

By : Niken Iarasati Winasih

LOAN DATA PREDICTION ANALYSIS

AGENDA

- Background
- Objective
- Data Collection
- Data Understanding
- Data preprocessing
- Modeling
- Insight

BACKGROUND

Checking the suitability of prospective borrowers against
predetermined criteria

Analyze the credit risk at which the borrower will
return the interest and principal rates

OBJECTIVE

conduct analysis on loan
customers

create a loan data
prediction analysis
model

provide insight and
formulate solutions

DATA COLLECTION

dataset :
loan_data_2007_2014.csv

Data shape:
(466285, 75)

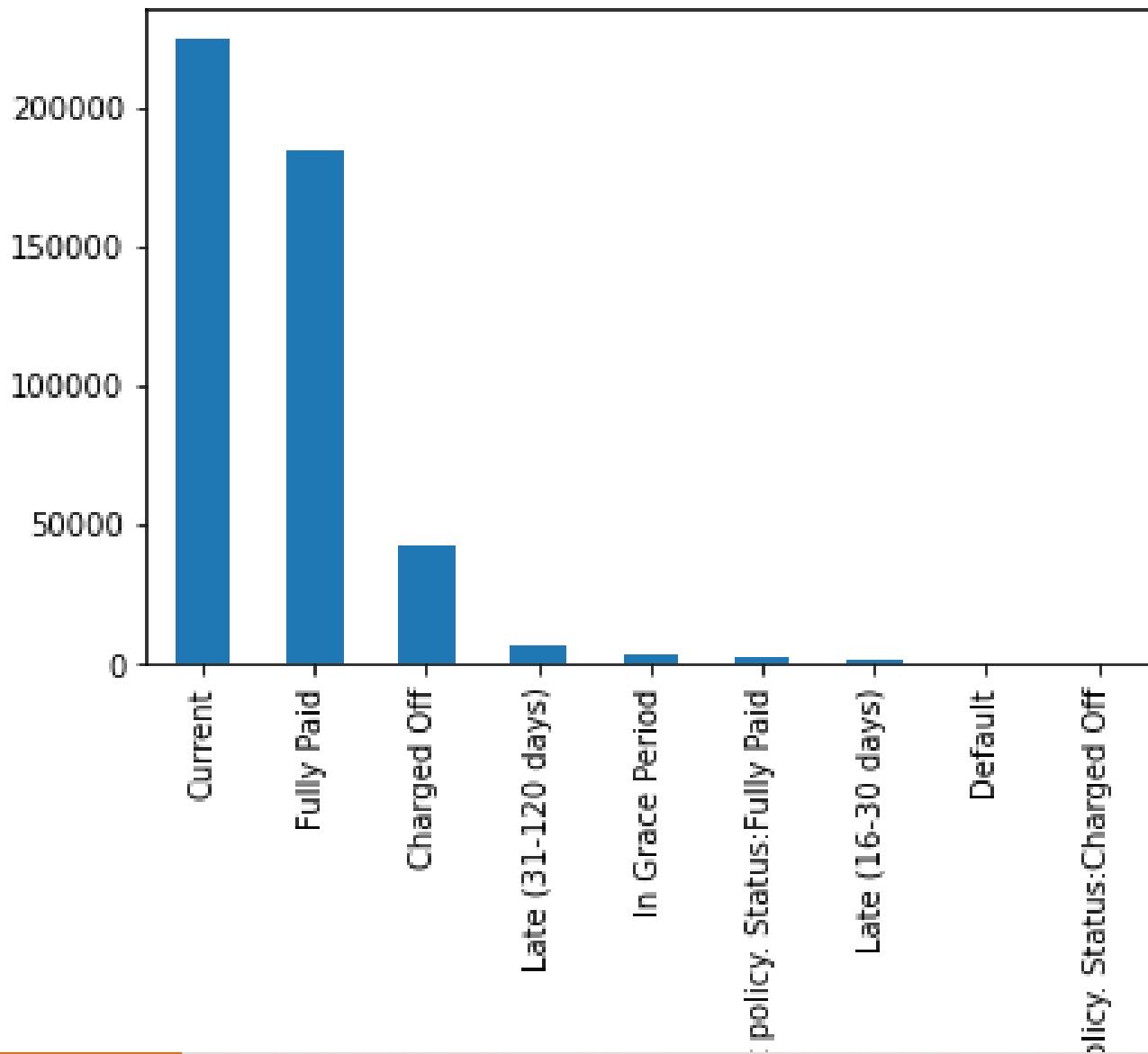
Review column : data
has a lot raw and NAN

Data Type :
Int, float, Object,
DateTime

DATA UNDERSTANDING

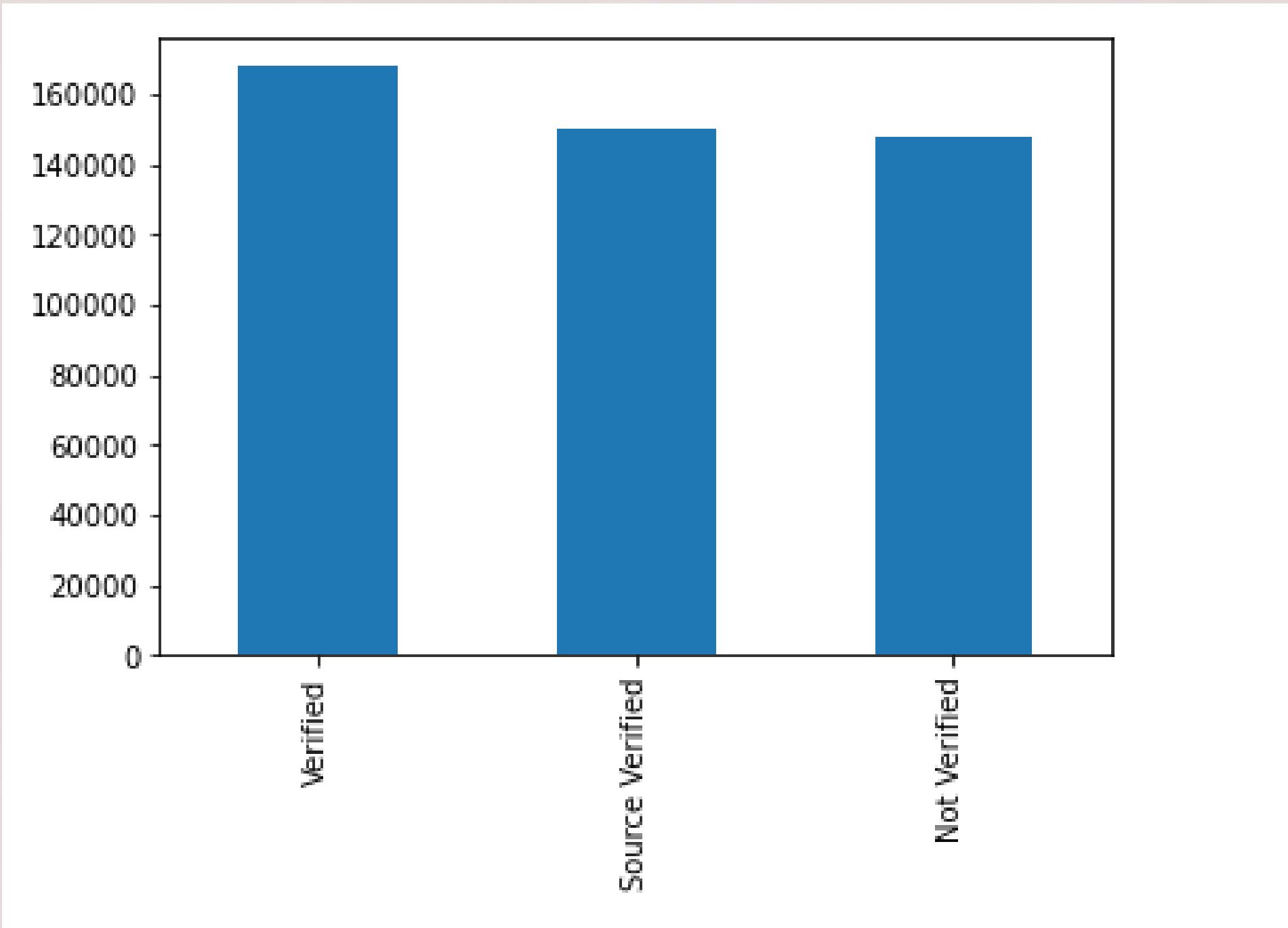
loan status

```
[13]: <AxesSubplot:>
```



```
Current  
Fully Paid  
Charged Off  
Late (31-120 days)  
In Grace Period  
Does not meet the credit policy. Status:Fully Paid  
Late (16-30 days)  
Default  
Does not meet the credit policy. Status:Charged Off  
Name: loan_status, dtype: int64
```

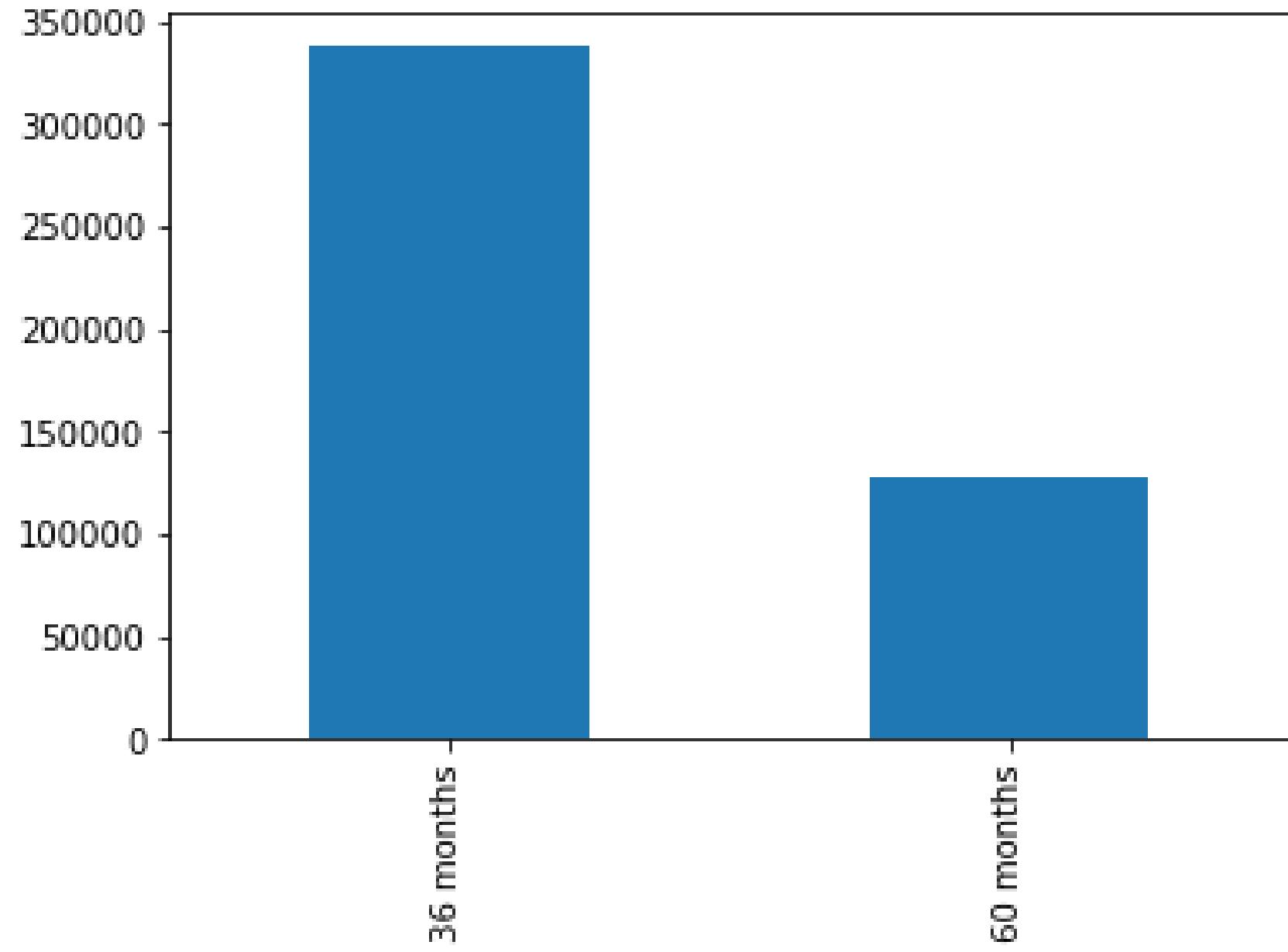
verivication status



Verified 168055
Source Verified 149993
Not Verified 148237
Name: verification_status, dtype: int64

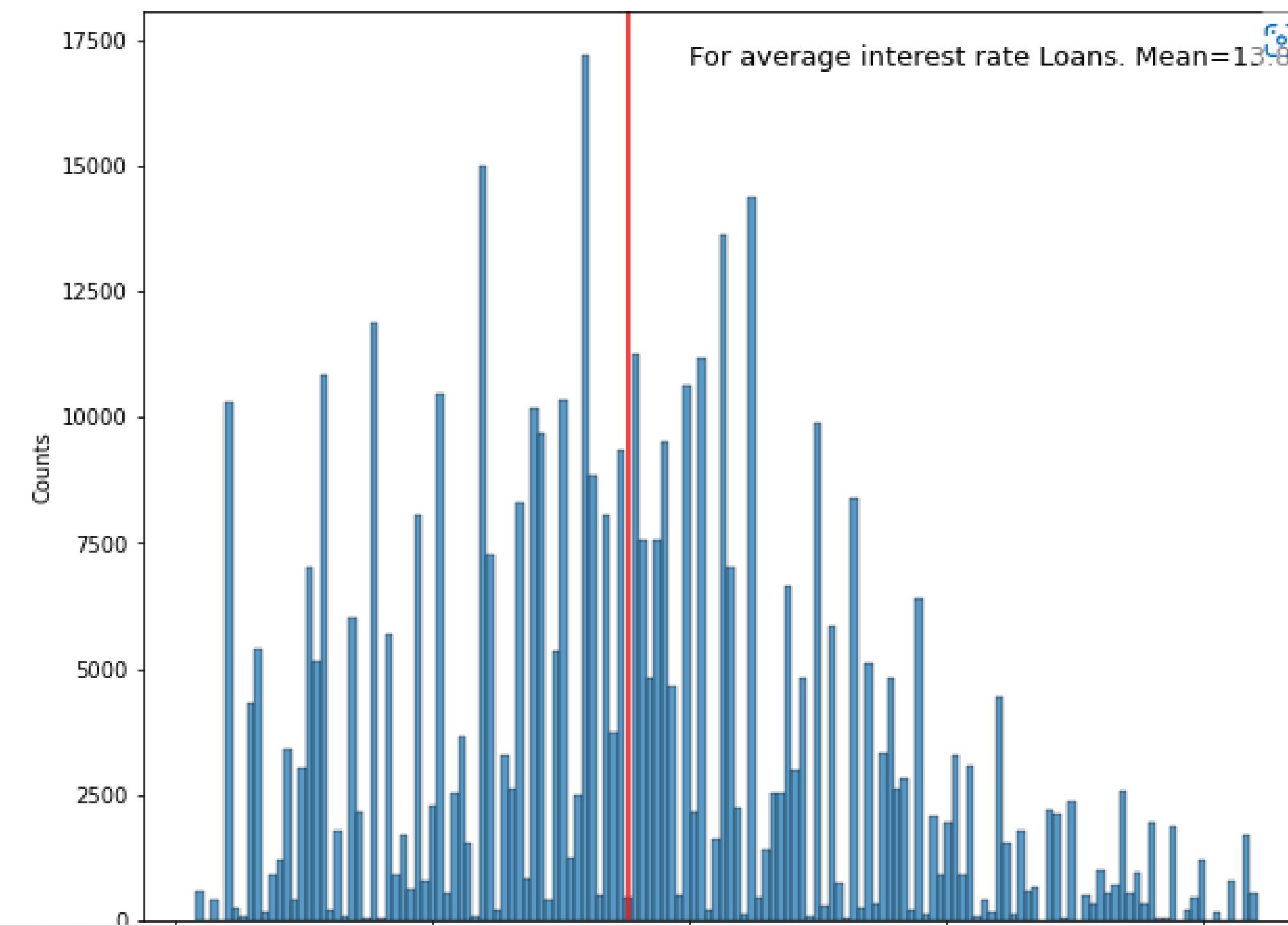
term

<AxesSubplot: >

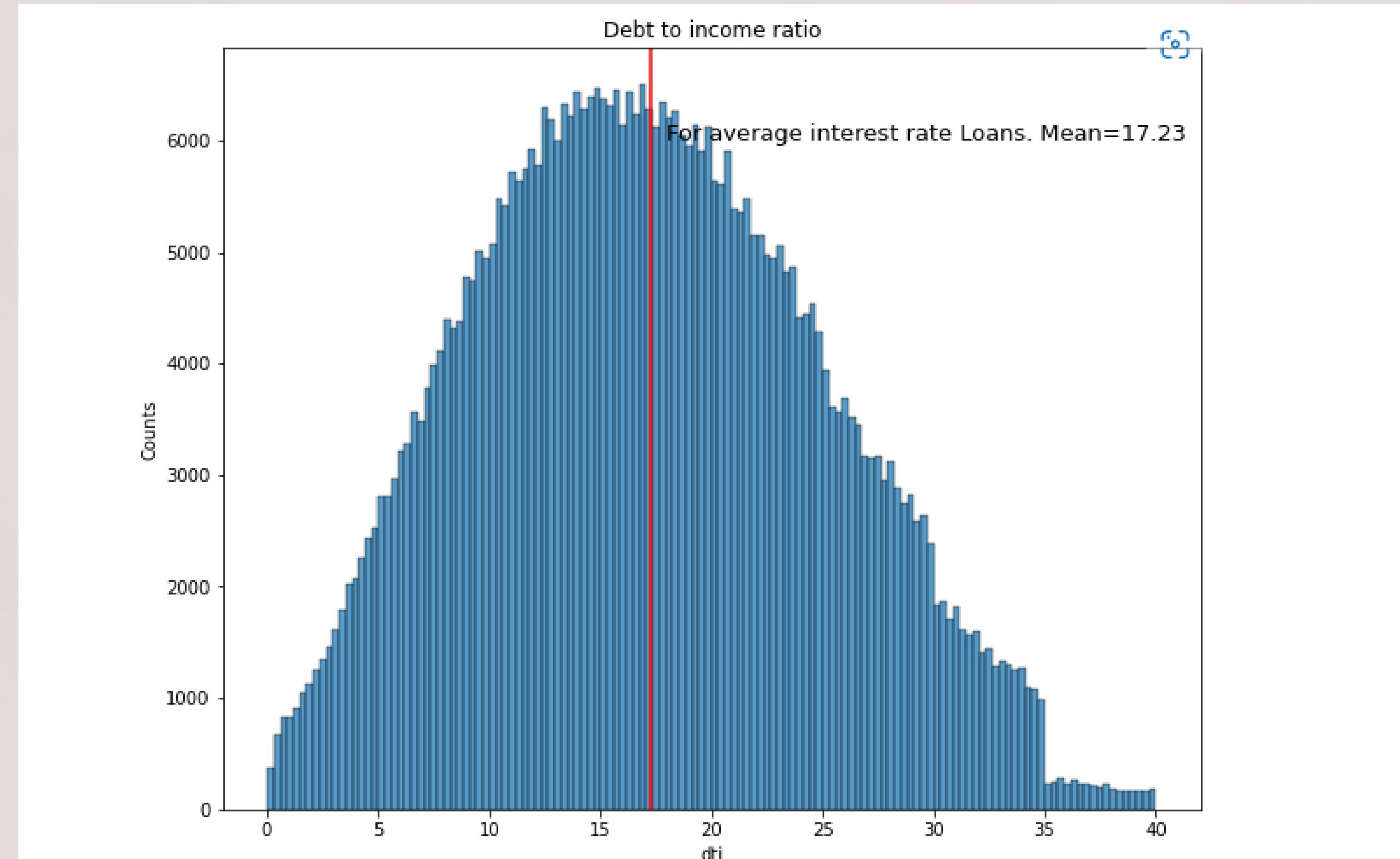


36 months 337953
60 months 128332
Name: term, dtype: int64

Average interest rate Loans

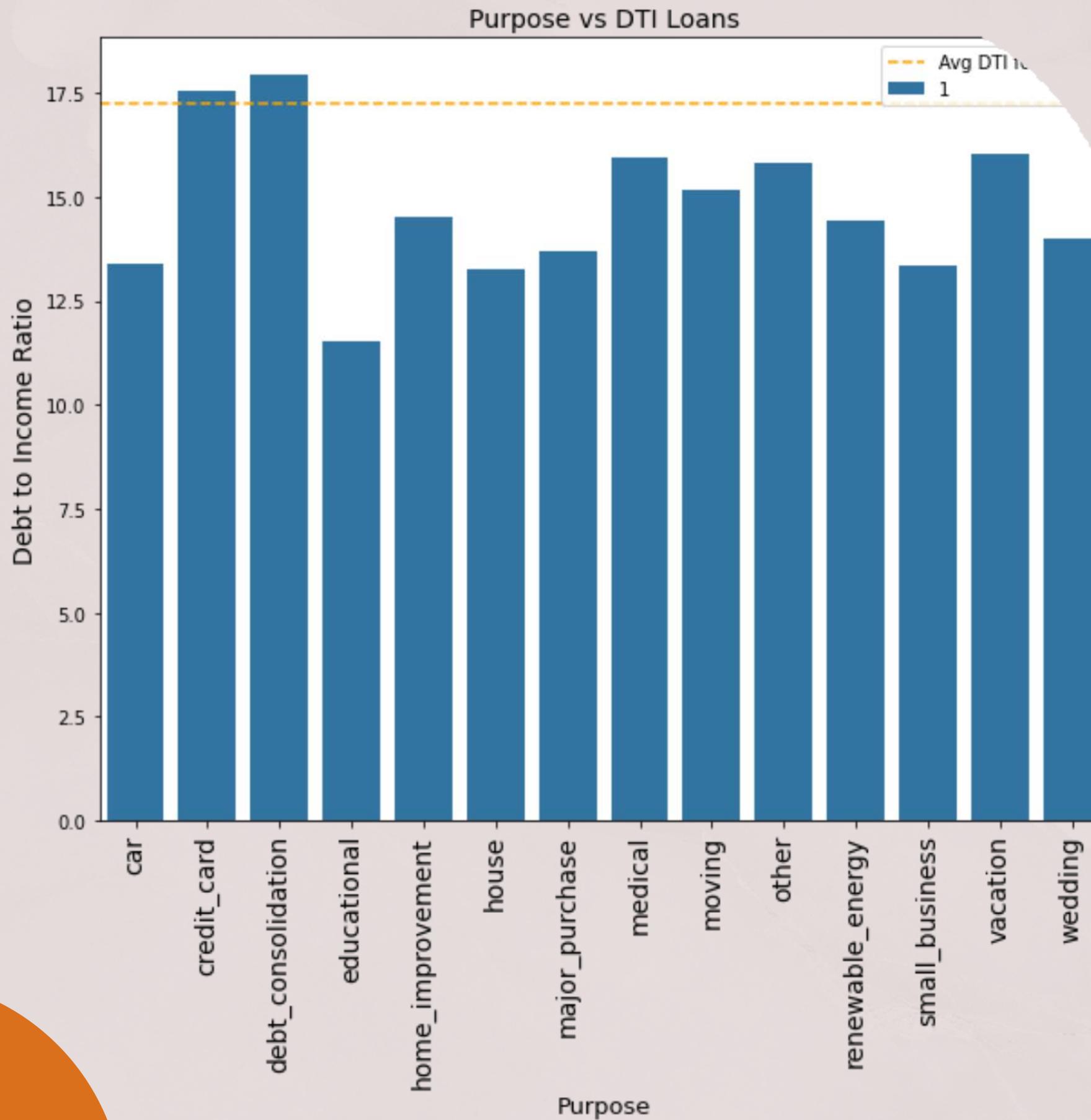


Average interest rate Loans



Rasio Hutang terhadap Pendapatan sebesar 17%

Purpose vs DTI



The higher the DTI ratio, the higher the risk.
Debt consolidation and credit card debt have above-average debt-to-income ratios.

DATA PREPOCESSING



Handling missing
value di emp_title
dengan 'other'



fill the nan column
in dti with 0



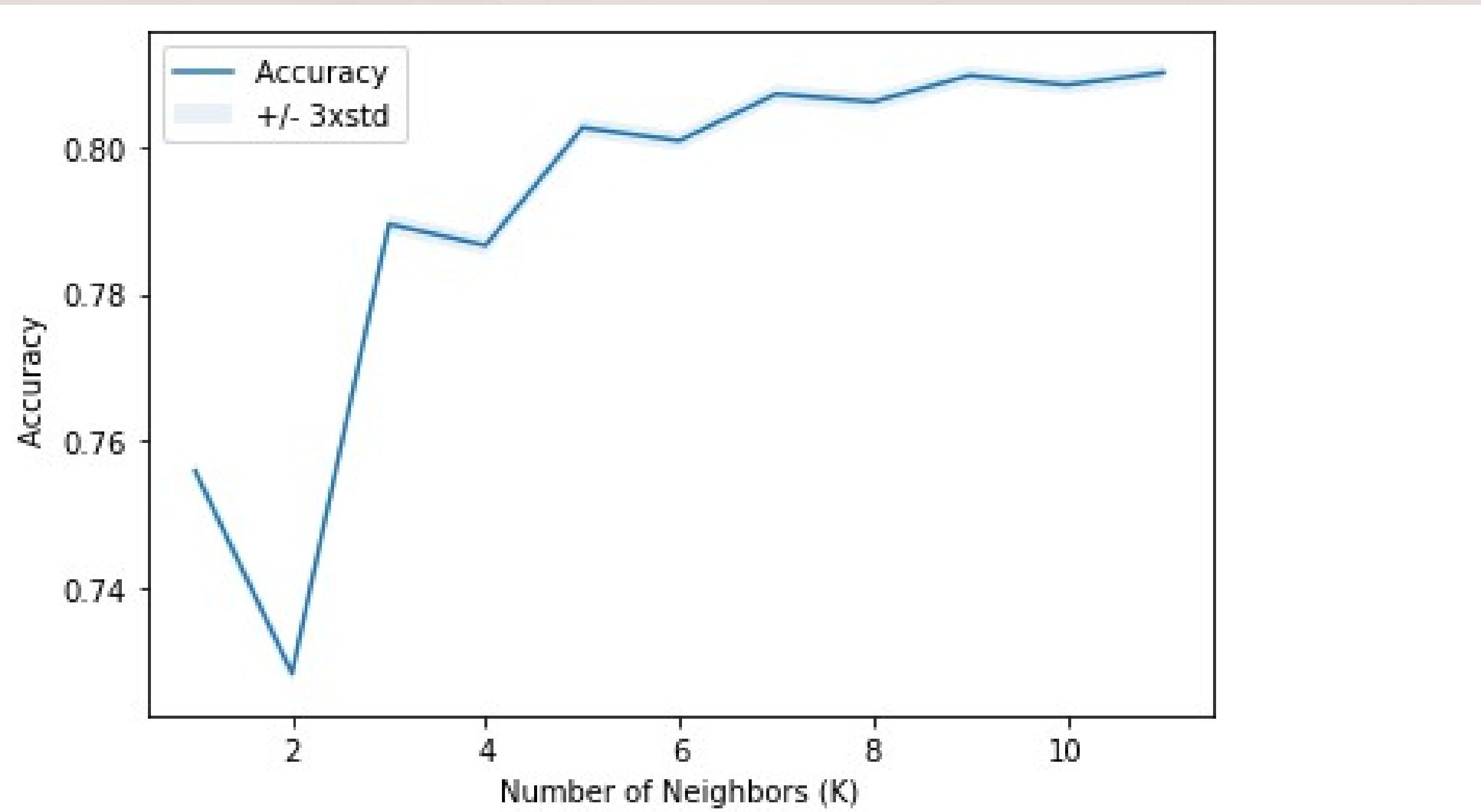
fill the nan column
in funded_amnt
with 0



scaling

MODELING

With KNN

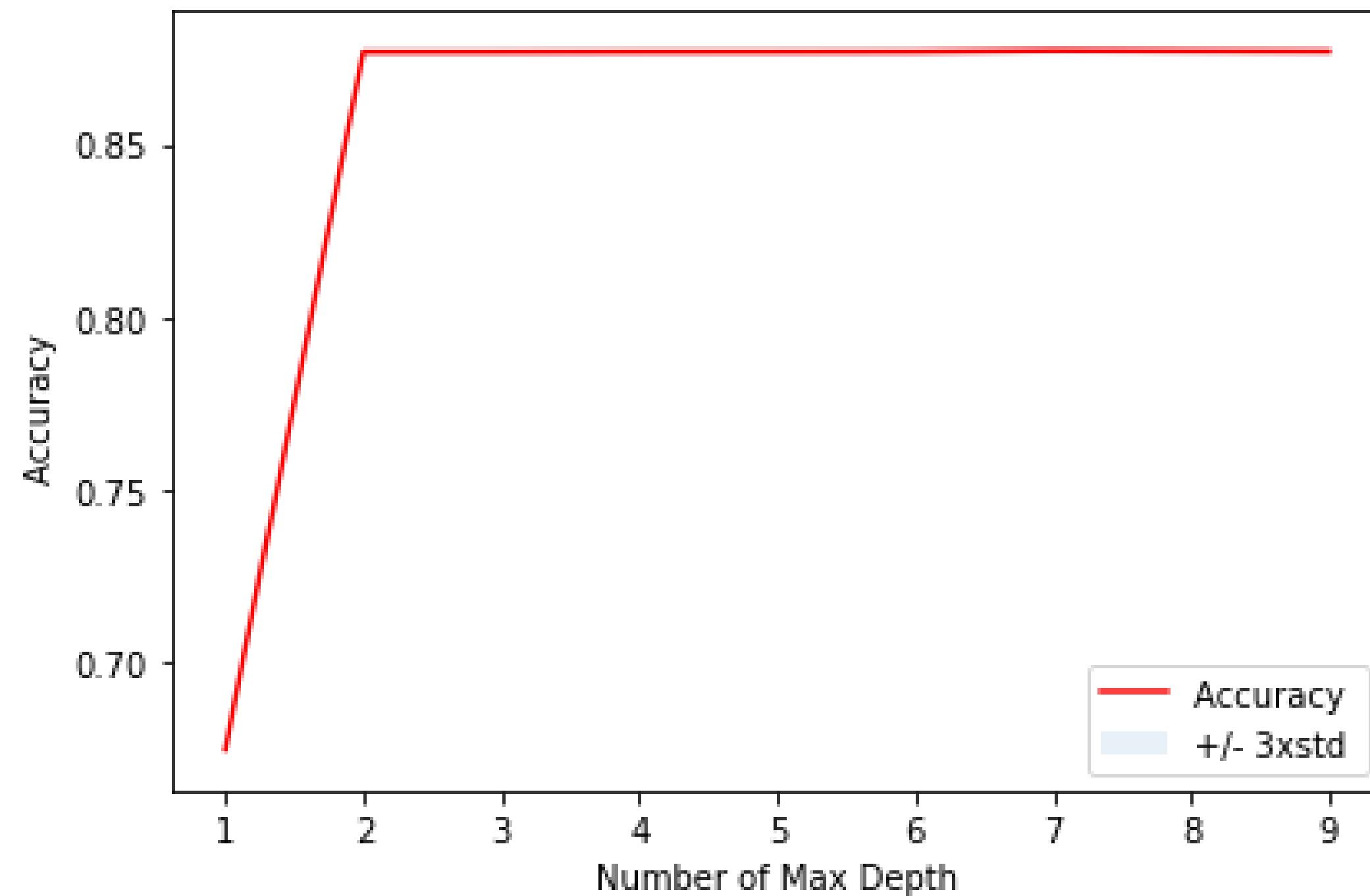


The best accuracy was with 0.8101268537482441 with k= 11

Train set Accuracy: 0.8394034764146391

Test set Accuracy: 0.8072316287249215

With Decision tree



Train set Accuracy: 0.8770360401900126
Test set Accuracy: 0.876877875119294

Debt Consolidation was identified as a primary risk factor, accounting for ~58% of identified high-risk cases.

Recommended a refined policy_code indexing system to distinguish between high-risk (0) and healthy (1) loan applications.

Feature independence was verified through correlation matrix analysis, ensuring model stability and avoiding multicollinearity.

Occupational analysis revealed a higher propensity for debt consolidation among specific professions, such as educators.

Recommended a refined policy_code indexing system to distinguish between high-risk (0) and healthy (1) loan applications.

INSIGHT

THANK YOU