Credit Risk Assessment

Rafif Noor Audryan

About Me

Fresh graduate from Universitas Sebelas Maret majoring in Economic Development. Has interest in data related role such as Data Analytic, Data Science, and Machine Learning. Proficient in using Spreadsheet, SQL, and Python. Skilled in Exploratory Analysis, Data Cleansing, Data Visualization, Machine Learning and Evaluation.

HERE IS MY LINKEDIN

Background

A lending company has a problem where it requires efficiency and speed in receiving loans from each customer.

As a Data Science Intern from ID/X Partners, we will to process data and create models that are able to predict and assess optimal credit applications and predict existing risks.

To facilitate the assessment, we will create a credit score based on the logistic regression model. Finally, we will provide solutions for lending companies how the insights we get.

PROBLEM

- It Takes A Long Time If We Do The Assessment Manually
- There Is No Definite Standard In Determining Credit Score
- More Customer Data We Need To Assess Next

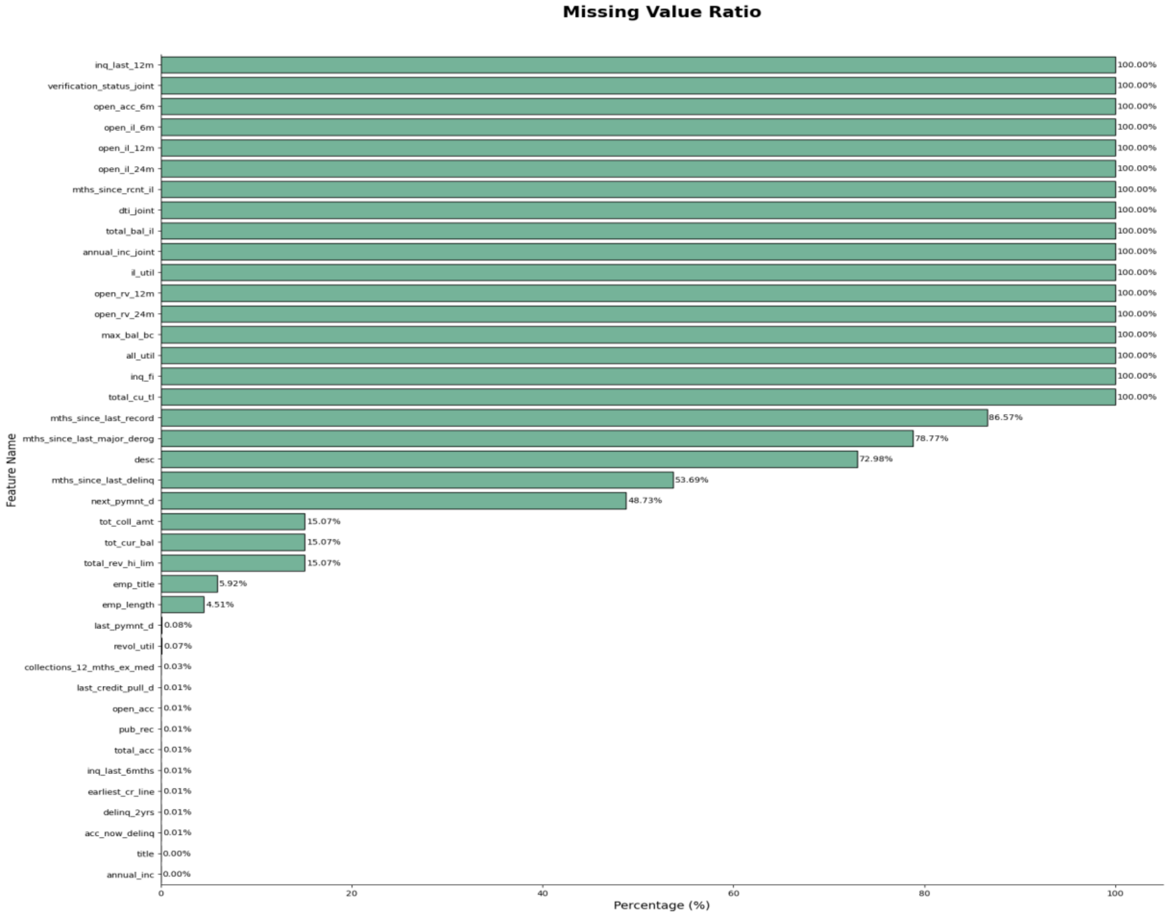
Dataset Overview

```
<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
```

- Dataset have 75 columns dan 466K rows
- There are 17 features whose data contains null data
- Some features have null data
- Loan Status to be set as target for the model has 9 unique values. To make prediction, will be formed with into 2 categorized ‘Good Loan’ with value ‘1’ and ‘Bad Loan’ with value ‘0’

01

Data Preprocessing



Handling Missing Value

- Based on this data , there are 17 feature have 100% missing value so it will drop
- Feature that have missing value more than 50% will drop because too avoid bias result on modeling
- Feature`tot_coll_amt`,`tot_cur_bal`,`total_rev_hi_lim` will replace missing value with "0" because assumption that customer didn't borrow loan again.
- Numerical feature will replace missing value with "Median“
- Categorical feature will replace missing value with "Mode"

For More Detail Check Here

Data Cleansing

01

Handle Unnecesarry Feature

Feature that contain free text, id, zip code will drop

02

Handle Feature that contain only one unique value

Feature ('pymnt_plan') which all have a value of one value it will drop it

03

Handle Features That Have A High Correlation Between Independent Features And Target Features

There are 7 features that have a high correlation (>0.8), these features will drop.

For More Detail Check Here

02

Feature Engineering

Feature Engineering

01

CHANGE DATA TYPE SOME FEATURE TO DATETIME AND ADD NEW FEATURE

4 feature will change data type to datetime, after that extract to create new feature:

1. pymnt_time : number of months between 'next_pymnt_d and 'last_pymnt_d'
2. credit_pull_year : number of years between 'last_credit_pull_d and 'earliest_cr_line',

02

FEATURE SELECTION USING WEIGHT OF EVIDENCE AND INFORMATION VALUE

There are 14 features that cannot be included in the model because feature have information value < 0.02 (useless predictive), feature have Information value > 0.5 (suspicious predictive), and feature that not make sense to bin.

03

ENCODE ALL FEATURES FOR THE MODEL WITH LABEL ENCODING AND ONE HOT ENCODING

There are 18 features that we will encode. Logistic regression have advantage to make best result if the data only contain binary value with 1 or 0 so numerical feature we will do various bin to create one hot encoding each feature bin

For More Detail Check [Here]()

03

Modelling

MODELLING

01 DEFINE FEATURE INDEPENDENT (X) AND TARGET (Y)

```
X = df_model.drop(['loan_status'], axis=1)
y = df_model['loan_status']
```

02 SPLIT DATA WITH RATIO 70% TRAIN : 30% TEST

```
#Split Dataset 70% Train : 30% Test
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3,random_state=24)
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

```
((326399, 126), (139886, 126), (326399,), (139886,))
```

03 HANDLING IMBALANCE TARGET USING SMOTE

```
# Handle Imbalance Target Using SMOTE
sm = SMOTE(random_state=24)
sm.fit(X_train, y_train)
X_smote, y_smote = sm.fit_resample(X_train, y_train)
X_smote.shape, X_train.shape, y_smote.shape, y_train.shape
```

```
((577036, 126), (326399, 126), (577036,), (326399,))
```

For More Detail Check Here

EVALUATION SCORE

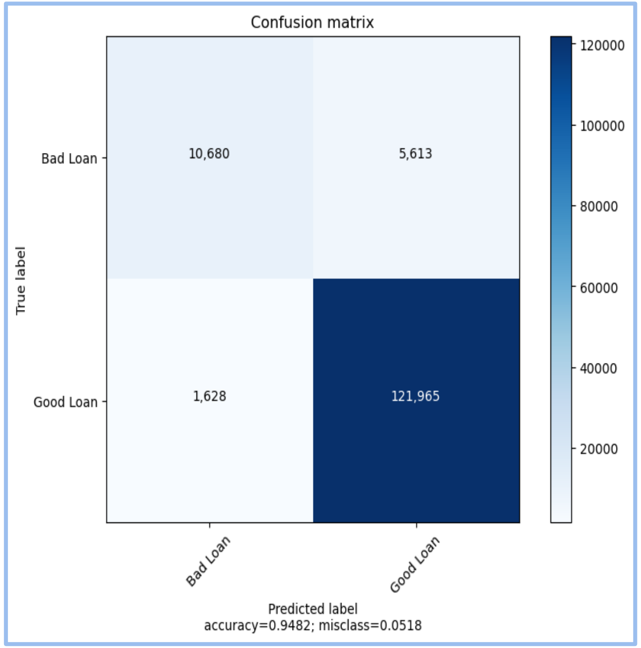
ALGHORITHM	AUC SCORE	ACCURACY SCORE
Logistic Regression	93.49%	94.90%
Logistic Regression with Hyperparameter Tuning	93.53%	94.83%

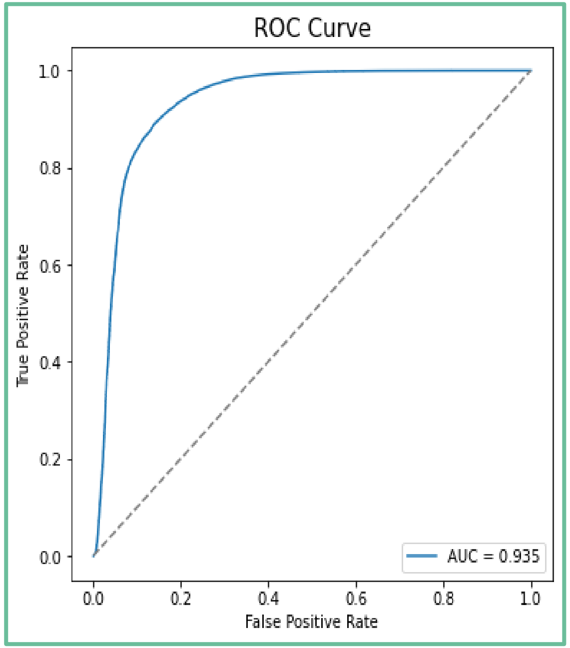
Metrics evaluation that important for this model is AUC SCORE.

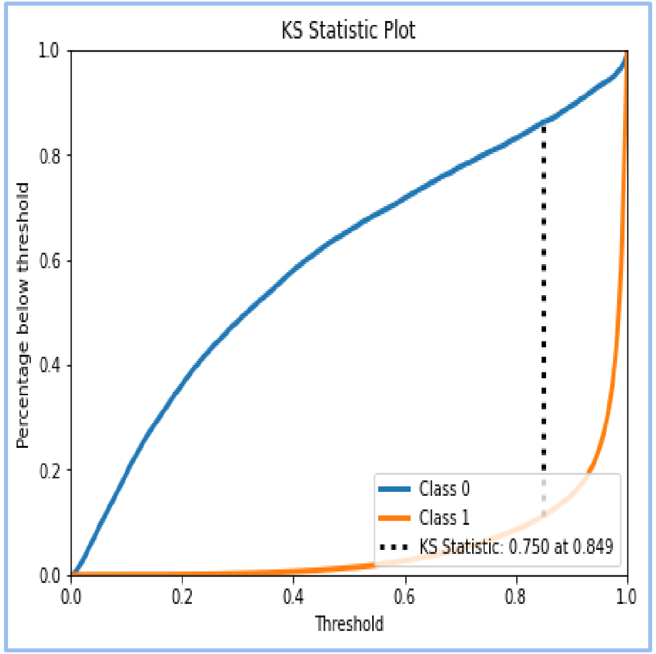
Logistic Regression with hyperparameter tuning get better result by 93.53% compared to non-tuning.

We decide to use Logistic Regression with hyperparameter tuning algorithm to get best prediction.

MODEL EVALUATION







04

CREDIT SCORE

SCORECARD

01 GET RESULT FROM COEFFICIENT OF LOGISTIC REGRESSION

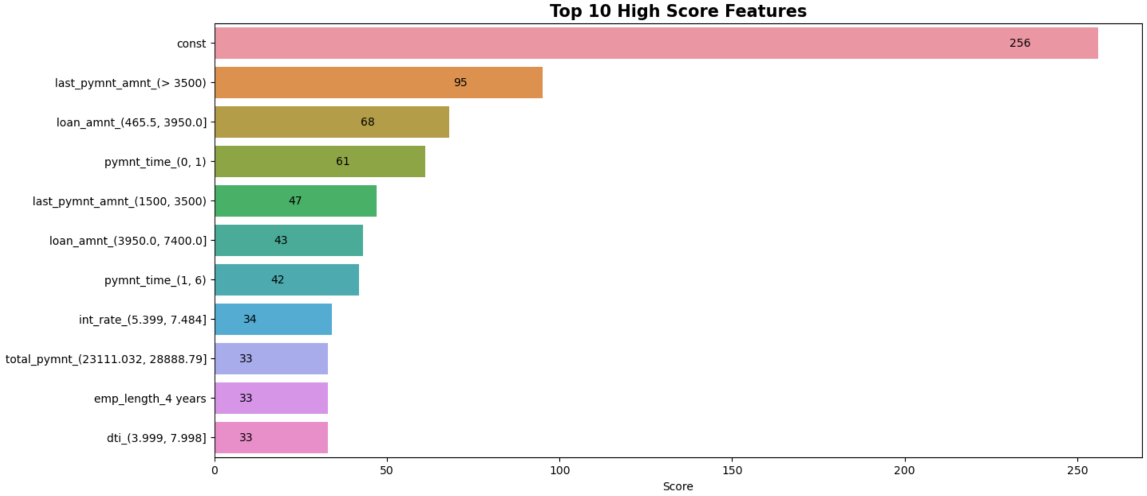
```
import statsmodels.api as sm
X2 = sm.add_constant(X_smote)
est = sm.Logit(y_smote, X2)
est2 = est.fit(method='bfgs')
print(est2.summary())
```

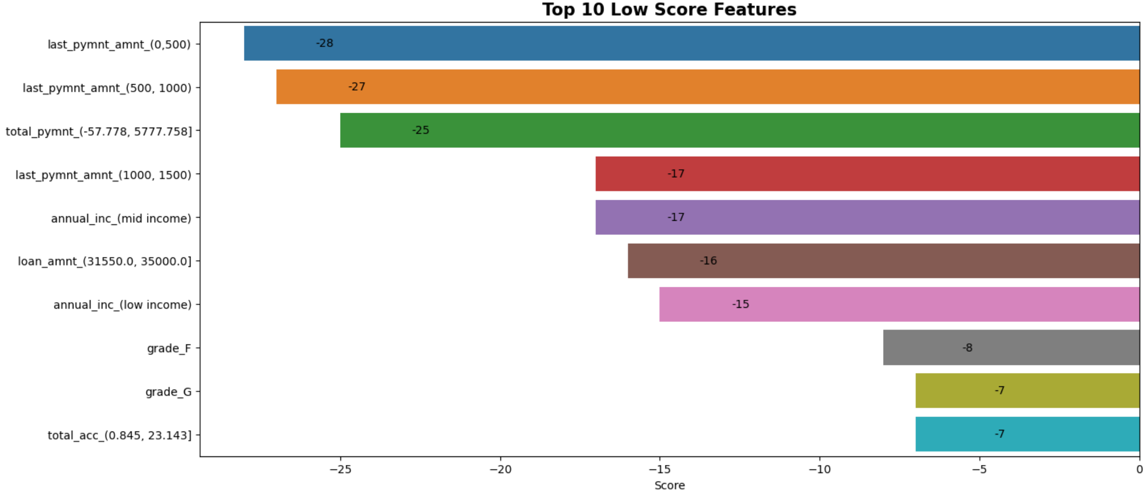
02 DEFINE MIN AND MAX SCORE BASED ON FICO SCORE (300-850)

```
# copy dataset
df_scorecard = df_importance.copy()

# define max and min score
min_score = 300
max_score = 850
```

```
# aggregate min and sum
min_sum_coef = df_scorecard.groupby('feature_name')['coef'].min().sum()

# aggregate max and sum
max_sum_coef = df_scorecard.groupby('feature_name')['coef'].max().sum()

# define credit score
df_scorecard['Score_Calculation'] = df_scorecard['coef'] * (max_score - min_score) / (max_sum_coef - min_sum_coef)

# adjust intercept values
df_scorecard['Score_Calculation'][0] = ((df_scorecard['coef'][0] - min_sum_coef) / ((max_sum_coef - min_sum_coef)) * (max_score - min_score) + min_score)

# round credit score
df_scorecard['Score_Final'] = df_scorecard['Score_Calculation'].round()
```

For More Detail Check [Here]





FEATURE IMPORTANCE

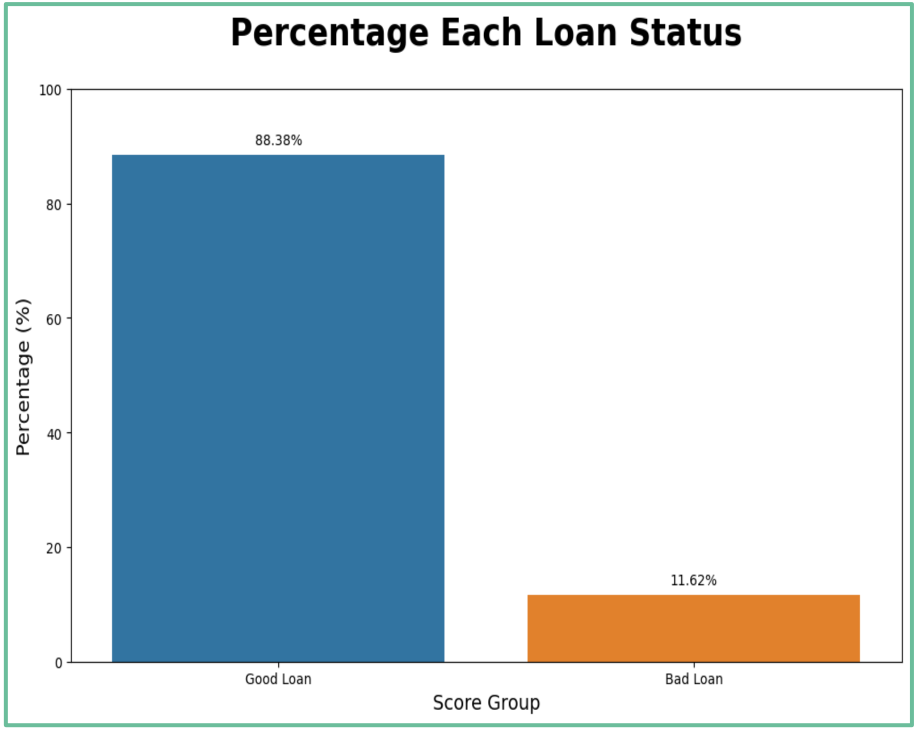
- As seen in the chart below, there are 10 features that have a highest scorecard to increase credit score.
- Meanwhile, there are 10 features that have the lowest scorecard that can reduce the credit score.
- For new customers, a base credit score is 256 that has been set based on the model we have created.

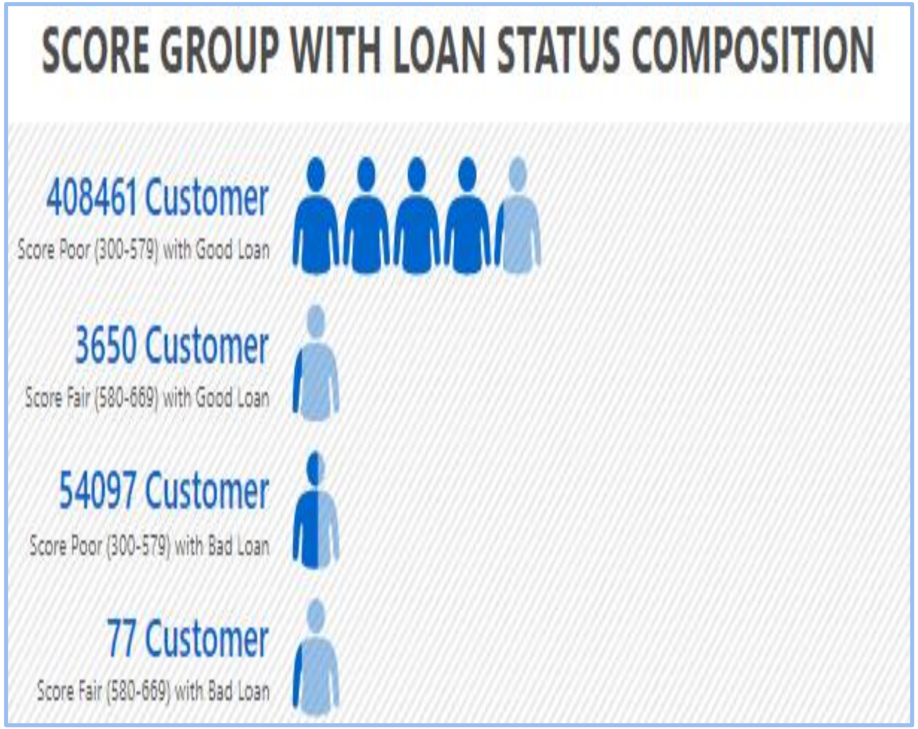
SAMPLE CREDIT SCORE

ID	Member _ID	Credit Score
15481037	17553386	526
313354	313351	418
26769586	29262614	347
4306237	5488553	466
36341665	39073098	436

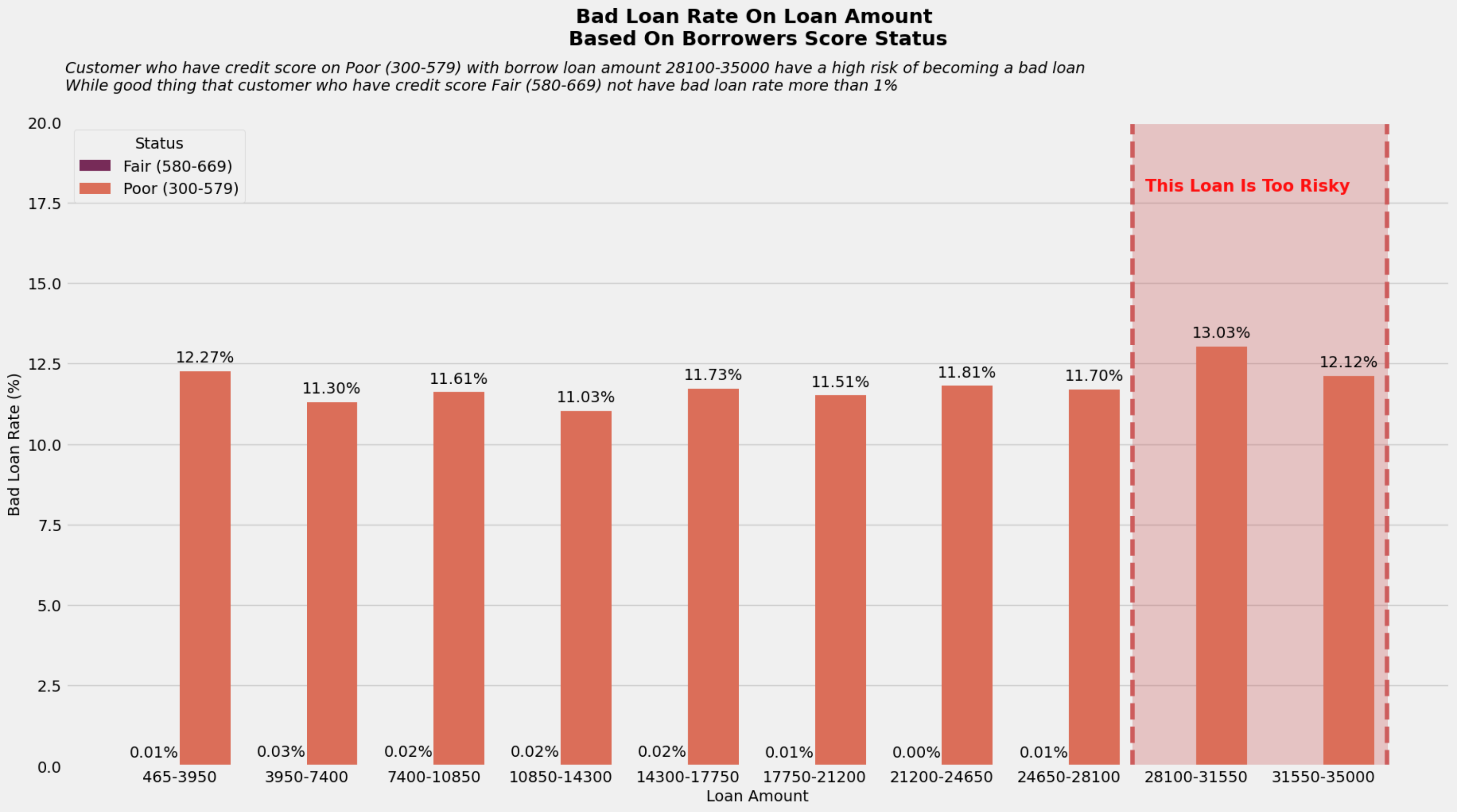
For More Detail Credit Score Result Here

BUSINESS INSIGHT

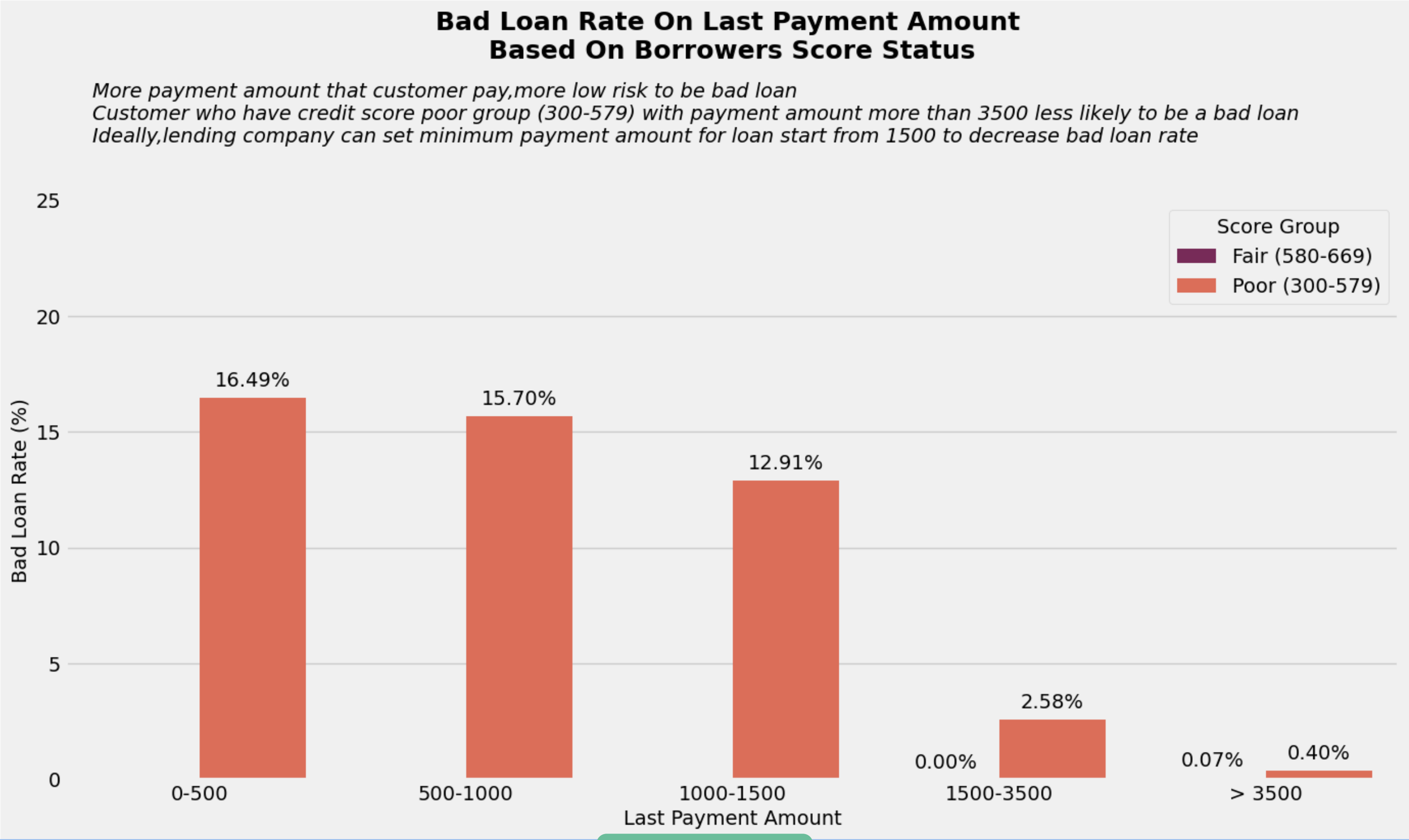




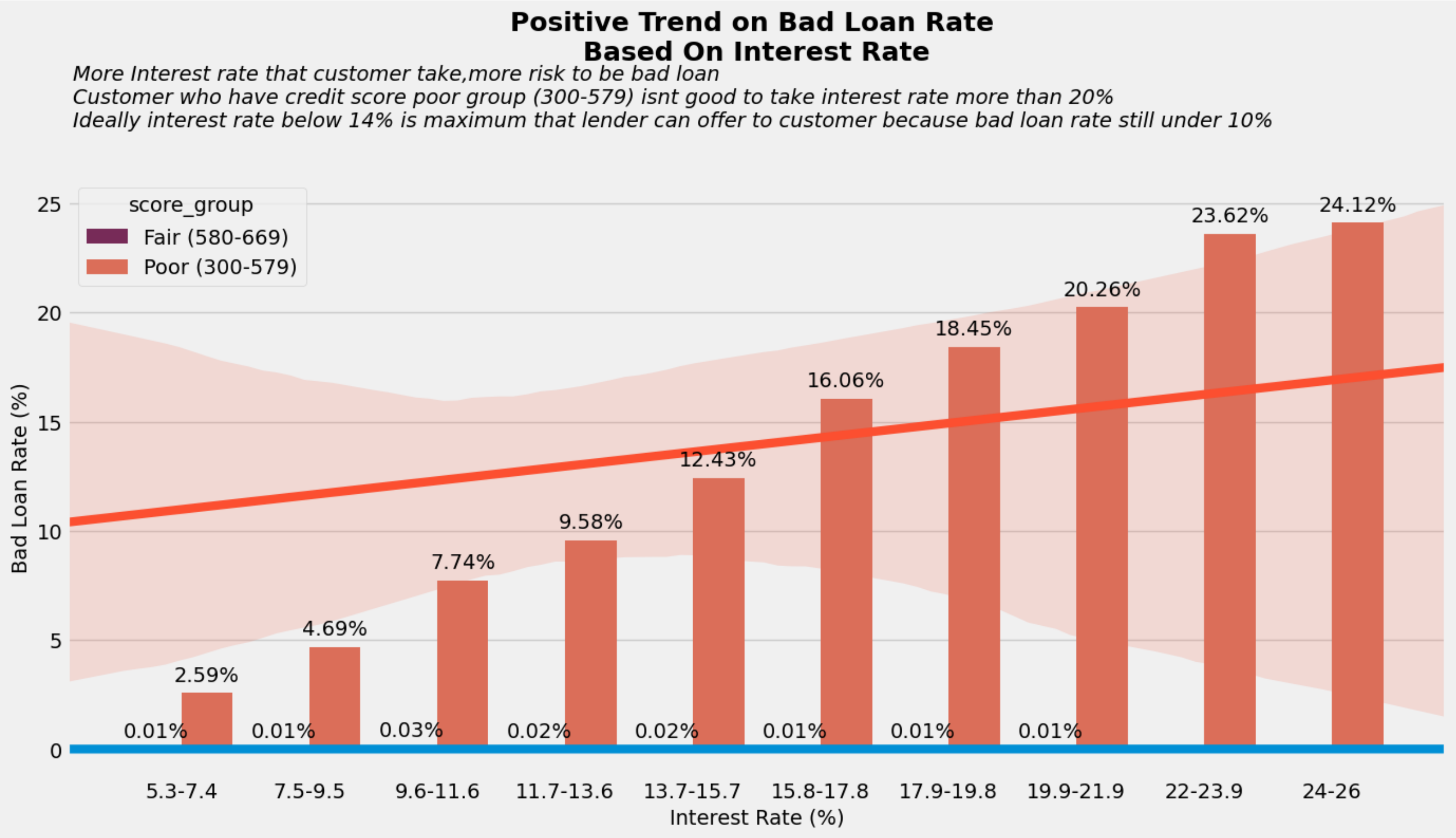
BUSINESS INSIGHT



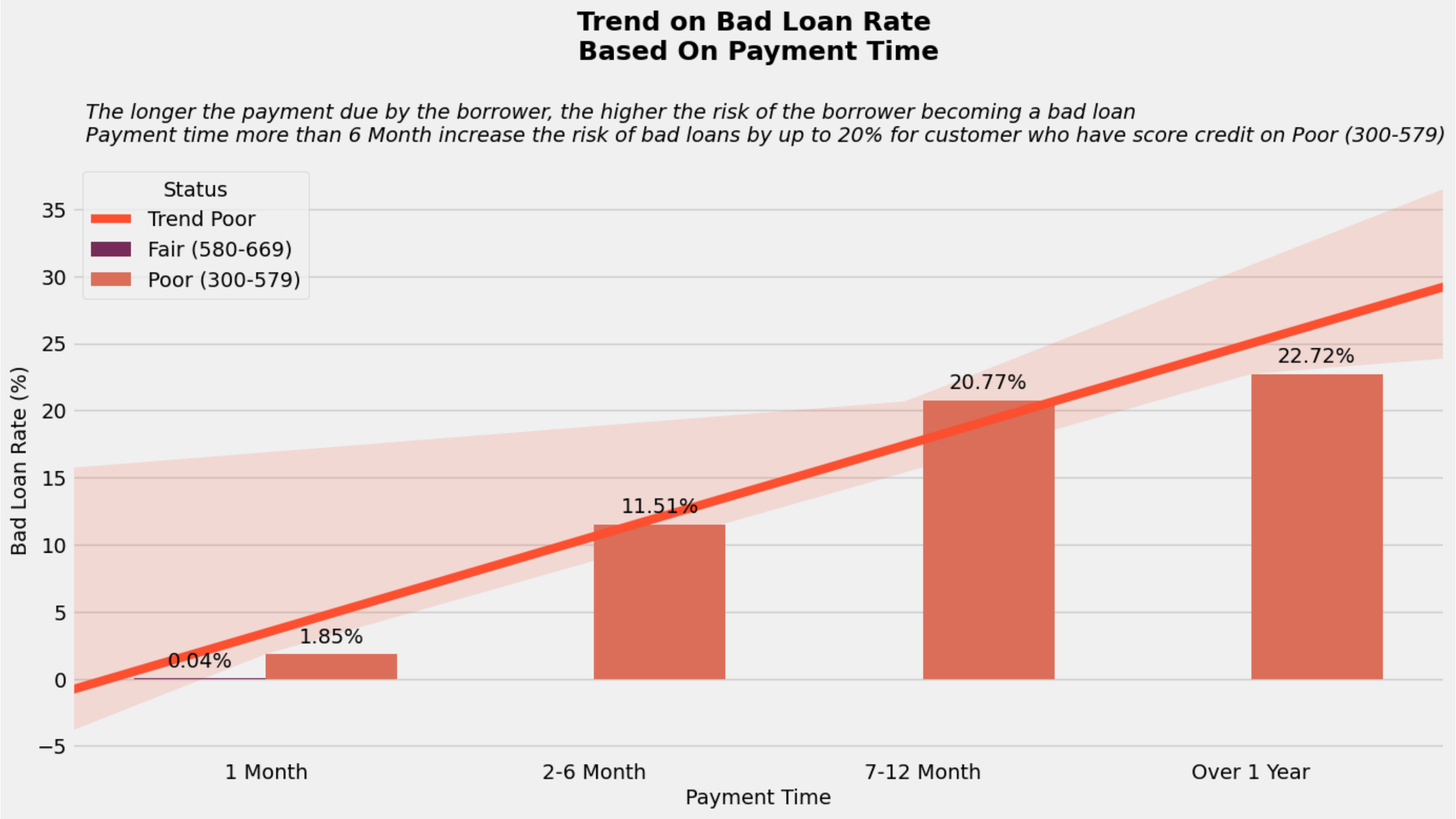
BUSINESS INSIGHT



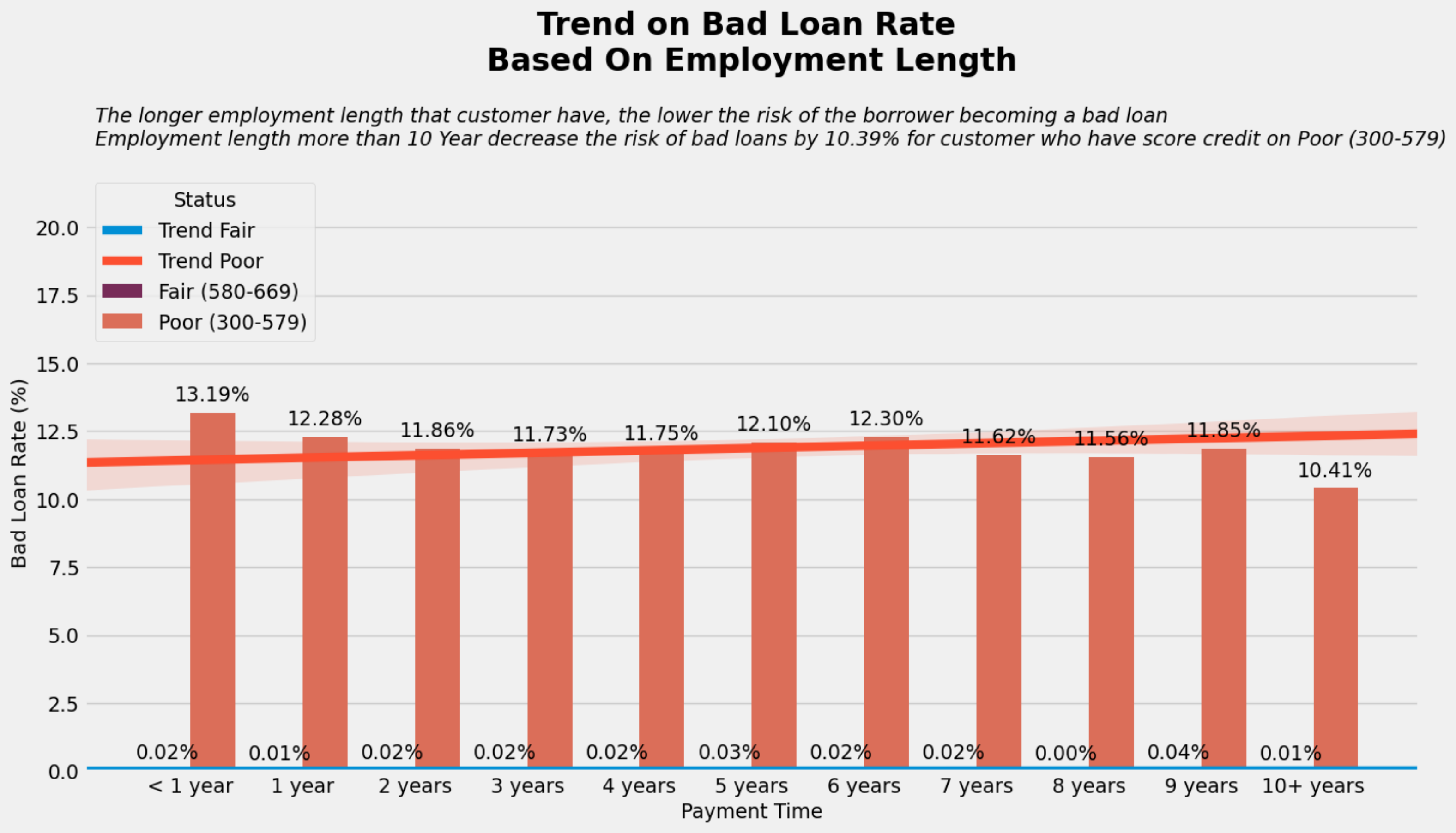
BUSINESS INSIGHT



BUSINESS INSIGHT



BUSINESS INSIGHT



SUMMARY

Loan Amount

The amount of the loan given is related to the interest rate that must be paid.
The larger the loan amount, the higher the interest rate that must be paid.
Loan amount more than 28100 not recommended to offer for customer.

Last Payment Amount

More payment amount that customer take, lower the risk of the customer becoming a bad loan.
Lending Companies can set a minimum amount that must be paid starting from 1500 for the amount of payment each time it is due.

Payment Time

The longer time that must be paid by the customer, the higher the risk of the customer becoming a bad loan.
Limiting the payment time max 6 years can reduce the risk of bad loans

Interest Rate

More interest rate that customer take, increasing more bad loan rate.
Ideally if lending companies want to keep bad loan low, they can offer interest rate below 14%.
Lending companies must avoid to offer loan with interest rate more than 20%.

Employment Length

It has been proven that the longer the customer's work experience, the more capable the customer is to repay the loan thereby increasing the good loan.

THANKS!

Carrier
freecreditreport.com
GOOD
POOR
Your credit score is
780
5311 885
JOHNATHAN