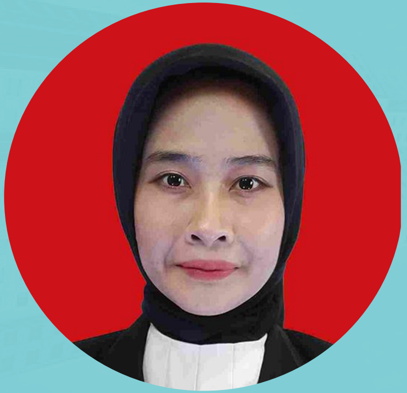
Rakamin Academy X id/x partners

Machine Learning Project

Credit Risk Assessment Using Logistic Regression

ID/X Partners Data Scientist Project Based Internship Program

Presented by
Nicken Shidqia Nurahman



Nicken Shidqia Nurahman

About Me

Civil engineer graduate with some experience in administration and project management, who is interested in data science.
Detail oriented, and time management person, and familiar with Microsoft Office, Python, SQL and Jupyter.
Motivated to continue to learn and grow as a professional.

My Experience

Rakamin Academy

Data Science Bootcamp Student – RAKAMIN ACADEMY
Oct 2023 - Now

Project Management Masters Degree Student – UNIVERSITAS INDONESIA
Sep 2021 - Sep 2023

Engineering Administration and Project Control Staff - PT. ISTAKA KARYA
Aug 2019 - Sep 2021

Project Control Intern - PT. ISTAKA KARYA
Feb 2019 - Jul 2019

Surveying Laboratory Assistant – UNIVERSITAS TRISAKTI
Jul 2017- Agust 2019

Case Study

Problem

A primary risk with corporate loans is **failing** in accurately **assessing credit risk**.

Disadvantage of Manual credit risk assessment

- **Subjectivity**
  Subjectivity can introduce bias and inconsistency in decision-making.

- **Time-Consuming**
  time-consuming especially when dealing with a large number of loan applications.

- **Risk of Error**
  Humans errors, such as data entry mistakes, miscalculations, or oversight of important details.

Challenges

Build a machine learning model that can predict credit risk assessment

Tool & Library Used

jupyter

pandas

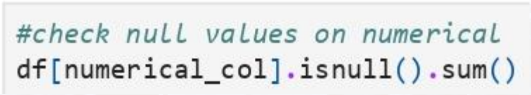
NumPy

scikit learn

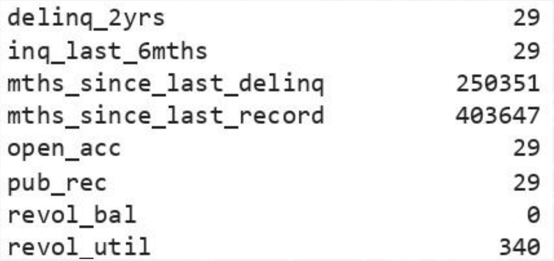
Data Cleaning

Data Cleaning

Missing Values



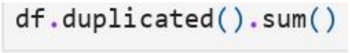
```
#check null values on numerical
df[numerical_col].isnull().sum()
```



```
delinq_2yrs                   29
inq_last_6mths                29
mths_since_last_delinq    250351
mths_since_last_record    403647
open_acc                      29
pub_rec                       29
revol_bal                      0
revol_util                   340
```

There are 32 numerical column and 7 categorical columns that have null values

Duplicated Values



```
df.duplicated().sum()
```

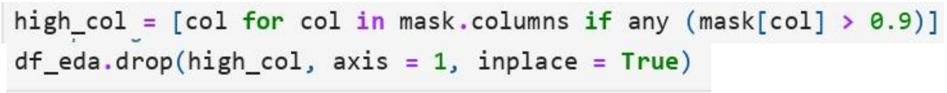
```
0
```

No duplicated value

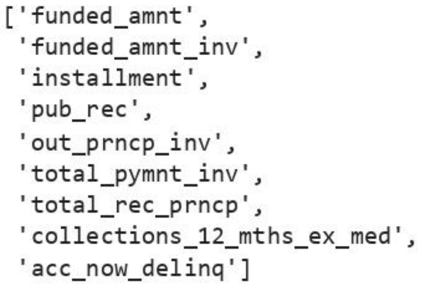
Handling missing value

- Drop feature that have missing value > 50%
- Replace missing values on feature tot_coll_amt, tot_cur_bal, total_rev_hi_lim with 0
- Replace missing values on numerical category with median
- Replace missing values on categorical with mode

Drop High Correlated Column



```
high_col = [col for col in mask.columns if any (mask[col] > 0.9)]
df_eda.drop(high_col, axis = 1, inplace = True)
```



```
['funded_amnt',
 'funded_amnt_inv',
 'installment',
 'pub_rec',
 'out_prncp_inv',
 'total_pymnt_inv',
 'total_rec_prncp',
 'collections_12_mths_ex_med',
 'acc_now_delinq']
```

Correlation coefficients whose magnitude are between 0.7 and 0.9 indicate variables which can be considered highly correlated

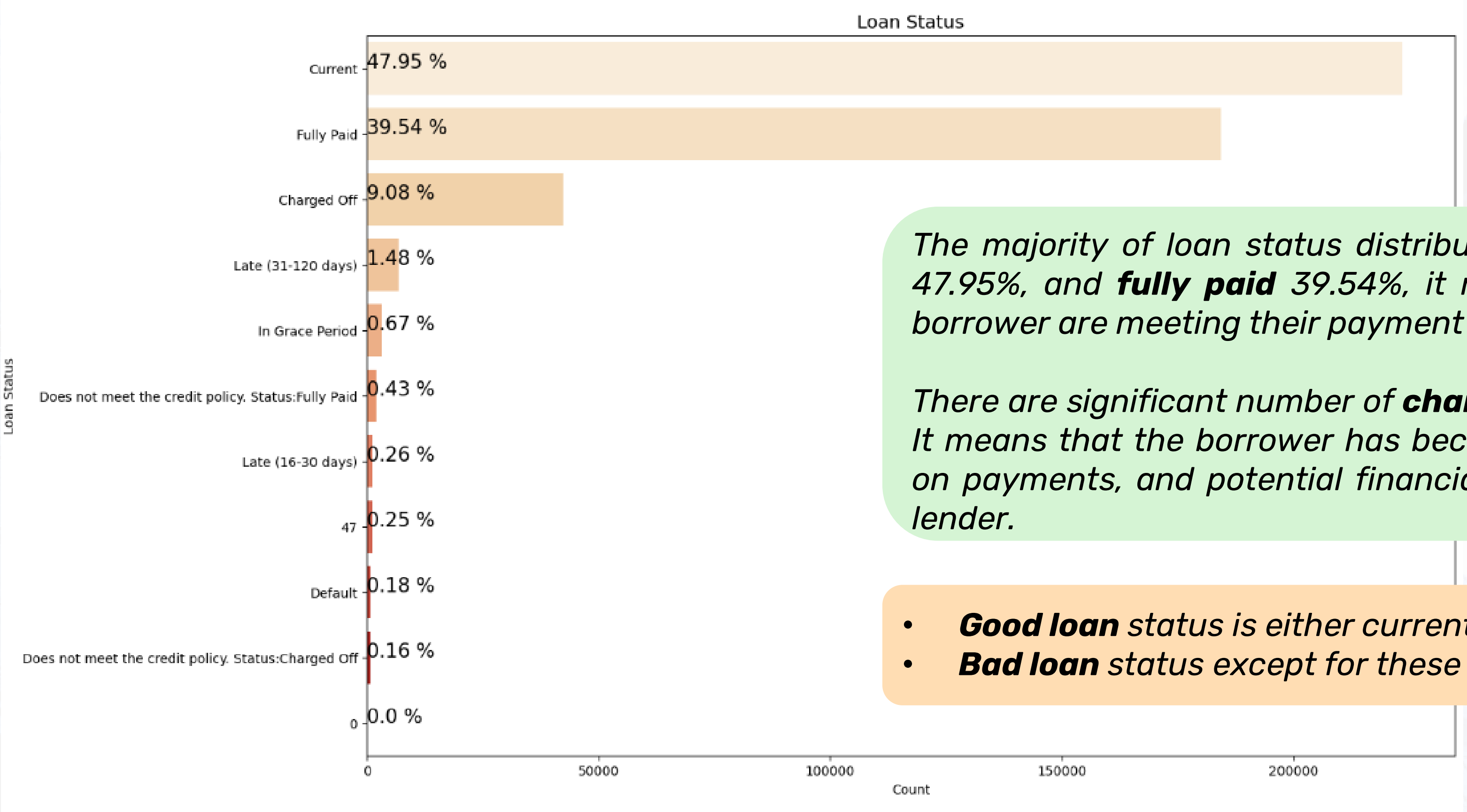
Correlation > 0.9 are gonna removed

Exploratory Data Analysis (Univariate Analysis)

Exploratory Data Analysis

Rakamin Academy

Applicants by Loan Status



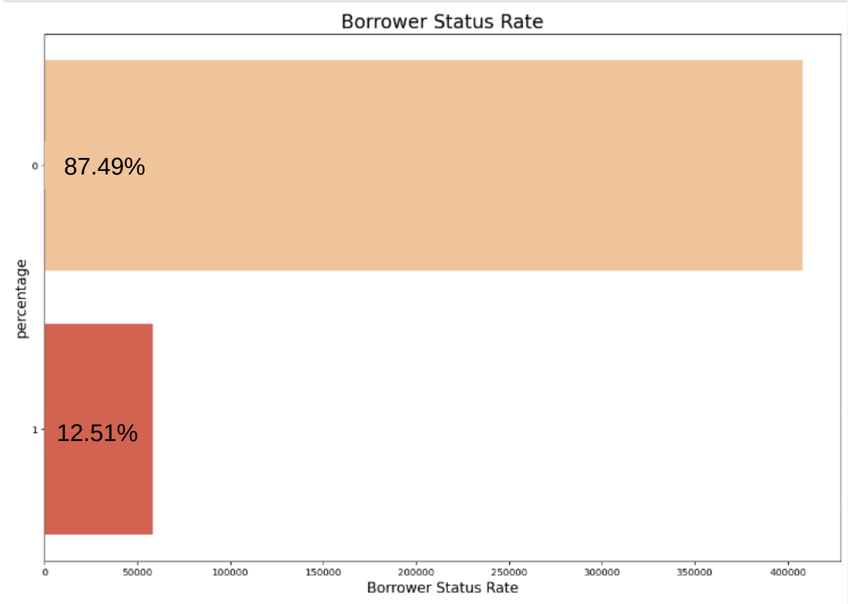
The majority of loan status distribution is ***current*** *47.95%, and* ***fully paid*** *39.54%, it means that the borrower are meeting their payment obligation.*

There are significant number of ***charged off*** *9.08%. It means that the borrower has become delinquent on payments, and potential financial loss from the lender.*

- ***Good loan*** *status is either current and fully paid.*
- ***Bad loan*** *status except for these 2 things.*

Exploratory Data Analysis

Applicants by Borrower’s Status Rate



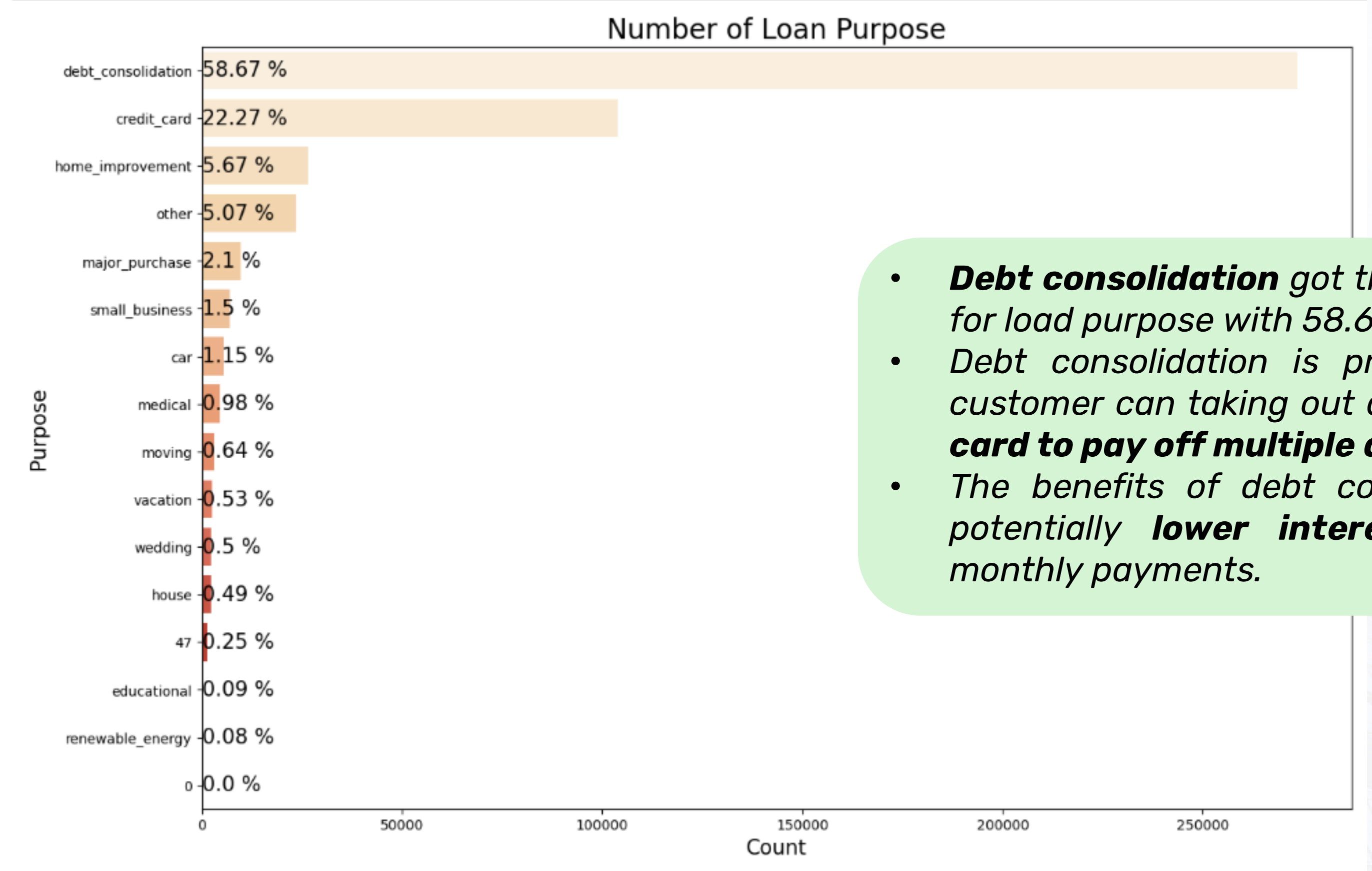
Good loan status got high percentage with 87.49%. It means that the bank's loan performing is good.

Bad loan status got low percentage with 12.51%. It means the bank need to analyzing the characteristic of the borrower, so they could identify early warnings sign, and implement the mitigation from failure of pay loans from customers

Exploratory Data Analysis

Rakamin Academy

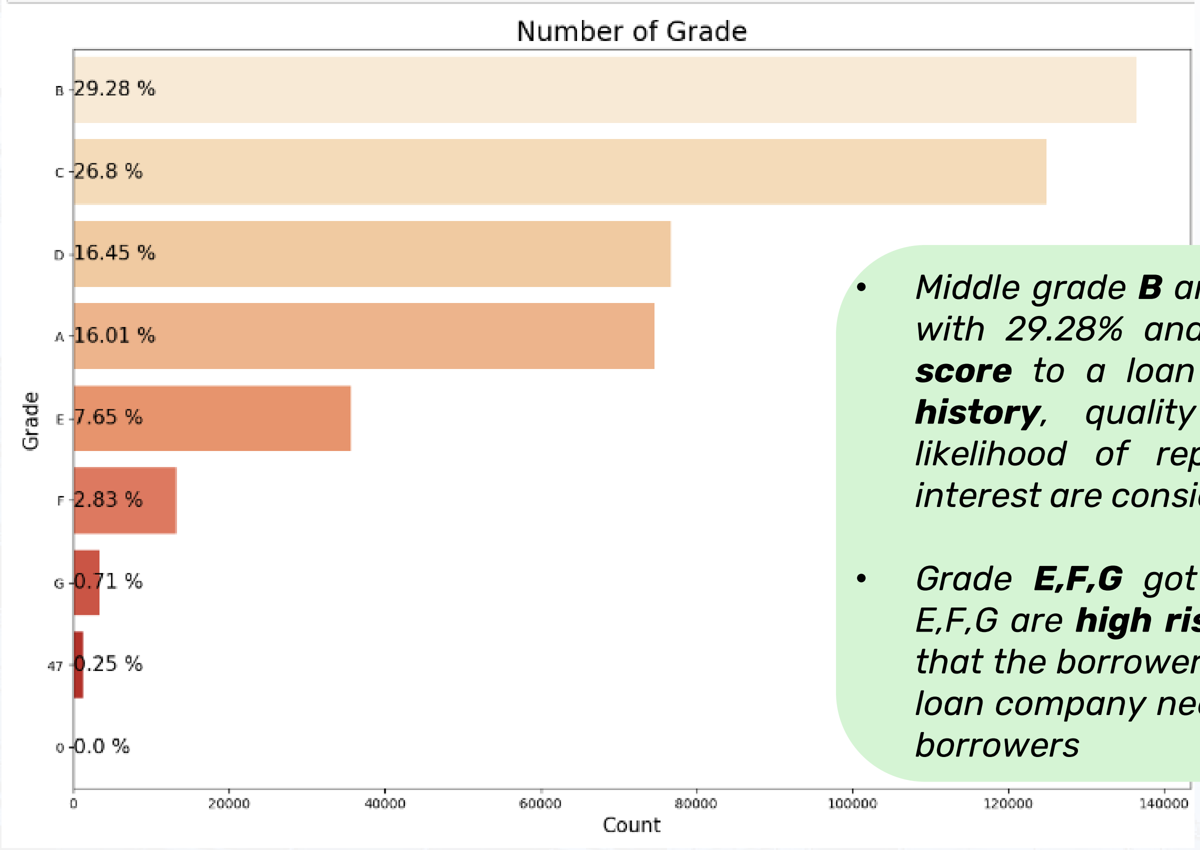
Applicants by Loan Purpose



- ***Debt consolidation*** *got the highest percentage for load purpose with 58.67%.*
- *Debt consolidation is preferred because the customer can taking out a single loan or* ***credit card to pay off multiple debts.***
- *The benefits of debt consolidation include a potentially* ***lower interest rate*** *and lower monthly payments.*

Exploratory Data Analysis

Applicants by Grade



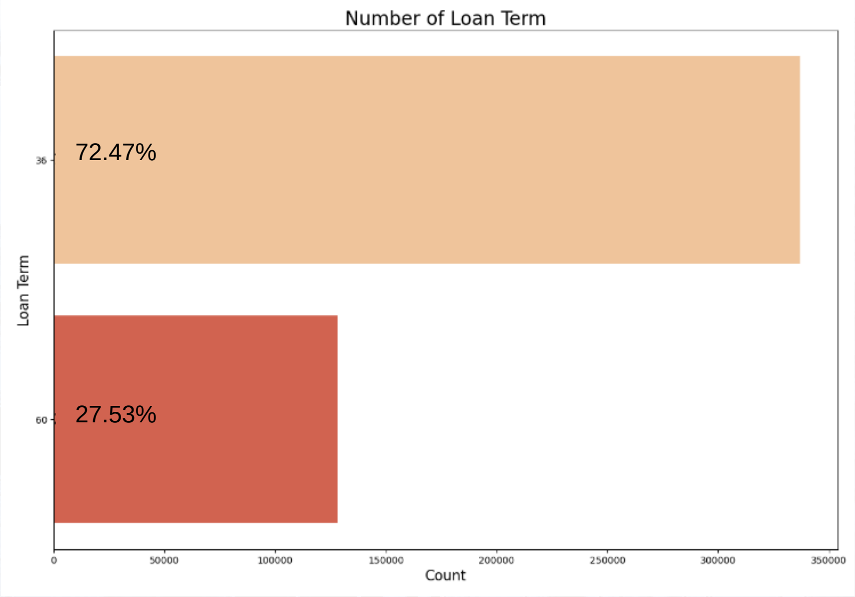
- *Middle grade **B** and **C** got the highest percentage with 29.28% and 26.8%. It means that **quality score** to a loan based on a borrower's **credit history**, quality of the collateral, and the likelihood of repayment of the principal and interest are considered moderate*

- *Grade **E,F,G** got the lowest percentage. Grade E,F,G are **high risk grade**, because the likelihood that the borrower will repay the loan is low. So the loan company need to tighten the criteria for loan borrowers*

Exploratory Data Analysis

Rakamin Academy

Applicants by Loan Term

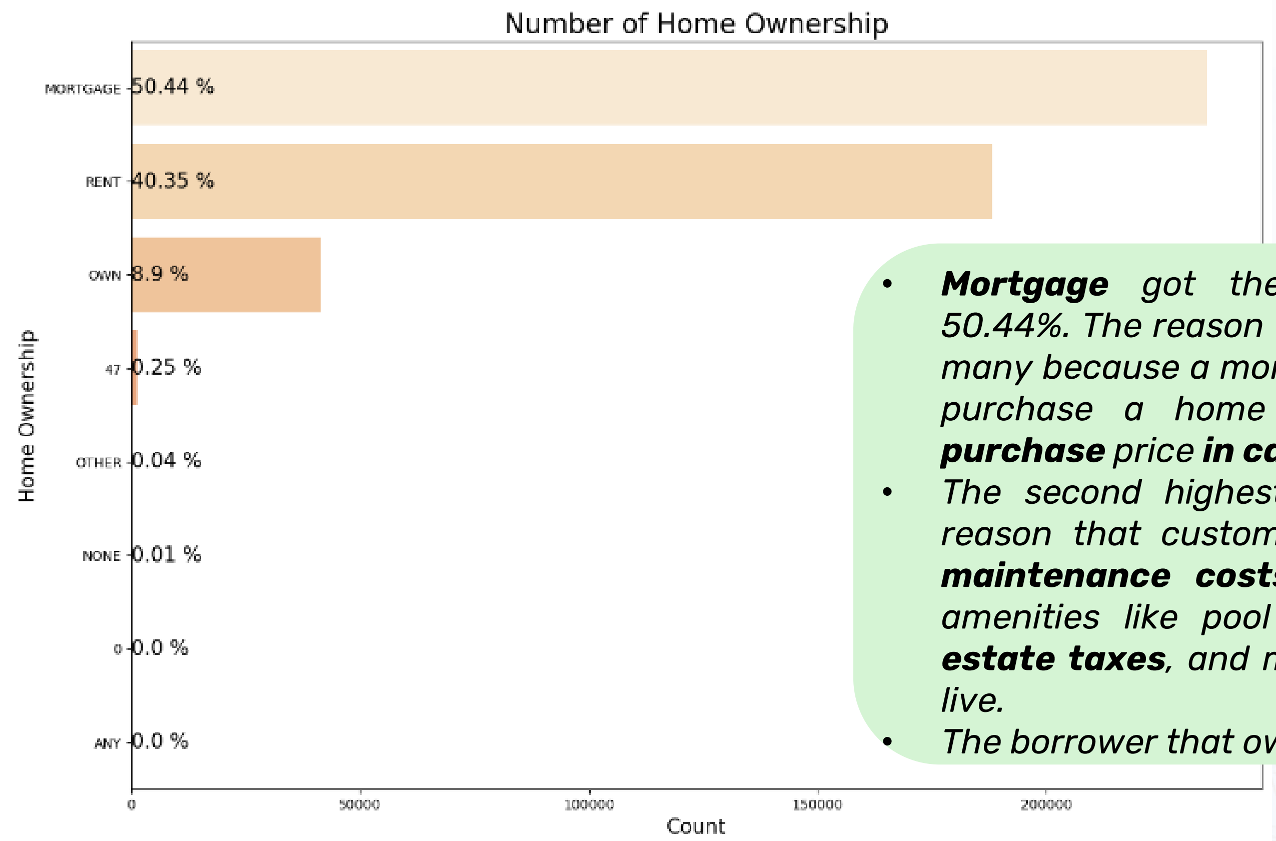


36 month *of loan term got the highest percentage with 72.47%.*
It means that ***short term loan are preferred*** *by borrowers rather than long term*

- *Compared to long term loans, the amount of* ***interest paid*** *is significantly* ***less****.*
- *These loans are considered less risky compared to long term loans because of a* ***shorter maturity date****.*
- *Short term loans are* ***the lifesavers of smaller businesses*** *or individuals who suffer from less than stellar credit scores*

Exploratory Data Analysis

Applicants by Home Ownership



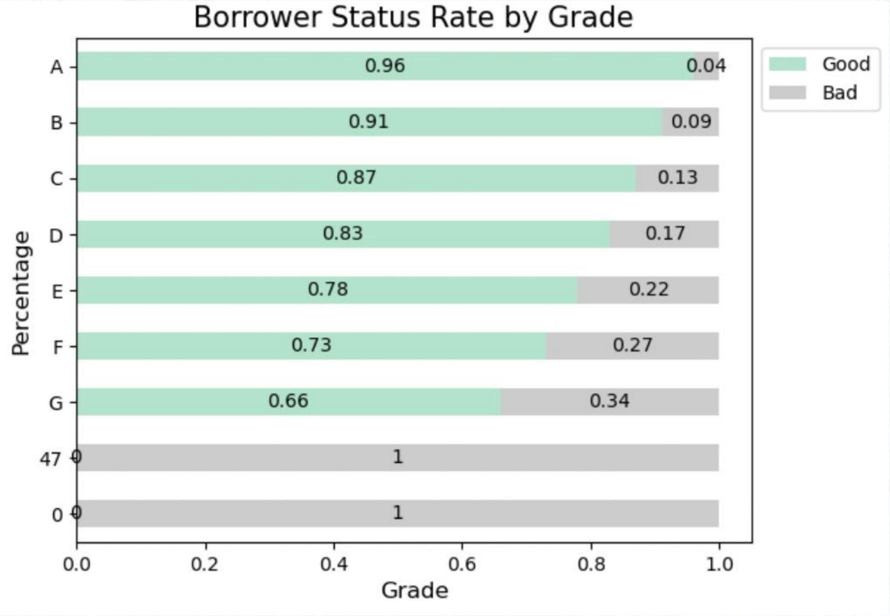
- ***Mortgage*** *got the highest percentage with 50.44%. The reason that mortgage customer is so many because a mortgage allows the customer to purchase a home* ***without paying the full purchase*** *price* ***in cash****.*
- *The second highest is* ***rent*** *with 40.35%. The reason that customer choose rent because* ***no maintenance costs*** *or repair bills, access to amenities like pool or fitness centre,* ***no real estate taxes****, and more flexibility as to where to live.*
- *The borrower that own their* ***houses*** *is only 8.9%.*

Exploratory Data Analysis (Bivariate Analysis)

Exploratory Data Analysis

Rakamin Academy

Borrower's Status Rate by Grade

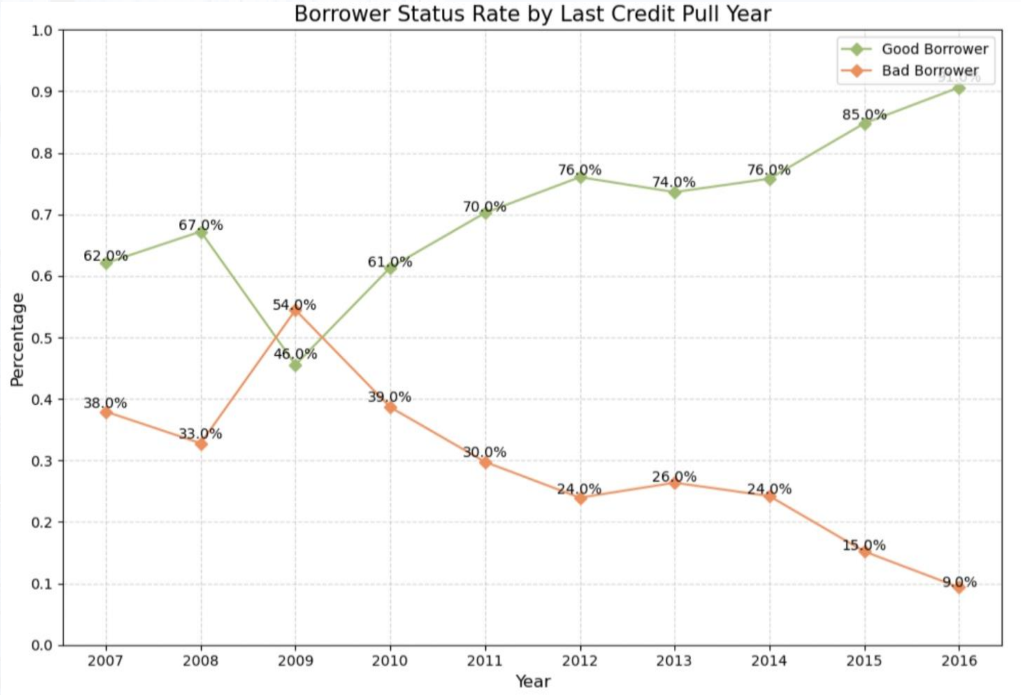


- ***Grade A*** *has the most of good borrowers with 96%, and has the least of bad borrowers with only 4%. So Grade A has the* ***least probability of loan default.***
- ***Grade G*** *has the least of good borrowers with 66%, and has the most of bad borrowers with 34%. So Grade G has the* ***most probability of loan default.***

The lower the quality of a grade, the higher the number of bad borrowers, which will lead to a higher possibility of loan default

Exploratory Data Analysis

Borrower’s Status Rate by Last Credit Pull Year



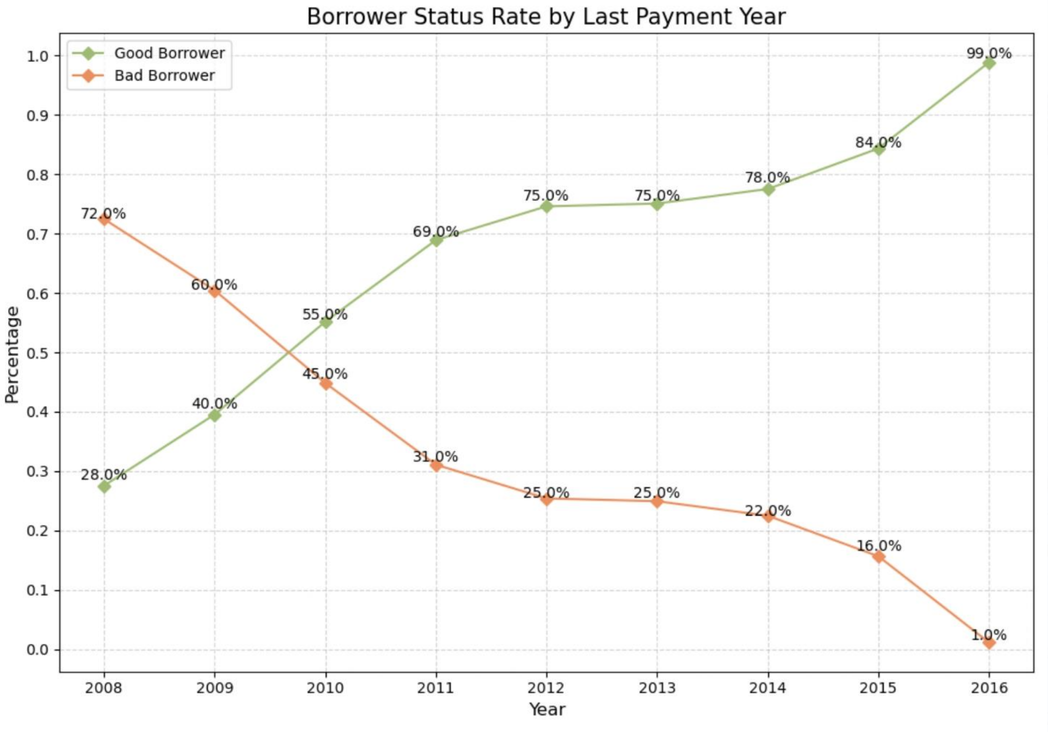
- *From 2007 to 2008, there was a slight increase 4% in the number of good borrowers. But from 2008 to 2009, there was a quite drastic decrease of 13%.*
- *The upward trend for good borrower started from 2009 to 2016. This means that many borrowers pay their loans on time.*

- *From 2007 to 2008, there was a slight decrease 5% in the number of bad borrowers. But from 2008 to 2009, there was a quite drastic increase of 21%.*
- *The downward trend for bad borrower started from 2009 to 2016.This would be good signal for lenders company.*

Exploratory Data Analysis

Rakamin Academy

Borrower's Status Rate by Last Payment Date



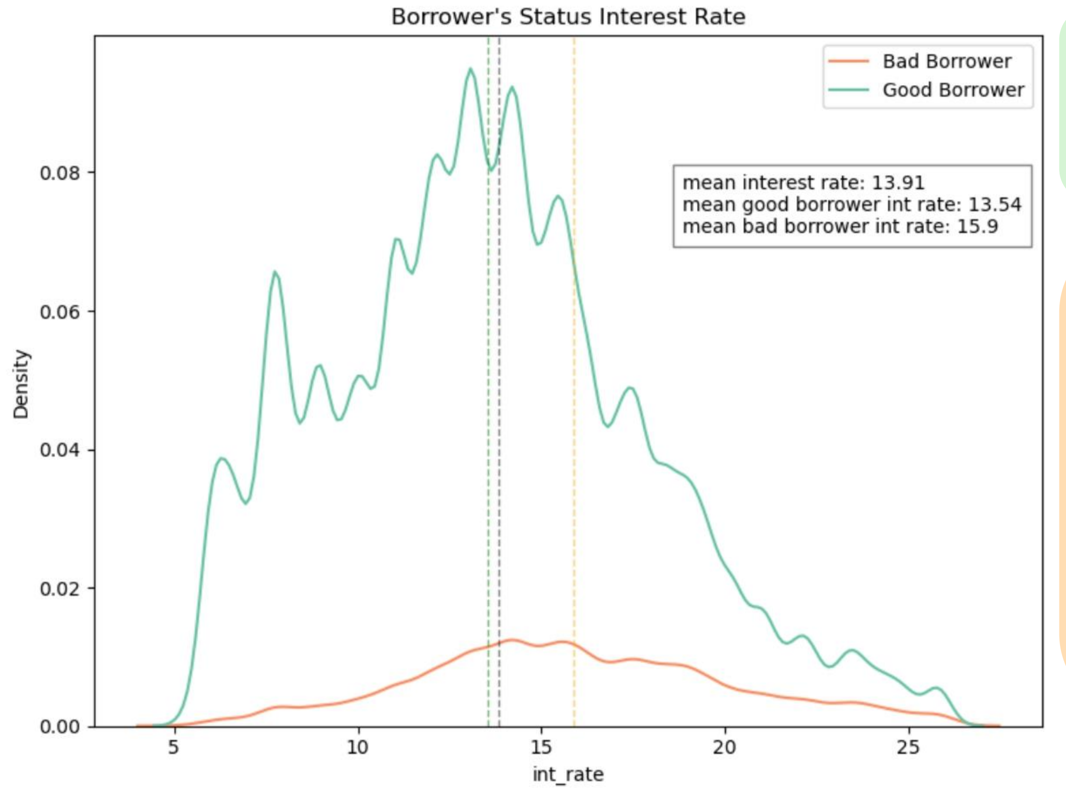
- *There is upward trend for good borrower by last payment year from 2008 to 2016. It means that the percentage of customers who have no difficulty in paying is getting higher. And this is good for the sustainability of the loan company's revenue.*
- *There is downward trend for bad borrower by last payment year from 2008 to 2016. It means that the company's performance is very good in selecting loan applications by borrowers.*

- *The difference percentage between good and bad borrowers for 2016 is really signicant with 98%. It means that the management could implement the stategy and policy in borrower eligibility and risk assessment.*

Exploratory Data Analysis

Rakamin Academy

Borrower's Status Interest Rate

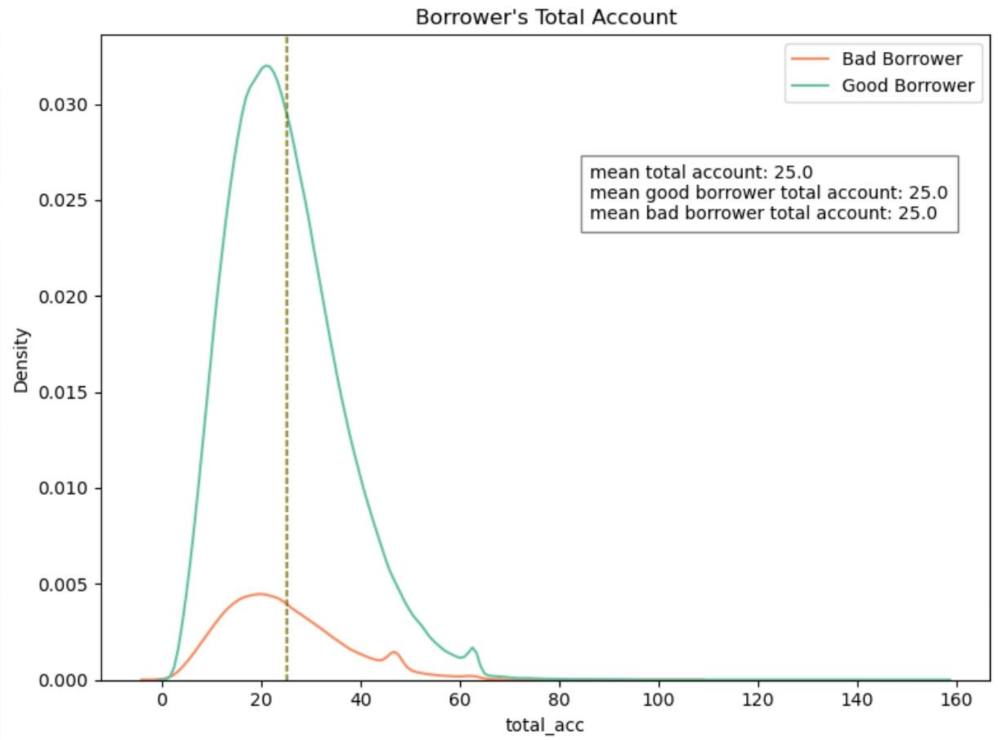


- *Average interest rate for all borrowers is 13.91%.*
- *Average interest rate for good borrowers is 13.54%.*
- *Average interest rate for bad borrowers is 15.9%.*

- *Based on creditninja.com(2023), For a good credit score, the average rate is 13% – 16%.*
- *The average loan interest rate for all borrowers at ID/X Partners is still relatively good because it is in the range of 13% – 16%.*
- *The reason why the average interest rate for bad borrowers is higher than for good borrowers is because the lower the borrower's credit score is, the higher the interest rates become to compensate for the increased risk the lender takes on.*

Exploratory Data Analysis

Borrower’s Total Account



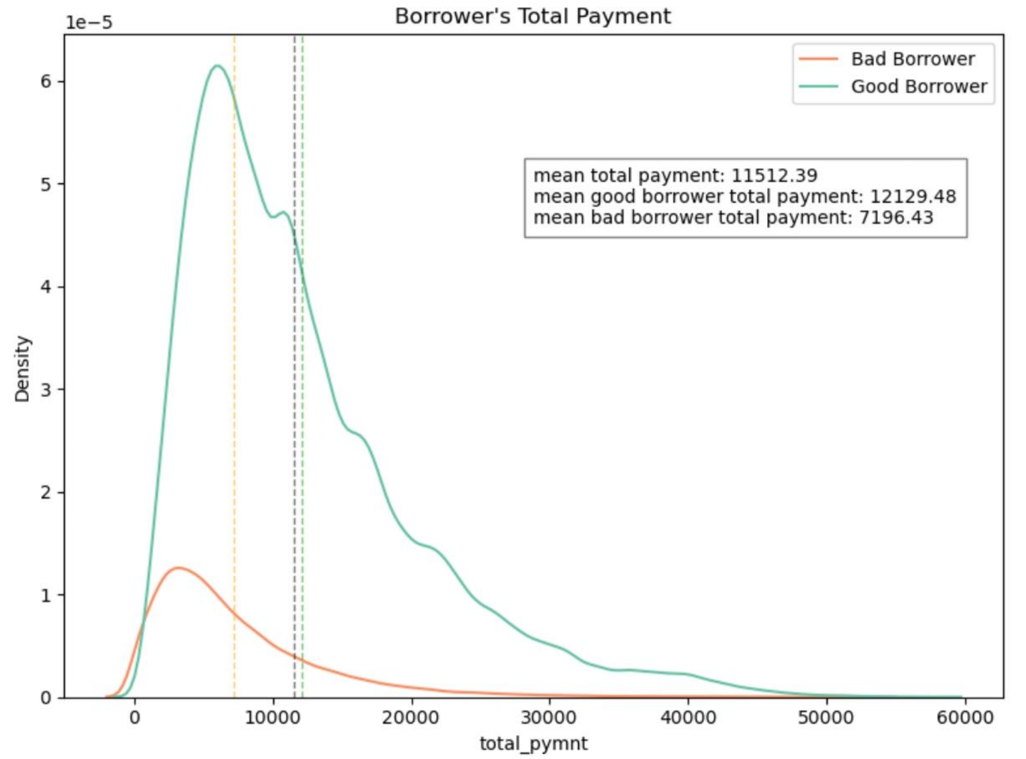
- *nerdwallet.com (2023) suggests that 5 or more accounts is a reasonable number to build toward over time.*

- *From the plot, average total account of good and bad borrower is 25 account, which is a lot more than recommended.*
- *Having too many open credit lines, even if borrower's not using them, can hurt their credit score by making the borrower's look more risky to lenders.*
- *Having multiple active accounts also makes it more challenging to control spending and keep track of payment due dates.*

Exploratory Data Analysis

Rakamin Academy

Borrower's Total Payment

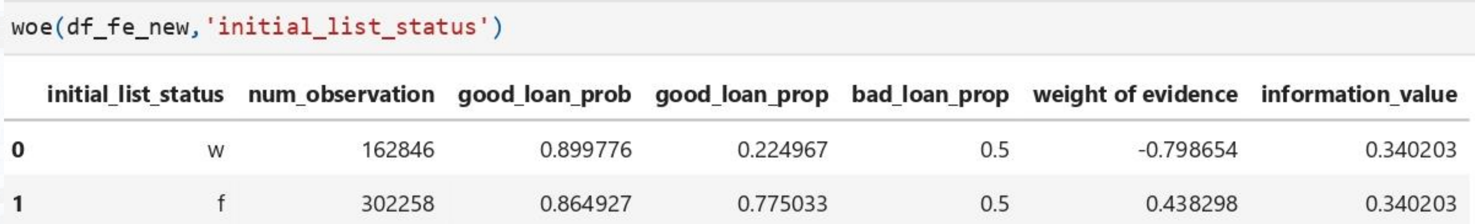


- *The average total amount funded for all customers is 11,512*
- *The average total amount funded for good borrowers is 12,129.*
- *The average total amount funded for bad borrowers is 7,196.*

- *Based on (investopedia.com, 2023), The average personal loan amount in America was 11,548 dollar in the second quarter of 2023 with average interest rate in Q2 is 11.48%.*
- *The average total amount funded in ID/X Partner is relevant with the average personal loan amount in America.*

Feature Engineering with Weight of Evidence (WOE) & Information Value (IV)

Weight of Evidence (WOE) & Information Value (IV)



```
woe(df_fe_new,'initial_list_status')
```

	initial_list_status	num_observation	good_loan_prob	good_loan_prop	bad_loan_prop	weight of evidence	information_value
0	w	162846	0.899776	0.224967	0.5	-0.798654	0.340203
1	f	302258	0.864927	0.775033	0.5	0.438298	0.340203

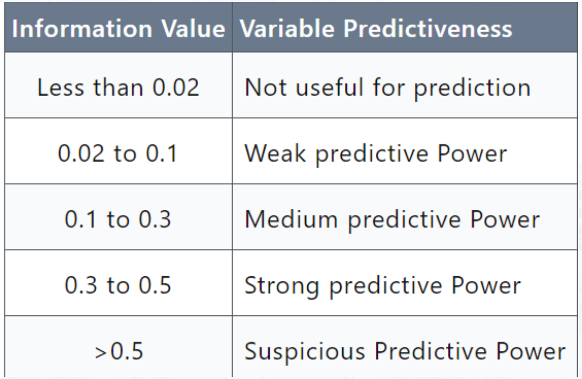
- *The most common logistic regression models a binary outcome (true/false)*
- *Weight of evidence (WOE) generally described as a measure of the separation of good and bad customers.*

$$\text{WOE} = \ln\left(\frac{\text{Distribution of Goods}}{\text{Distribution of Bads}}\right)$$

- *Positive WOE means Distribution of Goods > Distribution of Bads*
- *Negative WOE means Distribution of Goods < Distribution of Bads*

- *Information value (IV) is one of the most useful technique to select important variables in a predictive model. It helps to rank variables on the basis of their importance.*

$$\text{IV} = \sum (\text{\% of non-events} - \text{\% of events}) * \text{WOE}$$

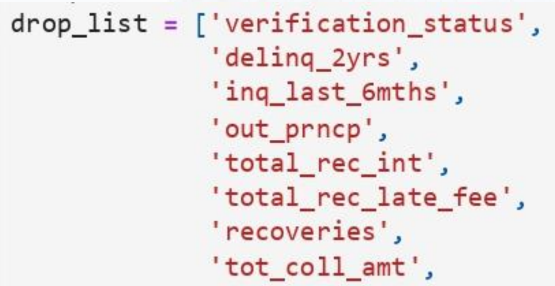


Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

Weight of Evidence (WOE) & Information Value (IV)

Drop Feature No Needed

- *Information Value (IV) < 0.02, The variable is Not useful for prediction*
- *Information Value (IV) > 0.5, The variable is Suspicious Predictive Power*



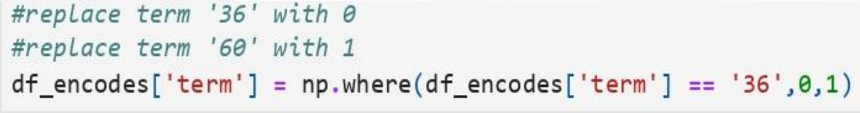
```
drop_list = ['verification_status',
             'delinq_2yrs',
             'inq_last_6mths',
             'out_prncp',
             'total_rec_int',
             'total_rec_late_fee',
             'recoveries',
             'tot_coll_amt',
```

- *Before Feature Engineering = 39 columns*
- *After Feature Engineering = 29 columns*

Feature Encoding

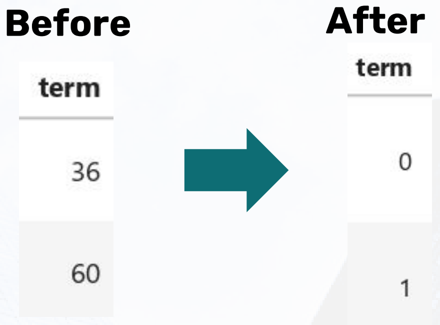
Categorical Encoding

Label Encoding



```
#replace term '36' with 0
#replace term '60' with 1
df_encodes['term'] = np.where(df_encodes['term'] == '36',0,1)
```

- *Machine learning models can only work with numerical values, so necessary to transform the categorical values to numerical, called **feature encoding**.*
- ***Label encoding** doesn't add any extra columns to the data but instead assigns a number to each unique value in a feature.*



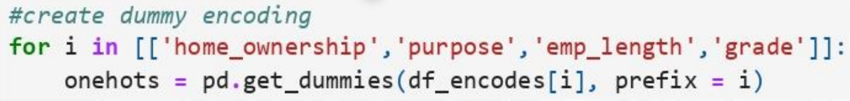
Before

term
36
60

After

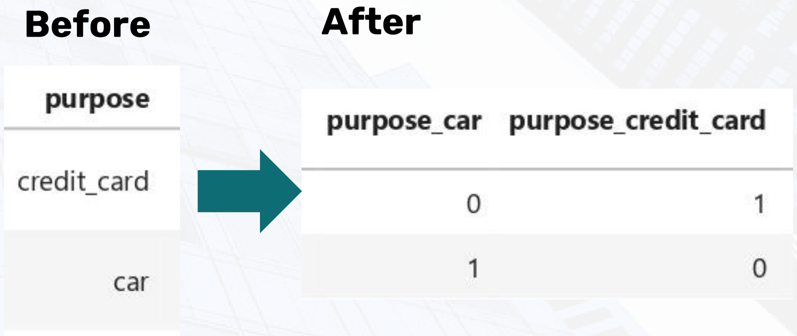
term
0
1

One Hot Encoding



```
#create dummy encoding
for i in [['home_ownership','purpose','emp_length','grade']]:
    onehots = pd.get_dummies(df_encodes[i], prefix = i)
```

- *One-Hot encoding technique is used when the features are nominal (do not have any order).*
- *In one hot encoding, for every categorical feature, a new variable is created (Dummy variables), either 0 or 1.*
- *0 represents the absence, and 1 represents the presence of that category*



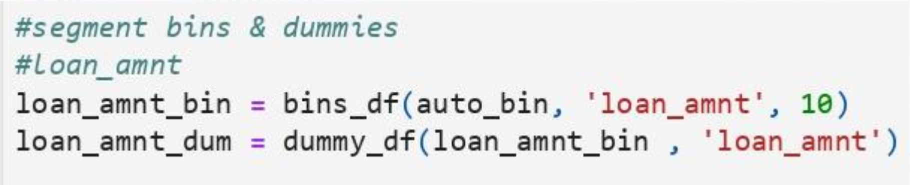
Before

purpose
credit_card
car

After

purpose_car	purpose_credit_card
0	1
1	0

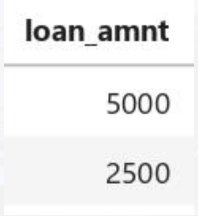
Numerical Encoding



```
#segment bins & dummies
#loan_amnt
loan_amnt_bin = bins_df(auto_bin, 'loan_amnt', 10)
loan_amnt_dum = dummy_df(loan_amnt_bin , 'loan_amnt')
```

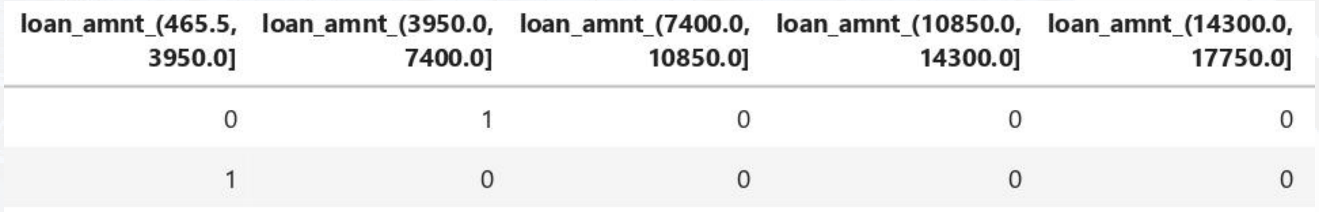
- *The simplest form of encoding numerical columns using Binarization*
- *In "binarization," continuous variables are transformed into binary values (0 or 1) based on a predetermined threshold*
- *Using this method, we can identify if a data point is above or below the threshold.*

Before



loan_amnt
5000
2500

After

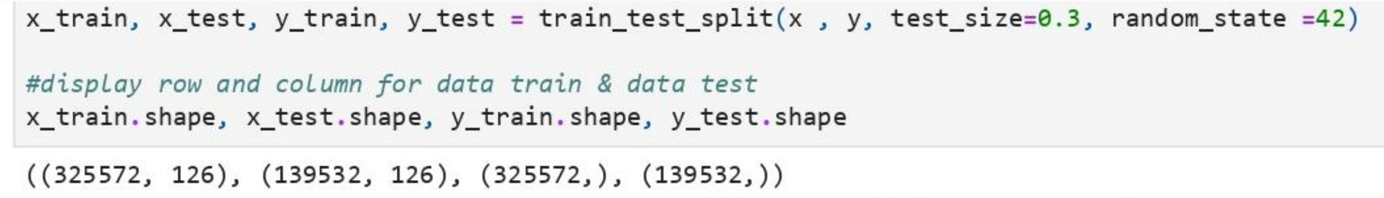


loan_amnt_(465.5, 3950.0]	loan_amnt_(3950.0, 7400.0]	loan_amnt_(7400.0, 10850.0]	loan_amnt_(10850.0, 14300.0]	loan_amnt_(14300.0, 17750.0]
0	1	0	0	0
1	0	0	0	0

Modelling

Split Dataset

70% training : 30% testing

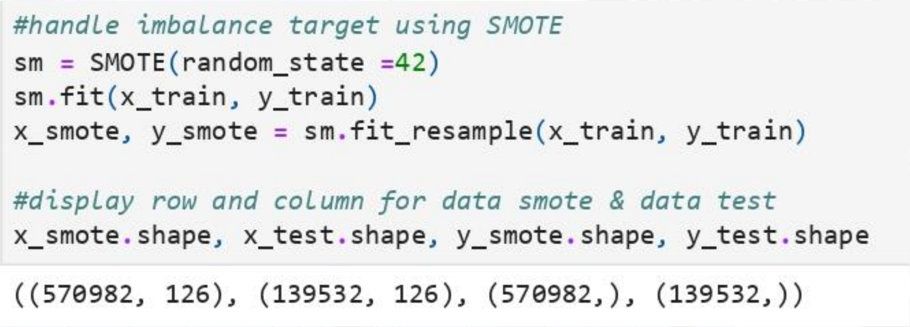


```
x_train, x_test, y_train, y_test = train_test_split(x , y, test_size=0.3, random_state =42)

#display row and column for data train & data test
x_train.shape, x_test.shape, y_train.shape, y_test.shape
```

```
((325572, 126), (139532, 126), (325572,), (139532,))
```

SMOTE

- ***Imbalanced Data*** *Distribution, generally happens when observations in one of the class are much higher or lower than the other classes*
- ***SMOTE*** *(synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem by randomly increasing minority class examples by replicating them*

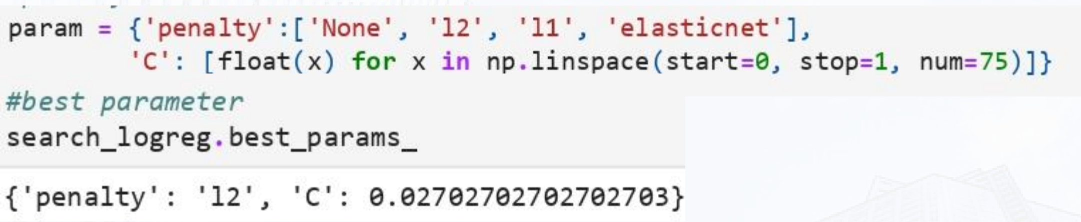


```
#handle imbalance target using SMOTE
sm = SMOTE(random_state =42)
sm.fit(x_train, y_train)
x_smote, y_smote = sm.fit_resample(x_train, y_train)

#display row and column for data smote & data test
x_smote.shape, x_test.shape, y_smote.shape, y_test.shape
```

```
((570982, 126), (139532, 126), (570982,), (139532,))
```

Train Model

Hyperparameter



```
param = {'penalty':['None', 'l2', 'l1', 'elasticnet'],
         'C': [float(x) for x in np.linspace(start=0, stop=1, num=75)]}
#best parameter
search_logreg.best_params_
```

```
{'penalty': 'l2', 'C': 0.02702702702702703}
```

- *If all training points are correctly classified then we have problem of **overfitting** (means doing perfect job on training set but performing very badly on test set, i.e. errors on train data is almost zero but errors on test data are very high).*

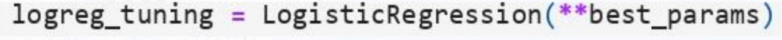
- *The regularization term adds a **penalty** to the model's coefficients, which reduces their magnitude and helps to prevent overfitting*

- *The “**C**” hyperparameter controls the strength of the regularization.*
- *A smaller value for “C” (e.g. C=0.01) leads to stronger regularization and a simpler model,*
- *while a larger value (e.g. C=1.0) leads to weaker regularization and a more complex model.*

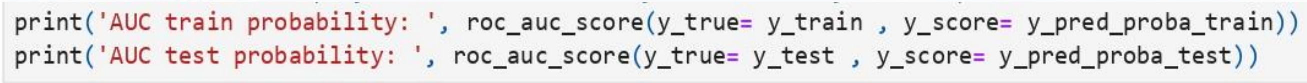
- *Best parameter we've got **is L2 (Ridge**) regularization with 'C' is 0.027 which is near to 0, and leads to **stronger regularization and a simpler model***

Train Model **Binary Logistic Regression**

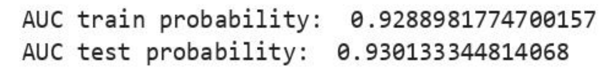
ROC/AUC



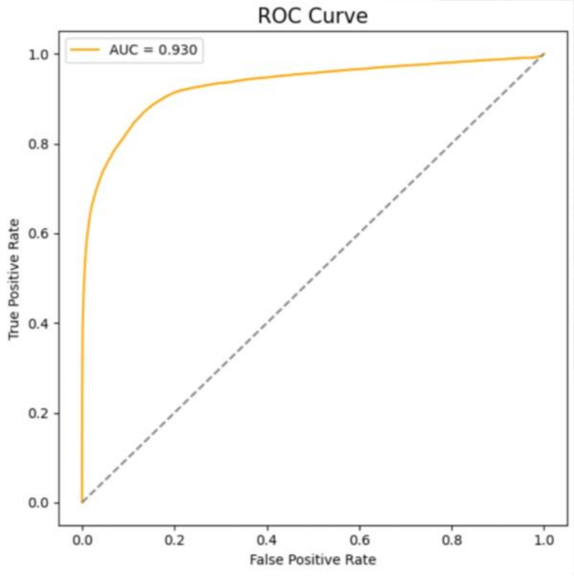
```
logreg_tuning = LogisticRegression(**best_params)
```



```
print('AUC train probability: ', roc_auc_score(y_true= y_train , y_score= y_pred_proba_train))
print('AUC test probability: ', roc_auc_score(y_true= y_test , y_score= y_pred_proba_test))
```



```
AUC train probability:  0.9288981774700157
AUC test probability:  0.930133344814068
```

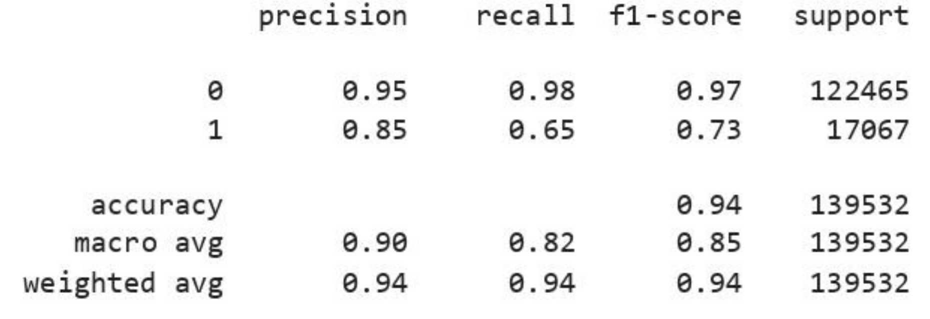


- ***Logistic regression*** *is the appropriate regression analysis to conduct when the dependent variable is binary, means variable has only 2 outputs.*
- ***ROC*** *is a probability curve and* ***AUC*** *represents the degree or measure of separability*
- *An* ***excellent model*** *has AUC near to the 1 which means it has a good measure of separability.*
- *A* ***poor model*** *has an AUC near 0 which means it has the worst measure of separability*

- *AUC score = 0.93, which is near to 1, indicates good performance*

Train Model

Classification Report



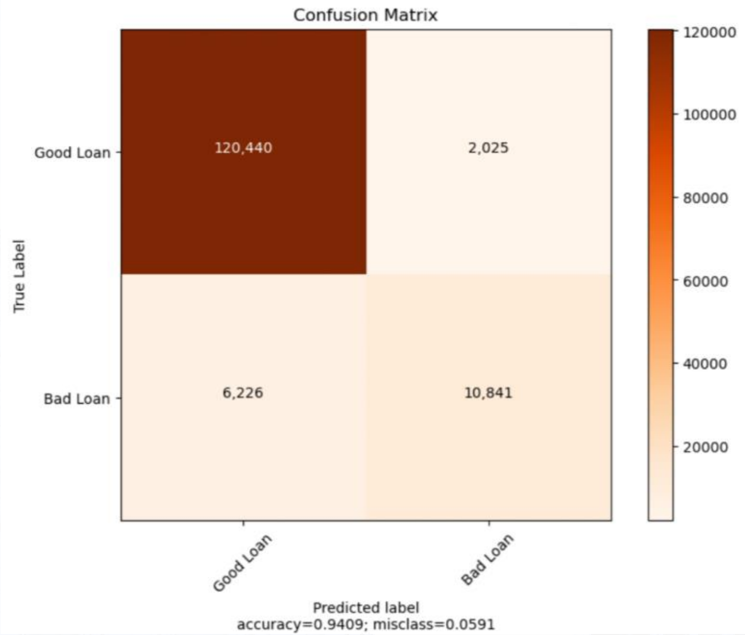
```
              precision    recall  f1-score   support

           0       0.95      0.98      0.97    122465
           1       0.85      0.65      0.73     17067

    accuracy                           0.94    139532
   macro avg       0.90      0.82      0.85    139532
weighted avg       0.94      0.94      0.94    139532
```

- ***Precision*** *tells us the accuracy of positive predictions.*
- *Out of all the loan status that the model predicted would get good loan, only 85% actually did.*

- ***Recall*** *tells us the fraction of correctly identified positive predictions*
- *Out of all the loan status that actually did get good loan, the model only predicted this outcome correctly for 65% of those loan status*

- *F1 score >= 0.7 considered good (spotintelligence.com, 2023)*
- ***F1 Score*** *= 0.73. So the model does a good job of predicting whether the loan status is considered good or bad*

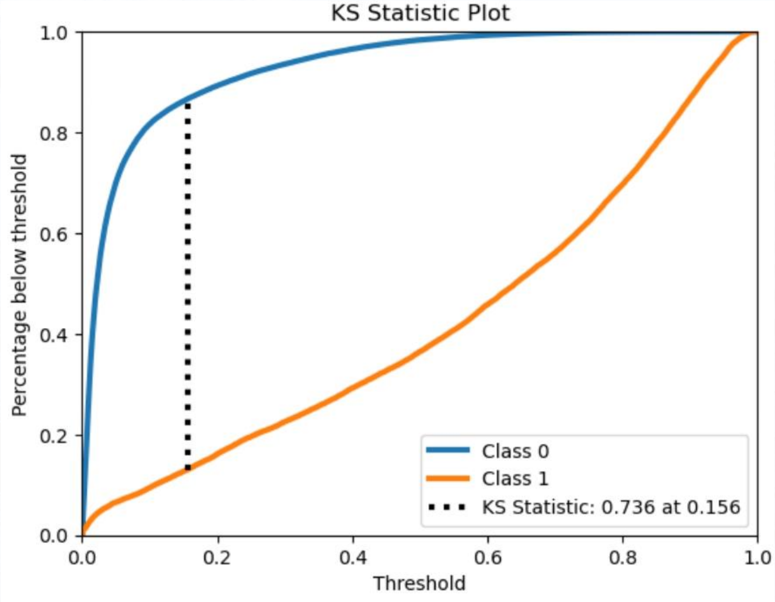
Confusion Matrix



- ***Confusion Matrix*** *is a performance measurement for machine learning classification problem*
- *Correct classifications are the diagonal elements of the matrix 120,440 for the positive class and 10,841 for the negative class*
- *Accuracy rate, which is the percentage of times a classifier is correct = 94.09%*

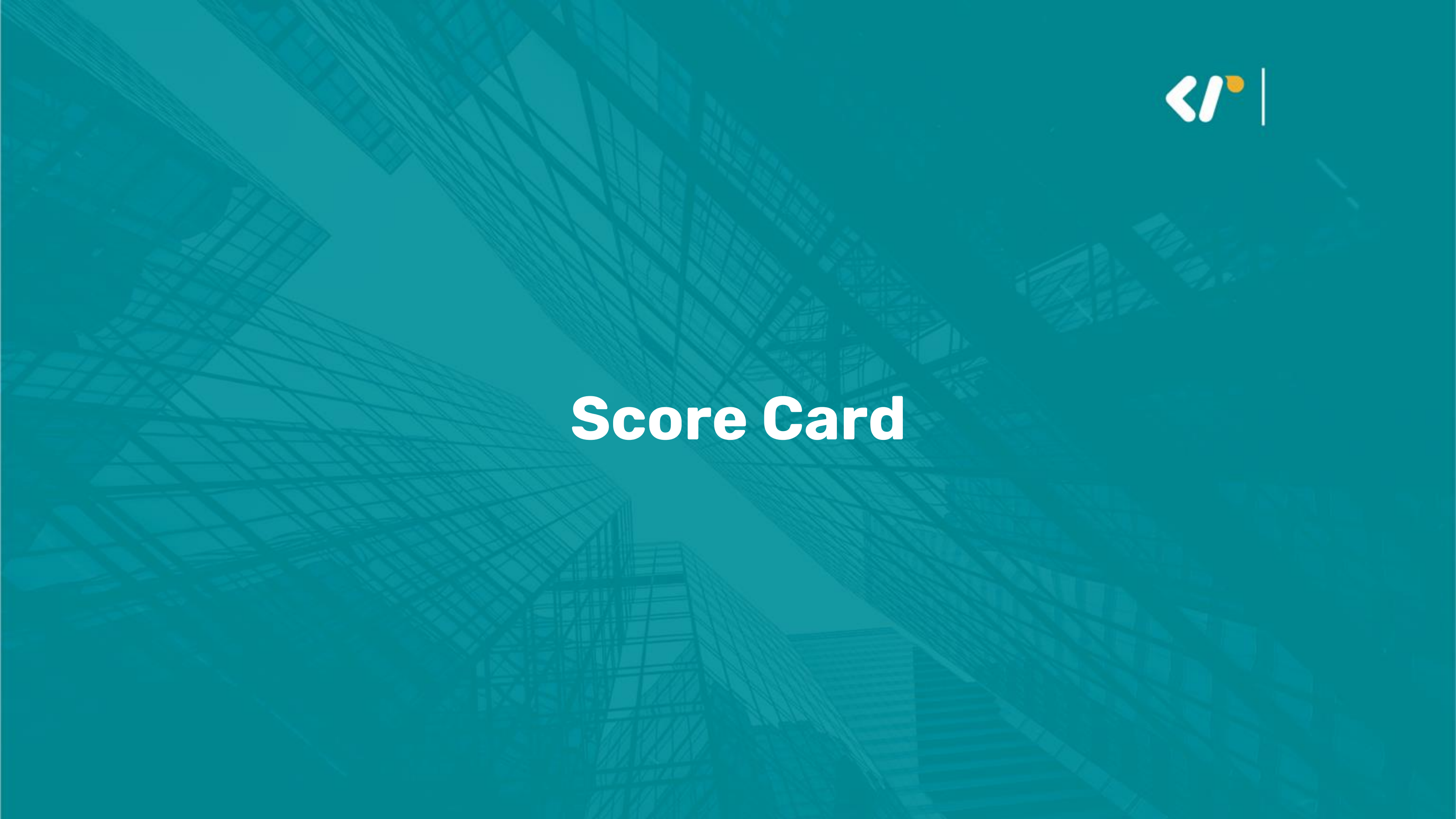
Train Model

Kolmogorov-Smirnov



- *K-S should be a high value (Max =1.0) when the fit is good and a low value (Min = 0.0) when the fit is not good*

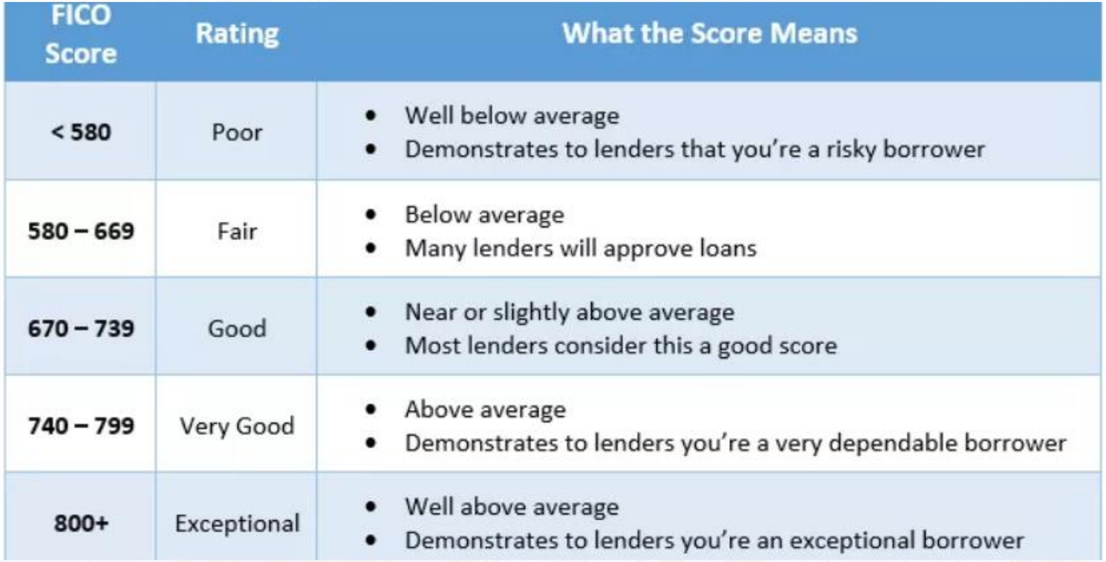
- *KS Statistic = 0.736*
- *Considered it as '**medium**' dataset, which mean even though it doesn't have perfect separation, but there is enough overlap to confuse the classifier, and has wide gap between the class CDF (positive & negative instances).*



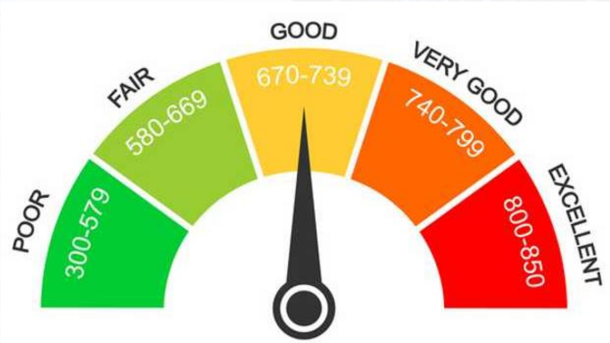
Score Card

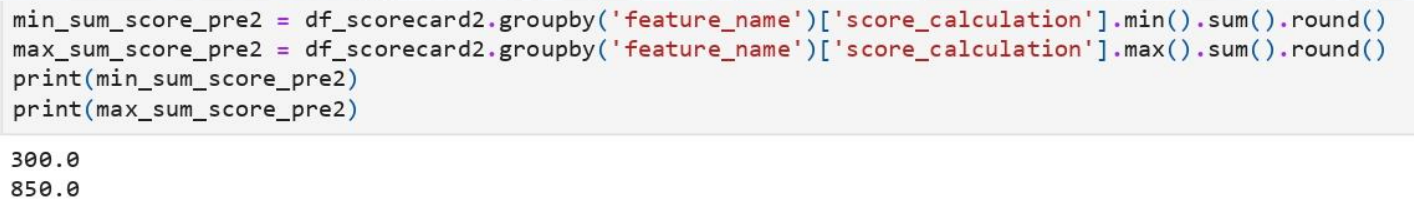
FICO Score

- ***FICO score** is a credit score created by the Fair Isaac Corporation (FICO)*
- *Lenders use borrowers' FICO scores along with other details on borrowers' credit reports to assess credit risk and determine whether to extend credit.*



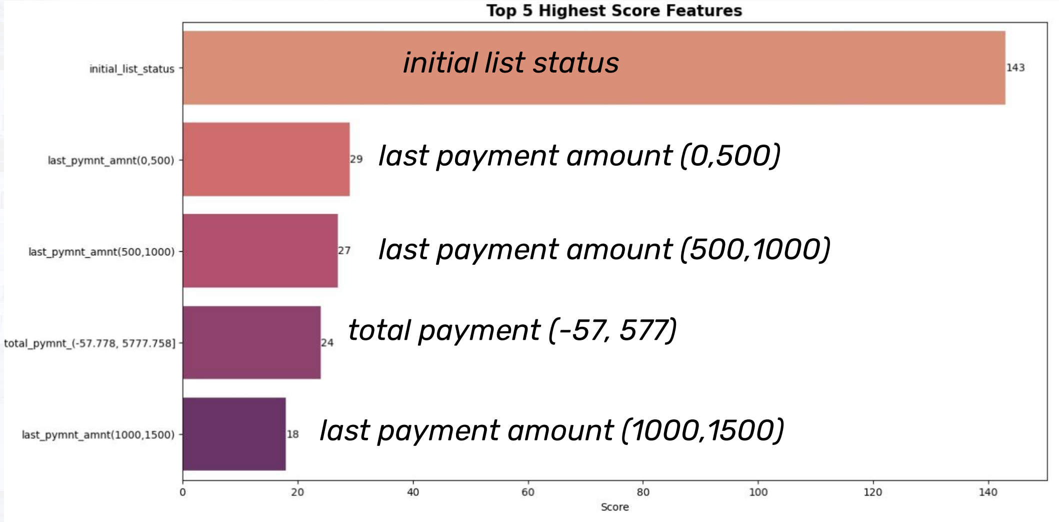
FICO Score	Rating	What the Score Means
< 580	Poor	• Well below average • Demonstrates to lenders that you're a risky borrower
580 – 669	Fair	• Below average • Many lenders will approve loans
670 – 739	Good	• Near or slightly above average • Most lenders consider this a good score
740 – 799	Very Good	• Above average • Demonstrates to lenders you're a very dependable borrower
800+	Exceptional	• Well above average • Demonstrates to lenders you're an exceptional borrower

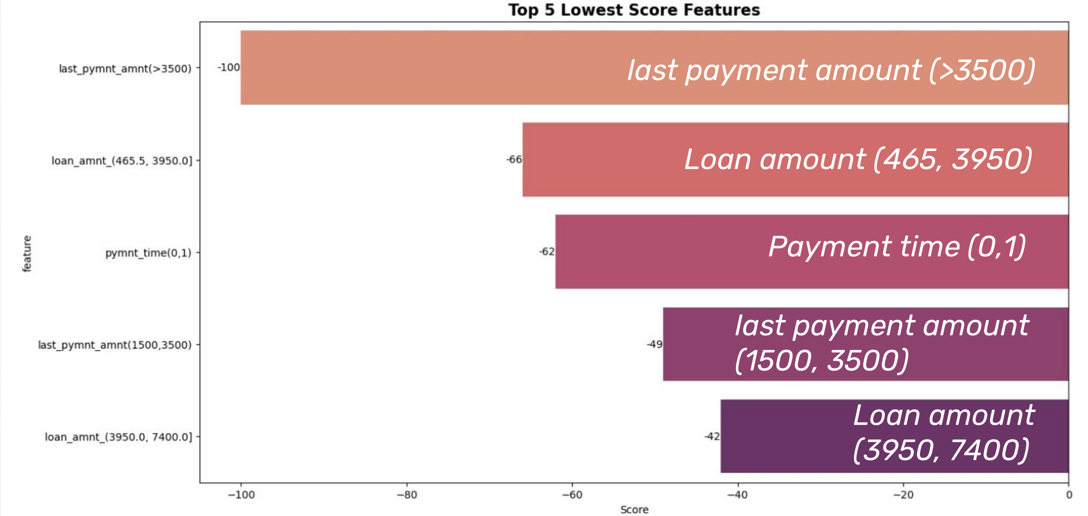




```
min_sum_score_pre2 = df_scorecard2.groupby('feature_name')['score_calculation'].min().sum().round()
max_sum_score_pre2 = df_scorecard2.groupby('feature_name')['score_calculation'].max().sum().round()
print(min_sum_score_pre2)
print(max_sum_score_pre2)
```

```
300.0
850.0
```





Top 5 Highest & Lowest Score Features

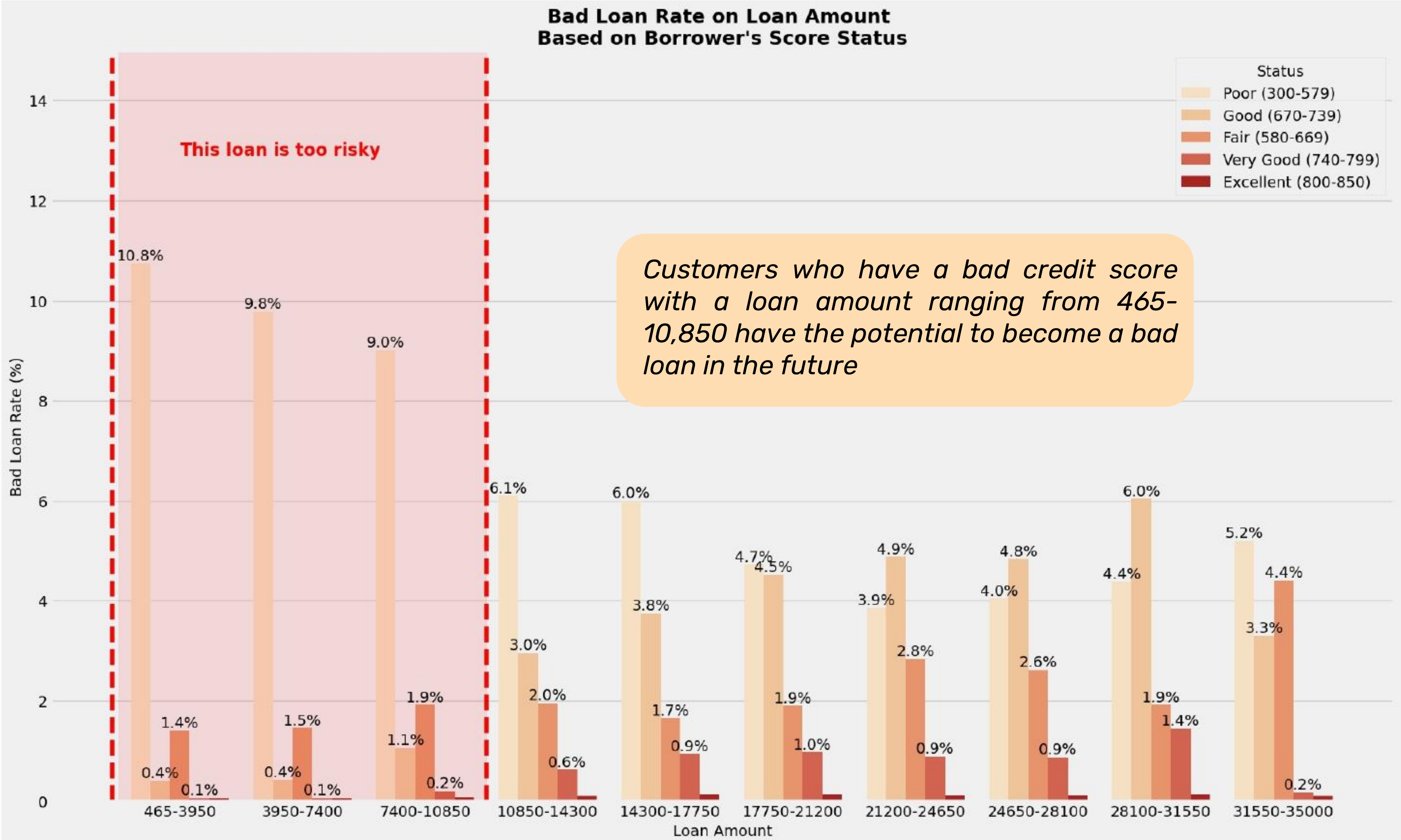
Rakamin Academy

Features that make contribution to increase or decrease credit score are:

- *initial list status*
- *last payment amount*
- *total payment*
- *loan amount*
- *payment time*

The result is in line with how FICO credit score is calculated that based on five factors:

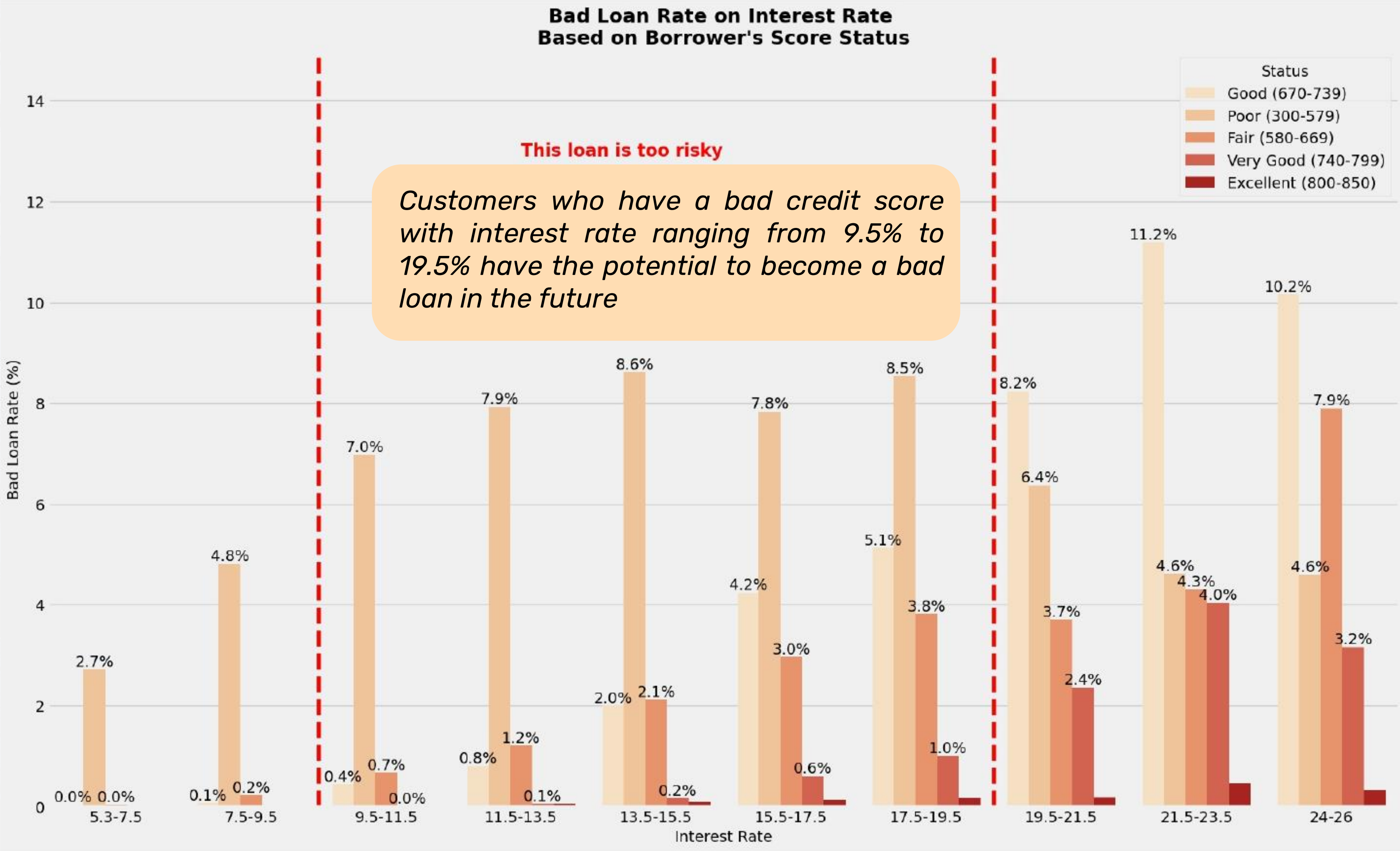
- *payment history,*
- *amount owed,*
- *new credit,*
- *length of credit history*
- *credit mix.*



Bad Loan Rate on Loan Amount
Based on Borrower's Score Status

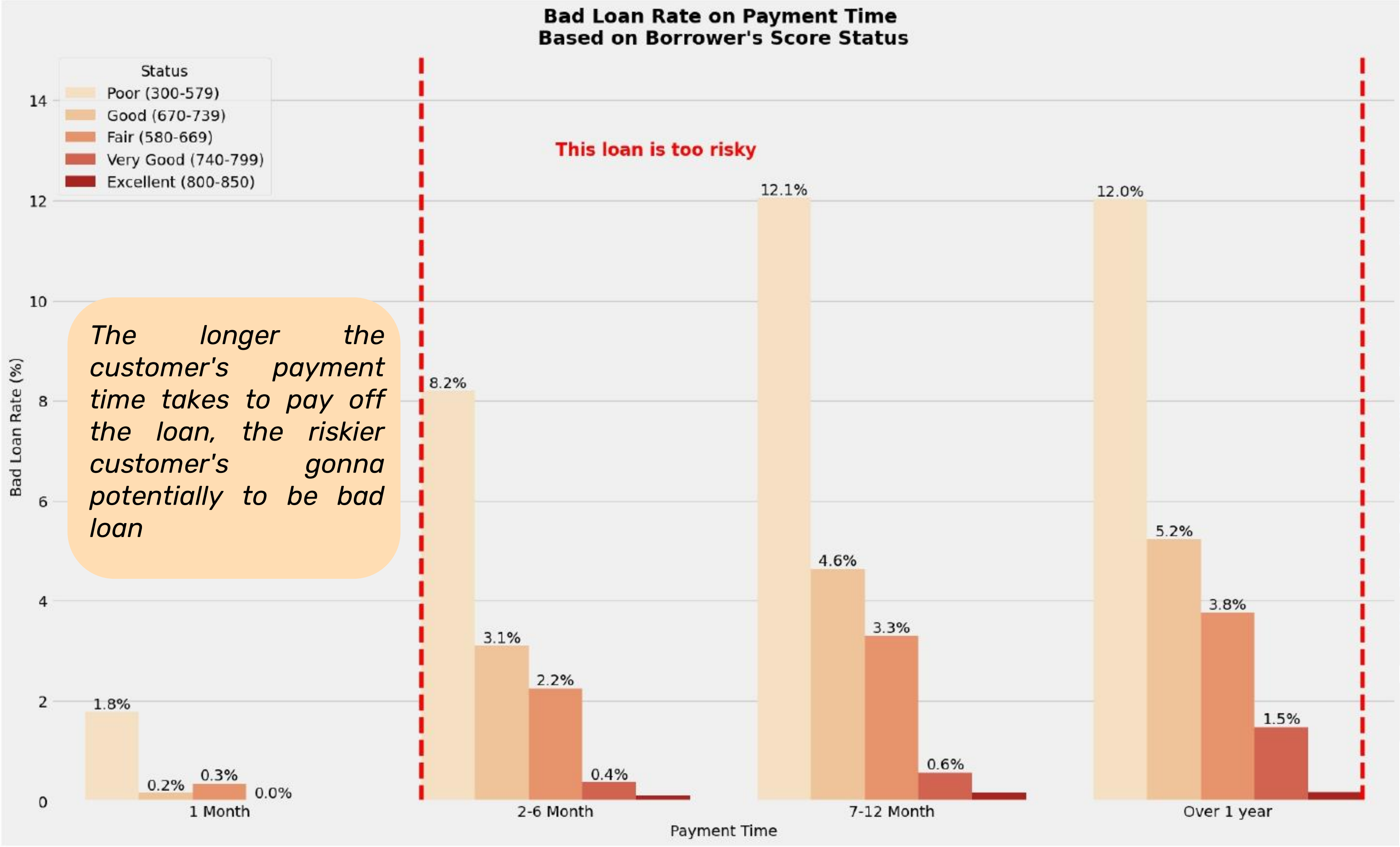
This loan is too risky

Customers who have a bad credit score with a loan amount ranging from 465-10,850 have the potential to become a bad loan in the future



This loan is too risky

Customers who have a bad credit score with interest rate ranging from 9.5% to 19.5% have the potential to become a bad loan in the future



The longer the customer's payment time takes to pay off the loan, the riskier customer's gonna potentially to be bad loan

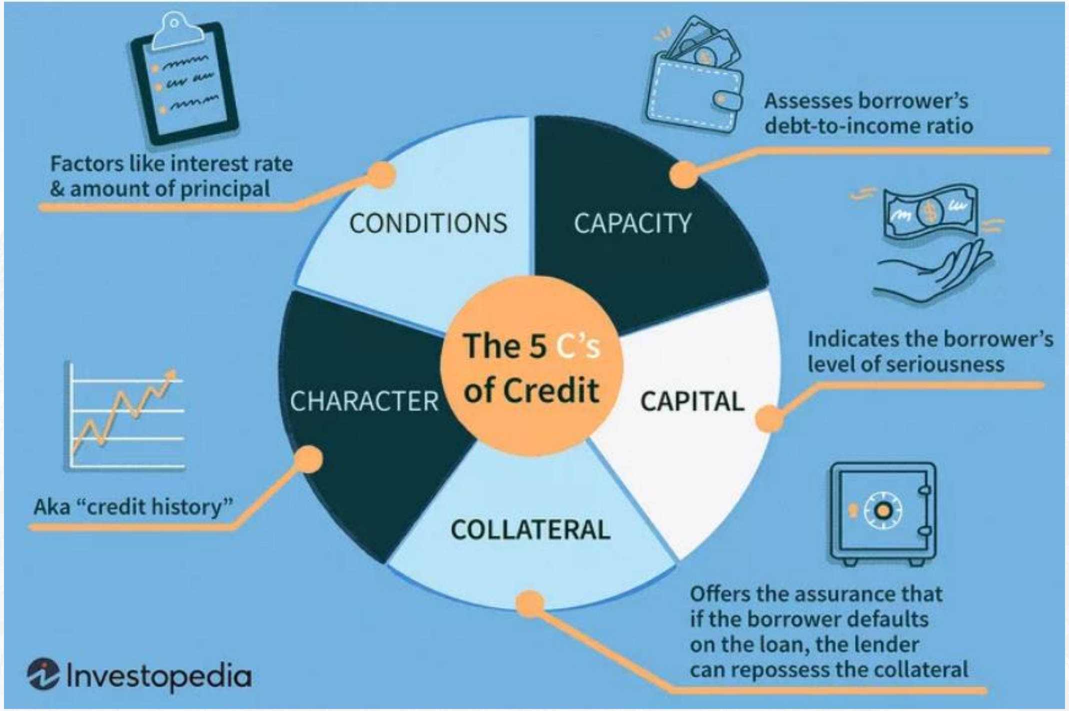
Recommendation

Recommendation

- *Loan companies can build a robust and effective* ***credit scoring model machine learning*** *using variety of methods and criteria to assess the creditworthiness of potential customers.*

- *The* ***goa****l is to minimize the risk of lending to individuals who are unlikely to repay their loans.*

- *One of method to evaluate a borrower incorporates both qualitative and quantitative measures is the* ***5 C's*** *of credit (Character, Capacity, Capital, Collateral, and Conditions)*



Factors like interest rate & amount of principal
Assesses borrower's debt-to-income ratio
CONDITIONS
CAPACITY
The 5 C's of Credit
Indicates the borrower's level of seriousness
CHARACTER
CAPITAL
Aka "credit history"
COLLATERAL
Offers the assurance that if the borrower defaults on the loan, the lender can repossess the collateral
Investopedia

Source : (Investopedia.com, 2023)

Link Portolio On Github :

https://github.com/nickenshidqia/Credit_Risk_Assessment_IDX_Partners

Thank You

Rakamin Academy X id/x partners