_amnt	home_ownership	zip_code,addr_state	total_acc	next_pymnt_d	open_acc_6m,	all_util
funded_amnt	annual_inc	dti	initial_list_status	last_credit_pull_d	open_il_6m	total_rev_hi_lim
funded_amnt_inv	verification_status	delinq_2yrs	out_prncp,out_prncp_inv	collections_12_mths_ex_med	open_il_12m	inq_fi
term	issue_d	earliest_cr_line	total_pymnt	mths_since_last_major_d erog	open_il_24m	total_cu_tl
int_rate	loan_status	inq_last_6mths	total_pymnt_inv	policy_code	mths_since_rcnt_il	inq_last_12m
installment	pymnt_plan	mths_since_last_delinq	total_rec_prncp	application_type	total_bal_il	
grade	url	mths_since_last_record	total_rec_int,total_rec_lat e_fee	annual_inc_joint	il_util	
sub_grade	desc	open_acc	recoveries,collection_rec overy_fee	dti_joint,verification_stat us_joint	open_rv_12m	

For more details, you can see dataset [here](#)



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 466285 entries, 0 to 466284
Data columns (total 75 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Unnamed: 0                            466285 non-null  int64
1   id                                     466285 non-null  int64
2   member_id                             466285 non-null  int64
3   loan_amnt                             466285 non-null  int64
4   funded_amnt                            466285 non-null  int64
5   funded_amnt_inv                       466285 non-null  float64
6   term                                  466285 non-null  object
7   int_rate                              466285 non-null  float64
8   installment                           466285 non-null  float64
9   grade                                 466285 non-null  object
10  sub_grade                             466285 non-null  object
11  emp_title                             438697 non-null  object
12  emp_length                             445277 non-null  object
13  home_ownership                         466285 non-null  object
14  annual_inc                             466281 non-null  float64
15  verification_status                   466285 non-null  object
16  issue_d                               466285 non-null  object
17  loan_status                           466285 non-null  object
18  pymnt_plan                             466285 non-null  object
19  url                                    466285 non-null  object
20  desc                                  125983 non-null  object
21  purpose                                466285 non-null  object
22  title                                 466265 non-null  object
23  zip_code                               466285 non-null  object
24  addr_state                             466285 non-null  object
25  dti                                    466285 non-null  float64
26  delinq_2yrs                             466256 non-null  float64
27  earliest_cr_line                       466256 non-null  object
28  inq_last_6mths                         466256 non-null  float64
29  mths_since_last_delinq                 215934 non-null  float64
30  mths_since_last_record                 62638 non-null  float64
31  open_acc                               466256 non-null  float64
32  pub_rec                                466256 non-null  float64
33  revol_bal                              466285 non-null  int64
34  revol_util                             465945 non-null  float64
35  total_acc                              466256 non-null  float64
```

```
36  initial_list_status                   466285 non-null  object
37  out_prncp                             466285 non-null  float64
38  out_prncp_inv                         466285 non-null  float64
39  total_pymnt                           466285 non-null  float64
40  total_pymnt_inv                       466285 non-null  float64
41  total_rec_prncp                       466285 non-null  float64
42  total_rec_int                         466285 non-null  float64
43  total_rec_late_fee                    466285 non-null  float64
44  recoveries                            466285 non-null  float64
45  collection_recovery_fee               466285 non-null  float64
46  last_pymnt_d                          465909 non-null  object
47  last_pymnt_amnt                       466285 non-null  float64
48  next_pymnt_d                          239071 non-null  object
49  last_credit_pull_d                   466243 non-null  object
50  collections_12_mths_ex_med           466140 non-null  float64
51  mths_since_last_major_derog           98974 non-null  float64
52  policy_code                           466285 non-null  int64
53  application_type                     466285 non-null  object
54  annual_inc_joint                      0 non-null      float64
55  dti_joint                             0 non-null      float64
56  verification_status_joint             0 non-null      float64
57  acc_now_delinq                        466256 non-null  float64
58  tot_coll_amt                          396009 non-null  float64
59  tot_cur_bal                           396009 non-null  float64
60  open_acc_6m                           0 non-null      float64
61  open_il_6m                            0 non-null      float64
62  open_il_12m                           0 non-null      float64
63  open_il_24m                           0 non-null      float64
64  mths_since_rcnt_il                    0 non-null      float64
65  total_bal_il                           0 non-null      float64
66  il_util                                0 non-null      float64
67  open_rv_12m                           0 non-null      float64
68  open_rv_24m                           0 non-null      float64
69  max_bal_bc                             0 non-null      float64
70  all_util                                0 non-null      float64
71  total_rev_hi_lim                       396009 non-null  float64
72  inq_fi                                 0 non-null      float64
73  total_cu_tl                             0 non-null      float64
74  inq_last_12m                           0 non-null      float64

dtypes: float64(46), int64(7), object(22)
memory usage: 266.8+ MB
```

Observation:

- In columns **annual\_inc\_joint**, **dti\_joint**, **verification\_status\_joint**, **open\_acc\_6m**, **open\_il\_6m**, **open\_il\_12m**, **open\_il\_24m**, **mths\_since\_rcnt\_il**, **inq\_last\_12m**, etc. all rows are null, will be dropped first
- In columns **desc**, **mths\_since\_last\_delinq**, **mths\_since\_last\_record**, **next\_payment\_d**, and **mths\_since\_last\_major\_derog** more than half of the rows have null values, may be imputed during preprocessing
- Columns **emp\_title**, **emp\_length**, **tot\_coll\_amt**, **tot\_cur\_bal**, and **total\_rev\_hi\_lim** have not too many null values, will be imputed during pre-processing
- In **title**, **earliest\_cr\_line**, **inq\_last\_6mths**, **open\_acc**, **pub\_rec**, **revol\_util**, **total\_acc**, **last\_pymnt\_d**, **pullasd\_credit**, **null\_credit** columns dropped during pre processing

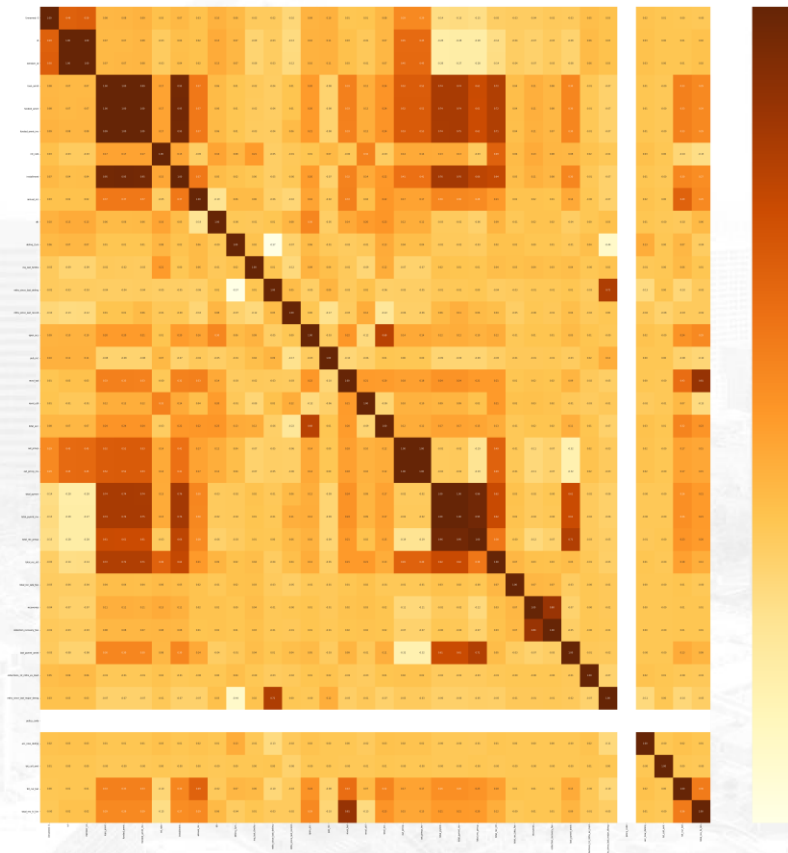
df[cats].describe()

	term	grade	sub_grade	emp_title	emp_length	home_ownership	verification_status	issue_d	loan_status	pymnt_plan	url	desc	purpose	title	zip_code	addr_state	earliest_cr_line	initial_list_status	last_pymnt_d	next_pymnt_d	last_credit_pull_d	application_type
count	466285	466285	466285	438697	445277	466285	466285	466285	466285	466285	466285	125983	466285	466265	466285	466285	466256	466285	465909	239071	466243	466285
unique	2	7	35	205475	11	6	3	91	9	2	466285	124436	14	63099	888	50	664	2	98	100	103	1
top	36 months	B	B3	Teacher	10+ years	MORTGAGE	Verified	Oct-14	Current	n	https://www.lendingclub.com/browse/loanDetail...	debt_consolidation	Debt consolidation	945xx	CA	Oct-00	f	Jan-16	Feb-16	Jan-16	INDIVIDUAL	
freq	337953	136929	31686	5399	150049	235875	168055	38782	224226	466276	1	234	274195	164075	5304	71450	3674	303005	179620	208393	327699	466285

Observation :

- The **Term** feature is dominated by **36 months, with a percentage of 72.4%**
- The **Grade** feature is dominated by **customers with grade B, with a percentage of 29.3%**
- The **sub\_grade** feature is dominated by **customers with grade B3, with a percentage of 6.7%**
- **1.1% of customers work as teachers**
- **43.04% of customers have worked for more than 10 years**
- **50.5% of customers have a residence with Mortgage status**
- **36% of customers have verified all their files**
- **48% of customers are currently on loan**
- **Total of 58.8% of customers took credit due to debt consolidation**





There are still too many features in this dataset; a Feature Selection is required to see the correlation between tables more clearly.

## Credit Risk Analysis

In this section, we will explain which types of credit risk of default we believe exist. Of course, default will occur at some point.

Some of the features that we believe can be examined from the lender's perspective are as follows:

- Loan amount proposed by Customer (**loan\_amnt**)
- The amount collected by the lender (**funded\_amnt**)
- Amount of money given to lenders by investors (**funded\_amnt\_inv**)
- Amount already paid to the lender (**total\_pymnt**)

Some of the characteristics that we believe are important in determining credit risk for customers applying for loans are as follows:

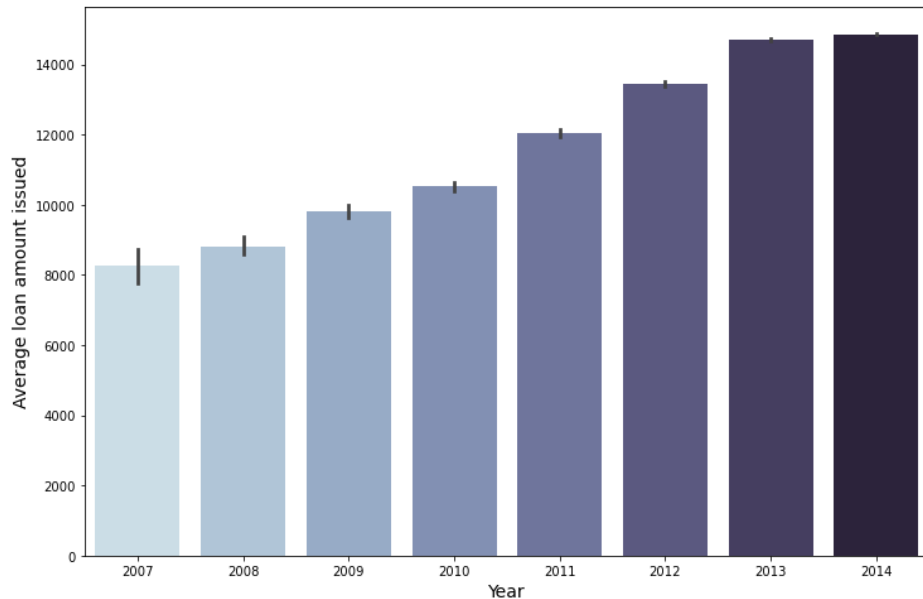
- Grade of Customer Credit Risk (**grade**)
- Annual Income for Credit Application Customers (**annual\_inc**)
- The reason for borrowing money (**purpose**).
- The status of home ownership (**home\_ownership**)
- Employment status (**emp\_title**)
- Work length (**emp\_length**)
- Term loan (**term**)
- The number of incidents of 30+ days in arrears in the borrower's credit file in the last two years (**delinq\_2yrs**).

### Annually, the average loan amount issued rises.

In the graphic image below, it can be seen that a significant increase indicates that many customers are taking loans and can be issued

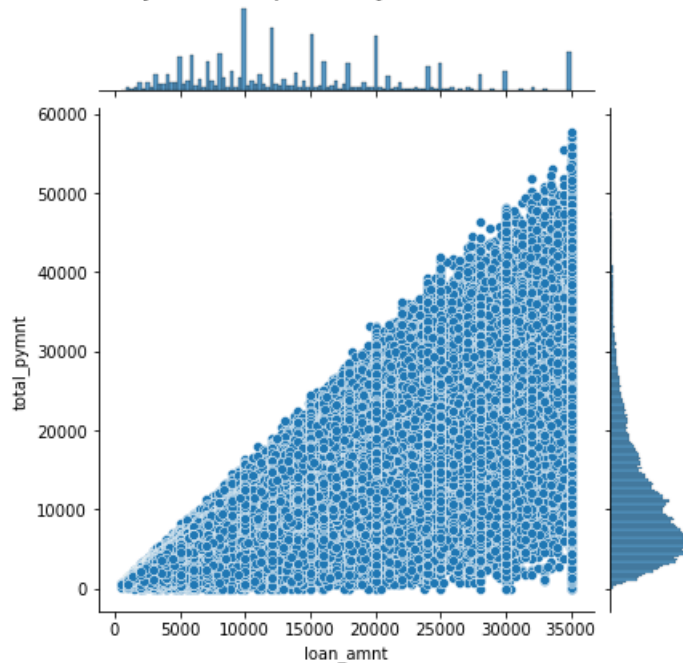


Observation:  
The average loan amount is **rising**  
**year by year**



### Loan Amount vs Total Payment

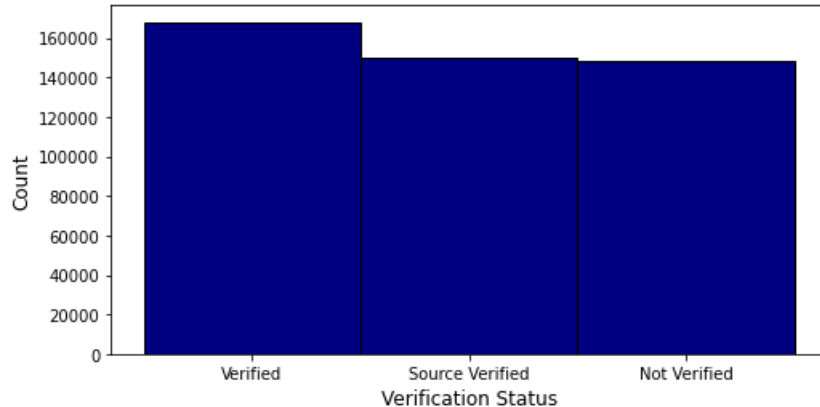
We see that feature Loan Amount and Total Payment are positively correlated



Observation:  
Loan Amount and Total Payment  
have a **positive correlation**

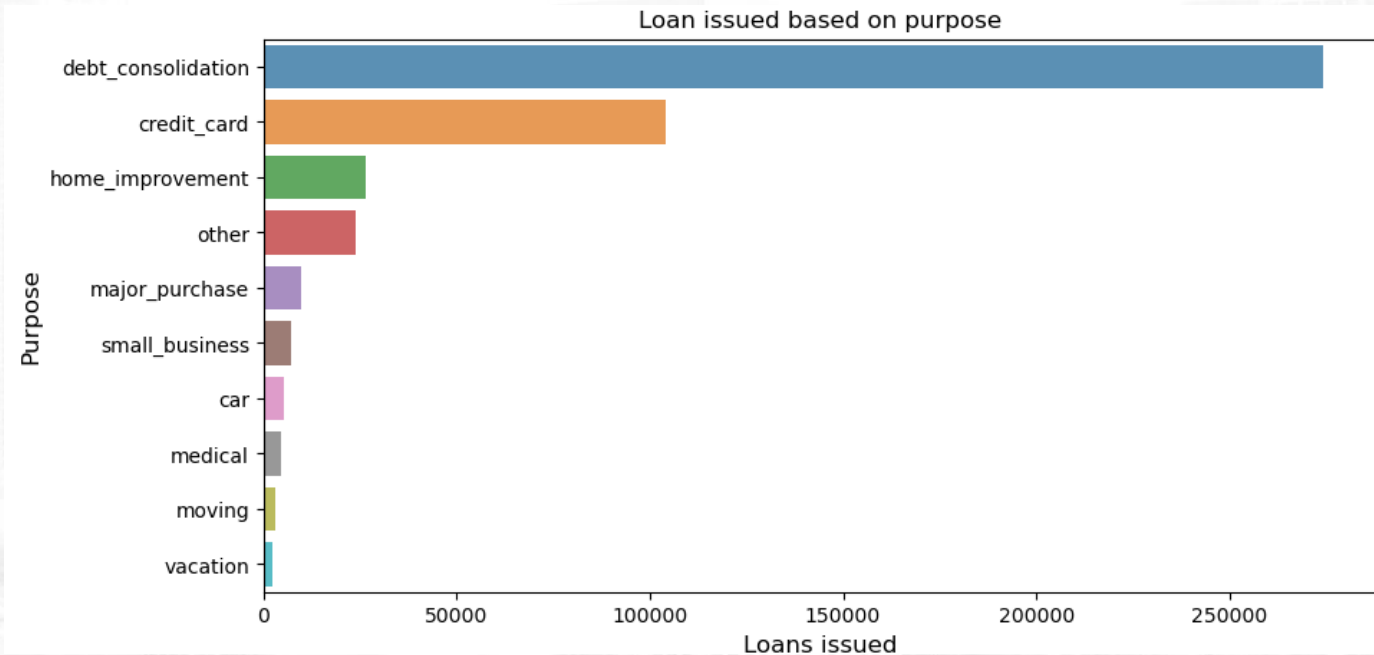
### Distribution of Verification Status

We see that the category of verification status is fairly distributed.



Observation:

- **36% customer is Verified**
- **31% customer is Not Verified**



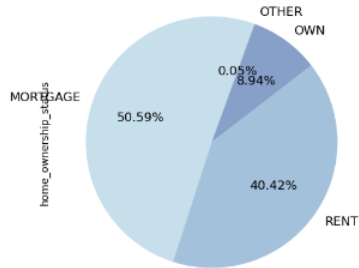
Observation:

- **Total of 58.8% of customers took credit due to debt consolidation**

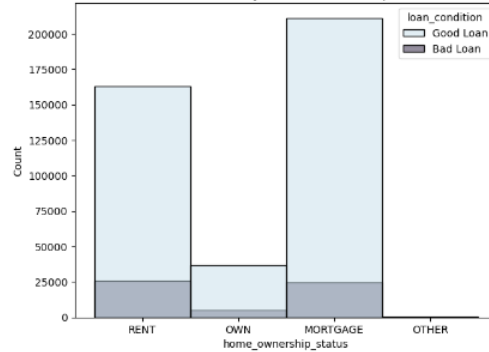


## Information on Home Ownership Status

Borrower Percentage by Home Ownership Status



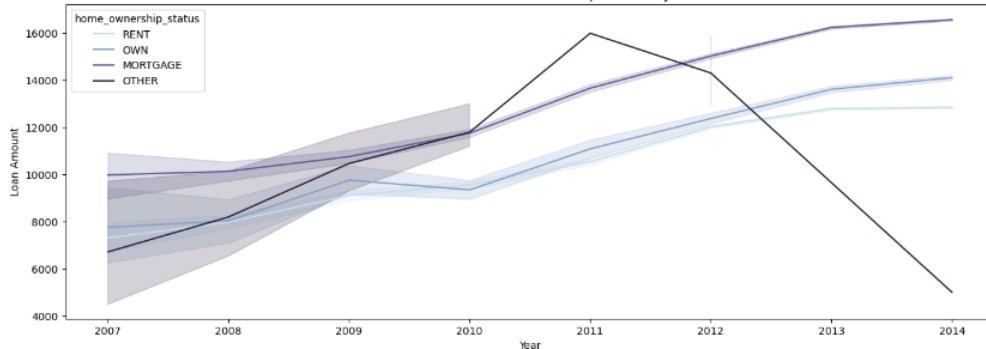
Loan Condition by Home Ownership Status



Observation:

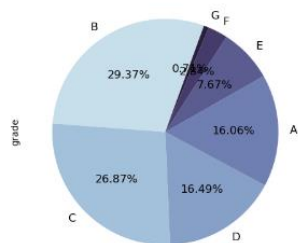
- **50.59%** of customers have a residence with **Mortgage status**

Loan Amount to Home Ownership Status by Year

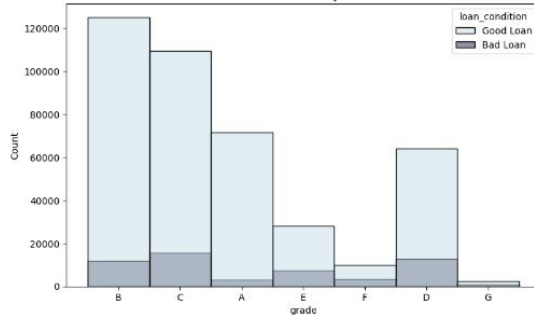


## Information on Credit Score

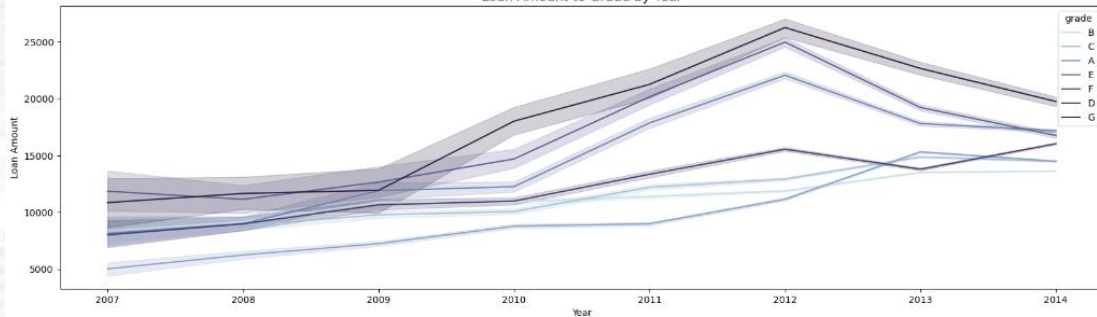
Borrower Percentage by Grade



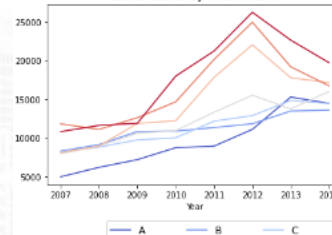
Loan Condition by Grade



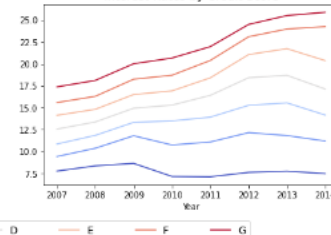
Loan Amount to Grade by Year



Loans issued by Credit Score



Interest Rates by Credit Score

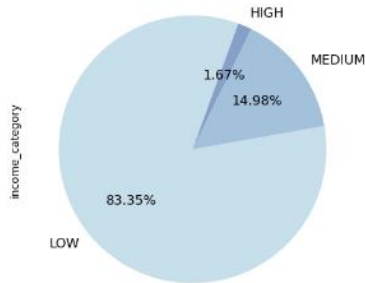


## Observation:

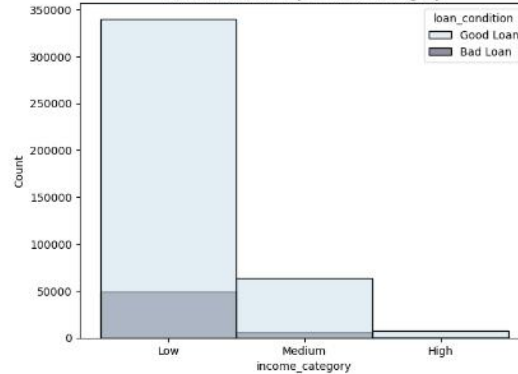
- The **Grade** feature is dominated by customers with grade **B**, with a **percentage of 29.37%**
- Customers with grade **G** tend to take out more loans than other Credit Score
- **Feature Grade** needs to be considered in the decision, whether customer loans can be good loaner or not. There are indications that customers with a 'G' credit score are at risk of default.

## Information on Income Category

Borrower Percentage by Income Category



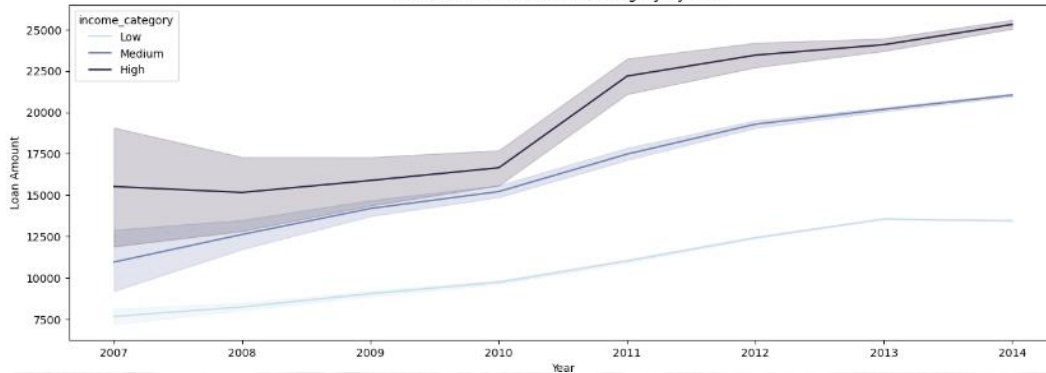
Loan Condition by Income Category



Observation:

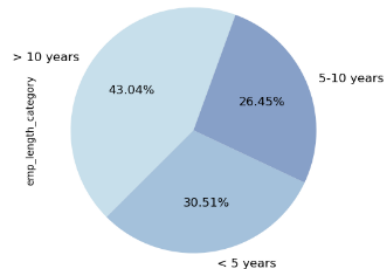
- The majority of customers are in the **Low Income Group, with a percentage of 83.35%**
- Customers with **High Income Group tend to take out more loans than other Income Groups**

Loan Amount to Income Category by Year

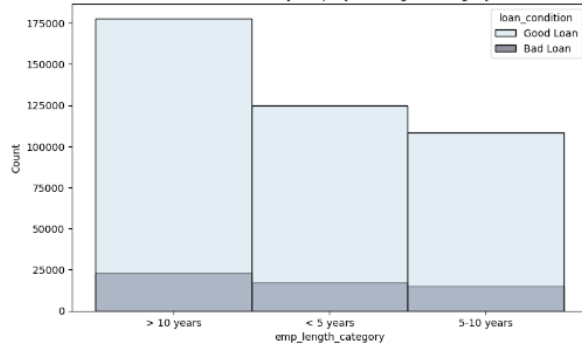


## Information on Employee Length Category

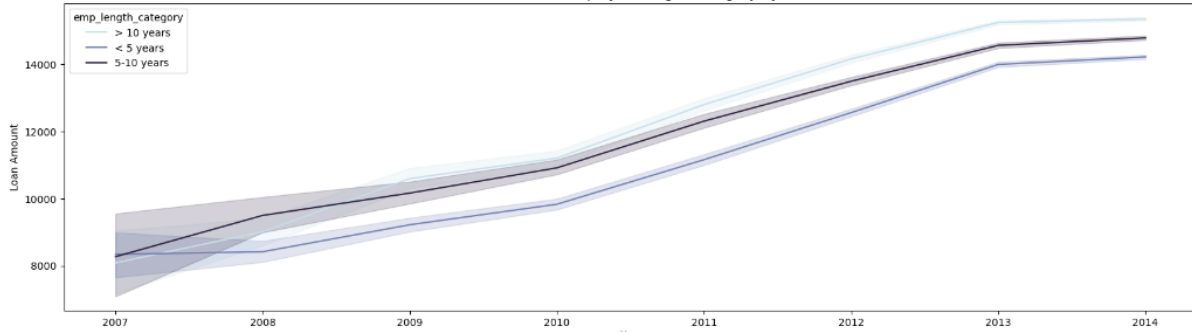
Borrower Percentage by Employee Length Category



Loan Condition by Employee Length Category



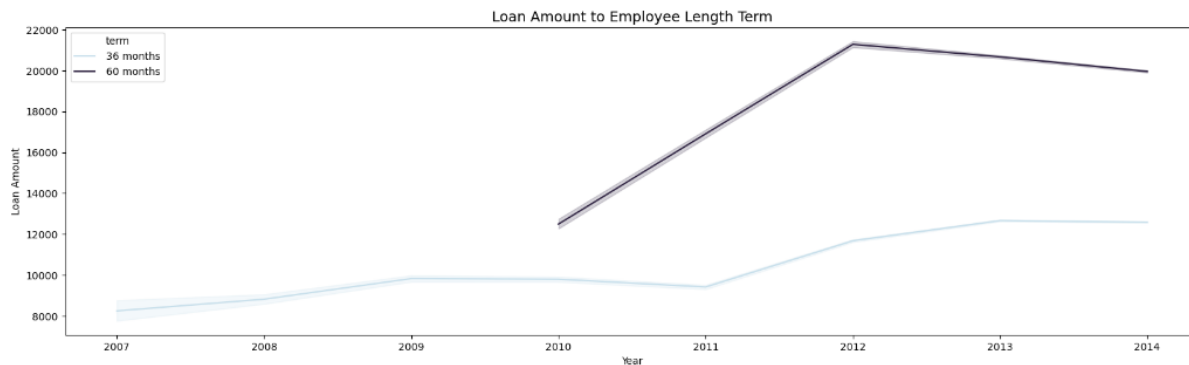
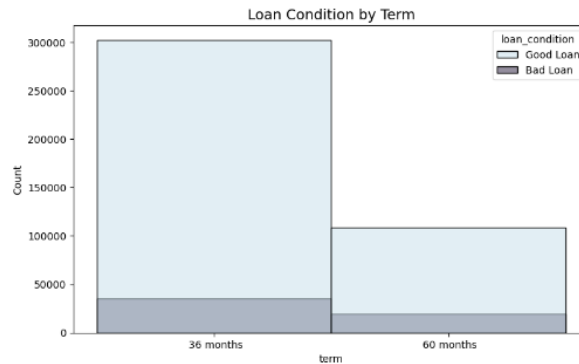
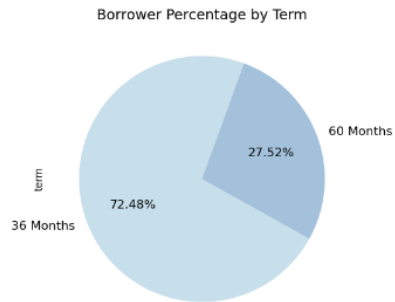
Loan Amount to Employee Length Category by Year



## Observation:

- 43.04% of customers have worked for more than 10 years
- **Customers who have worked for more than 10 years more likely to loan higher**

## Information on Term

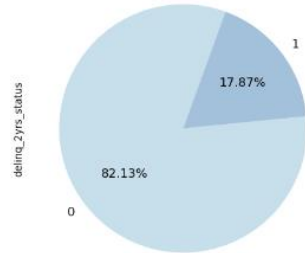


## Observation:

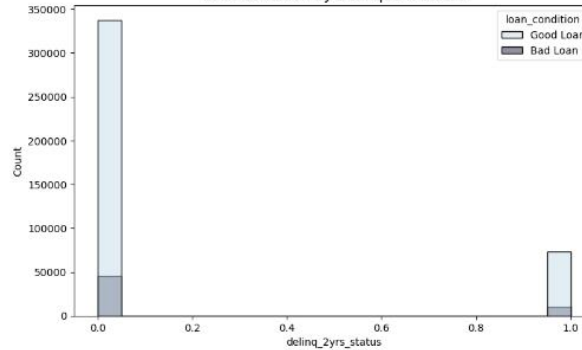
- Total of **72.48%** of customers choose a payment term of **36 months**
- Customers who choose a payment term of **60 months** tend to take more loans than customers with a payment term of **36 months**

## Information on Delinquent Status

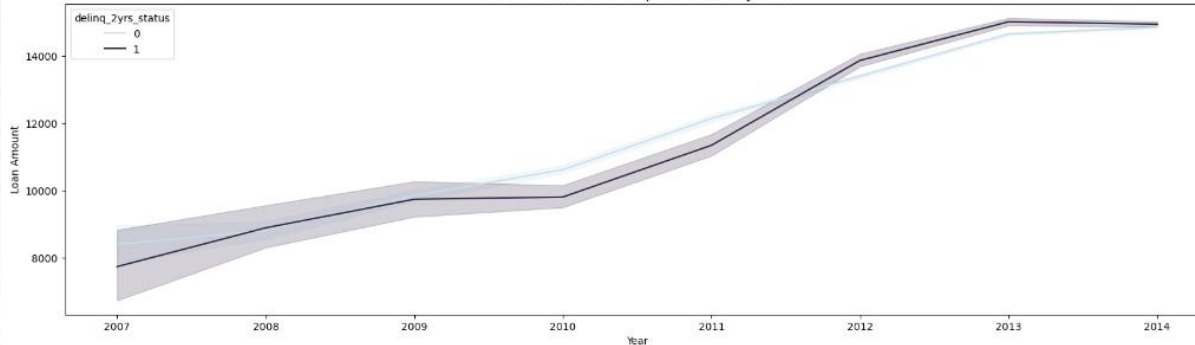
Delinquent Status Percentage



Loan Condition by Delinquent Status



Loan Amount to Delinquent Status by Year



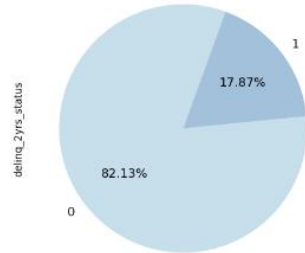
## Observation:

- **82,13% Customers have no arrears status in the last 2 years**
- Customers who have delinquent status tend to loan more than customers who do not have arrears

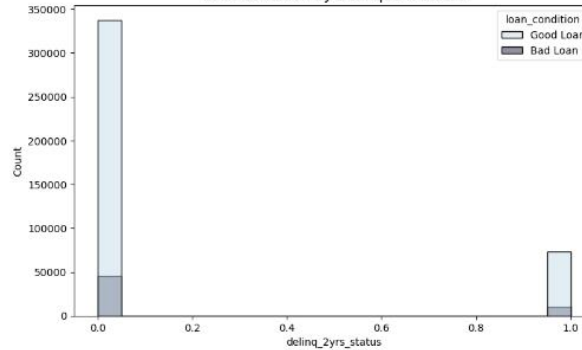


## Information on Delinquent Status

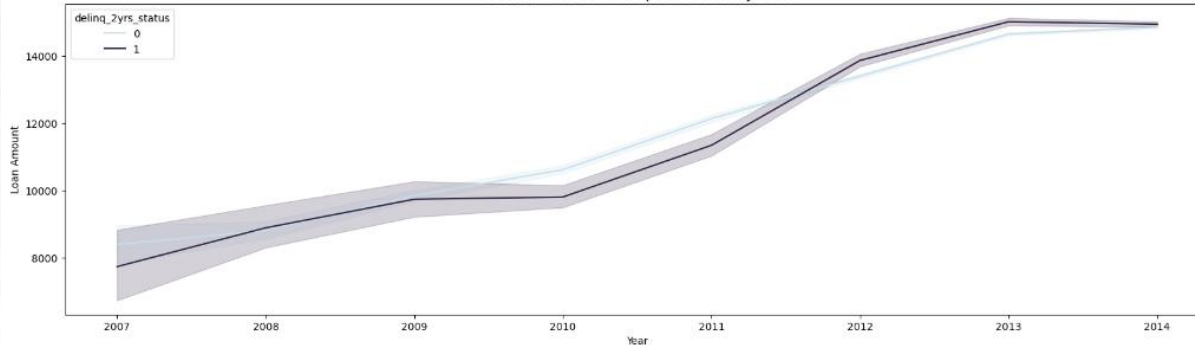
Delinquent Status Percentage



Loan Condition by Delinquent Status



Loan Amount to Delinquent Status by Year



## Observation:

- **82,13% Customers have no arrears status in the last 2 years**
- Customers who have delinquent status tend to loan more than customers who do not have arrears

## Feature Selection

Feature that we need is:

- **loan\_amnt**
- **term**
- Grade
- emp\_length\_category
- home\_ownership\_status
- income\_category
- verification\_status
- loan\_condition
- purpose
- delinq\_2yrs\_status

We set **loan\_condition** as a Target for this modelling

```
df_selection.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 466285 entries, 0 to 466284
Data columns (total 10 columns):
#   Column              Non-Null Count  Dtype  
---  -
0   loan_amnt           466285 non-null  int64  
1   term                466285 non-null  object  
2   grade               466285 non-null  object  
3   emp_length_category 466285 non-null  object  
4   home_ownership_status 466285 non-null  object  
5   income_category      466285 non-null  object  
6   verification_status  466285 non-null  object  
7   loan_condition       466285 non-null  object  
8   purpose              466285 non-null  object  
9   delinq_2yrs_status   466285 non-null  int64  
dtypes: int64(2), object(8)
memory usage: 35.6+ MB
```

## Data Cleaning

```
df_selection.isna().sum()

loan_amnt      0
term            0
grade           0
emp_length_category 0
home_ownership_status 0
income_category 0
verification_status 0
loan_condition  0
purpose         0
delinq_2yrs_status 0
dtype: int64
```

Data is Clean without Null Values

## Encoding

```
df_selection.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 466285 entries, 0 to 466284
Data columns (total 36 columns):
 #   Column                                Non-Null Count  Dtype  
---  --
 0   loan_amnt                             466285 non-null  int64  
 1   term                                 466285 non-null  object  
 2   grade                                466285 non-null  int64  
 3   emp_length_category                  466285 non-null  object  
 4   home_ownership_status               466285 non-null  object  
 5   income_category                     466285 non-null  int64  
 6   verification_status                 466285 non-null  object  
 7   loan_condition                      466285 non-null  int64  
 8   purpose                             466285 non-null  object  
 9   delinq_2yrs_status                  466285 non-null  int64  
10  term_36_months                      466285 non-null  uint8  
11  term_60_months                      466285 non-null  uint8  
12  emp_length_category_5-10 years      466285 non-null  uint8  
13  emp_length_category_< 5 years      466285 non-null  uint8  
14  emp_length_category_> 10 years      466285 non-null  uint8  
15  home_ownership_status_MORTGAGE      466285 non-null  uint8  
16  home_ownership_status_OTHER         466285 non-null  uint8  
17  home_ownership_status_OWN           466285 non-null  uint8  
18  home_ownership_status_RENT          466285 non-null  uint8  
19  verification_status_Not Verified    466285 non-null  uint8  
20  verification_status_Source Verified  466285 non-null  uint8  
21  verification_status_Verified        466285 non-null  uint8  
22  purpose_car                         466285 non-null  uint8  
23  purpose_credit_card                 466285 non-null  uint8  
24  purpose_debt_consolidation          466285 non-null  uint8  
25  purpose_educational                 466285 non-null  uint8  
26  purpose_home_improvement            466285 non-null  uint8  
27  purpose_house                       466285 non-null  uint8  
28  purpose_major_purchase              466285 non-null  uint8  
29  purpose_medical                     466285 non-null  uint8  
30  purpose_moving                      466285 non-null  uint8  
31  purpose_other                       466285 non-null  uint8  
32  purpose_renewable_energy            466285 non-null  uint8  
33  purpose_small_business              466285 non-null  uint8  
34  purpose_vacation                    466285 non-null  uint8  
35  purpose_wedding                     466285 non-null  uint8  
dtypes: int64(5), object(5), uint8(26)
memory usage: 47.1+ MB
```

Encoding Strategy for Feature Selection:

- Label Encoding: **grade** and **income\_category**
- One Hot Encoding: **term**, **emp\_length\_category**, **home\_ownership\_status**, **verification\_status**, **purpose**



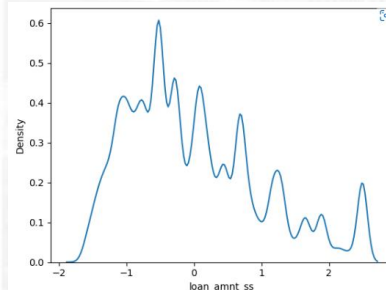
## Standardization

```
list_col_abnormal = []

for i in df_selection:
    if df_selection[i].min() != 0 and df_selection[i].max() != 1:
        list_col_abnormal.append(i)

print(list_col_abnormal)

['loan_amnt']
```



Standardization  
loan\_amnt column

## Model Comparison

Evaluation	Model			
	Logistic Regression (Default)	Decision Tree (Default)	Decision Tree (HP)	Random Forest (Default)
Accuracy (Test Set)	0,88	0,83	0,88	0,85
Accuracy (Train Set)	0,88	0,93	0,88	0,93
Precision (Test Set)	0,88	0,89	0,88	0,88
Precision (Train Set)	0,88	0,94	0,88	0,93
Recall (Test Set)	1,00	0,92	1,00	0,95
Recall (Train Set)	1,00	0,98	1,00	0,99
F1-Score (Test Set)	0,94	0,90	0,94	0,92
F1-Score (Train Set)	0,94	0,96	0,94	0,96

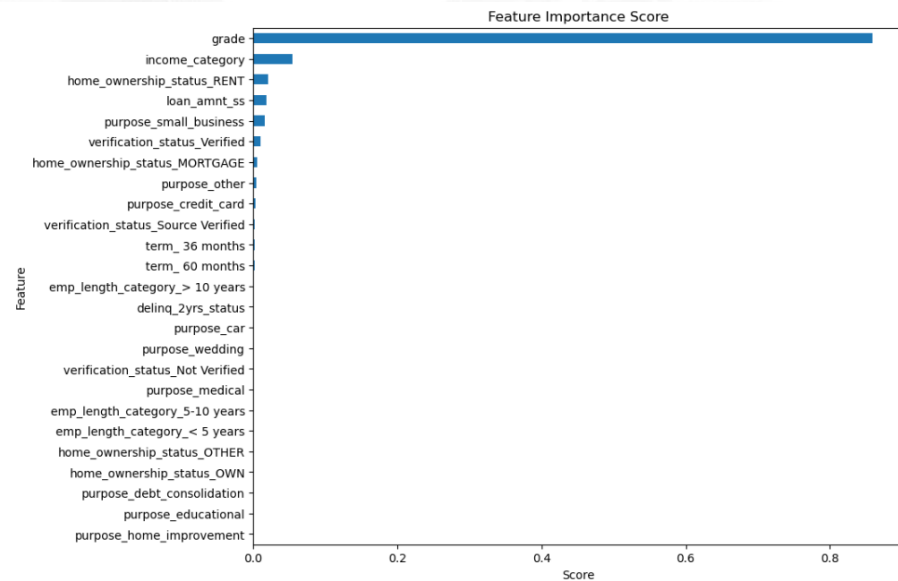
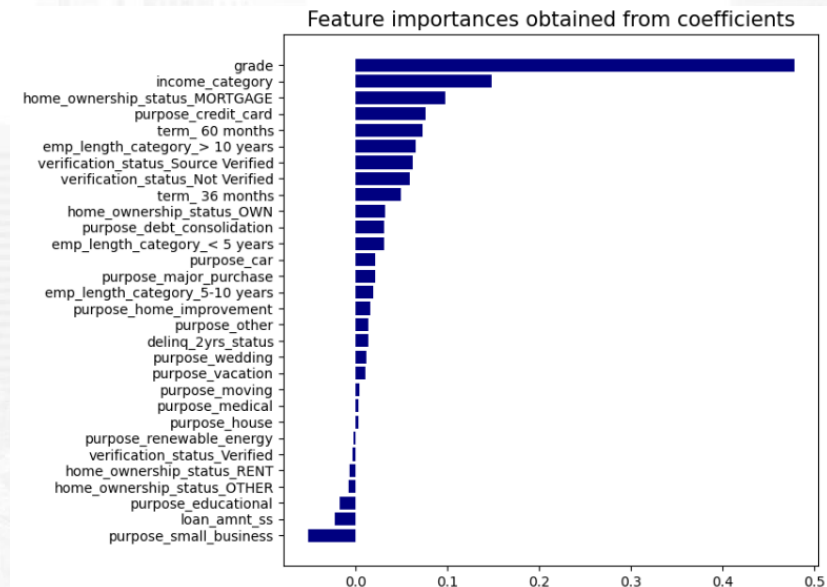
Model Evaluation:

### Precision

From the results of the comparison above, we determine that the **Decision Tree and Logistic Regression** are the models to be used, because the Precision value is quite high and best fit. We will look at the feature importance to make credit risk management

For more details, you can see jupyter notebook [here](#)

## Feature Importance



As we can see, the Decision Tree results after Hyperparameter Tuning and the default Logistic Regression results have the same evaluation value (Precision) but differ in Feature Importance. Based on exploratory data analysis and feature importance, we make two recommendations for credit risk management.



## Credit Risk Management with Log Reg Model (Based on Feature Importance and EDA):

1. The most important factor in determining whether a borrower is eligible for a Good Loan or a Bad Loan is their credit score. According to EDA results, Credit Score **6** is the lowest grade that is risky and may result in a default.
2. Income is the second most important factor in determining whether a loan is good or bad. According to the EDA results, bad loaners are dominated by Low Income, but this needs to be reviewed further with several other features.
3. Mortgage home owners are preferred borrowers. Borrowers with Mortgage home status, according to our analysis, are more accustomed to paying an installment.
4. Credit card borrowers are prioritized over other types of borrowers. Credit card users are accustomed to making installment payments, and the risk of default may be reduced as a result.
5. Borrowers with a 60-month repayment period are more likely to have their loans approved. The longer the loan period, the easier it is to pay the installments with low installments to make it easier for customers and reduce the risk of default. The interest rate rises as the payment period lengthens.
6. Loan Amount is a relatively minor feature. The greater the loan amount requested by the customer, the greater the risk of default.



### Credit Risk Management with Decision Tree Model (Based on Feature Importance and EDA):

1. The most important factor in determining whether a borrower is eligible for a Good Loan or a Bad Loan is their credit score. According to EDA results, Credit Score **\*\*G\*\*** is the lowest grade that is risky and may result in a default.
2. Income is the second most important factor in determining whether a loan is good or bad. According to the EDA results, bad loaners are dominated by Low Income, but this needs to be reviewed further with several other features.
3. Customers who own rental properties are more likely to have their loans approved. Customers with rental housing status, according to the analysis, are more accustomed to paying in installments. Customers with mortgage home ownership status are also eligible for approval.
4. Customers who want to borrow money to start a small business are more likely to be approved based on the Decision Tree model results. Furthermore, several lending purposes, such as credit cards and other purposes, can be considered.
5. Customers with approved loan verification status are given priority consideration for loan approval. Customers with "Source Verified" verification status are also eligible for approval.
6. Borrowers with a 60-month repayment period are more likely to have their loans approved. The longer the loan period, the easier it is to pay the installments with low installments to make it easier for customers and reduce the risk of default. The interest rate rises as the payment period lengthens.

# Thank You



Created by:

**Muhammad Ariq Arfina**

[ariqarfina05@gmail.com](mailto:ariqarfina05@gmail.com)

LinkedIn : [Muhammad Ariq Arfina](#)

Github : [ariqarfina](#)