

Problem Statement

Company and investors are not able to identify the applicants profile properly before loan approval and hence they are facing both business and financial loss.



Objective

0,18

cso.

.60

260

Sec. As

0.

0.

3

·\$2.

SEY

62.08

.1

810.68

595,08

0 0.00 CA1000

101.101

25.78

128 851

11.882

00 100

485.5h

48.88

%.

20.00

124

cto

080.00

115,48

10.64

161.09

.چن

30.20

·c,v

100

36.01

· (2)

950.00

17.00

S

20.00

C9. C7

3

.28

.0

20.00

0,000

500.00

कुटिं देश

Yes. 175.26

2.78

.go

6860 W

280

6

6000

97. 168 63. 176

0,7%

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (risky applicants) at a higher interest rate, etc.

35.

101

32.029

29.131

33.323

38.

20.282

12.101

\$2.00

10.02

100.00

600 576 121.78

11.12

012.10

25.05

3000

35.00

72.69

637.24

C.C.

S.O.

611.13

20.0

26,00

061 100.10

1.

808.76

15.61

60° AN

25.72

111.89

V5.24

168.050

020301

25.76

ver les

11,14

200

05.76

A15.08

000

CNS. 21

32.023

23.323

20.22.47

19.518

200

250

VV. V80

16.312

20.030

15.11

12.036

51.345

14.000

6.818

1.60%

15.020

67.21

38.743

20.382

12.101

91:901

7.321

0.088

21.661

19.250

34.609

38.131

ds.33

128 298

152.69

5. V. ON . OC.

96%

.00

40.0%

ch 000 00.00

81.67

192.68

150.10

500.65

121.08

C57.00

53.020

23.221

W. ZV

72.098

369

0.973

155

73.601

10.105

100.34

15.500

3.765

54.322

13.46

36.561

38.161

19.64

28.801

2.518

17.17

24.584

91.7h

989

43.613

32.121

10.10

57.322

VO.38

81.69

8 The Sale 905. AN

600

LY CO.

80.

012:21

15.278

52.025

V.205

10.765

80.

5.161

2.5%

13.004

9.033

88. 1855

75%.78

W.28

012.68

20.00

110.00

42.00

89.00

2007

162.61

25.500

12.926

ON.

60.37A

84.197

821

B.

21.90

00.NS

95.052

D.

6.

60.032

38.

58.21

1.922

22.

0.012

VV.

IIA

3.

151.734

30

31.

97.397

31.780

30.534

1.50

10.

00.

8.

15.

13.291

2002

60.00

851.14

125.60

635.54

% Als. O. 000

26.8

076.46

165.5h

465.46

96.08 100

O. VO

200,7

65.50

00.130

82.623

66.781

1.18

16.911

16.5Kg

86.71

92.179

89.589

45.082

67.307

03.668

63.575

14.601

0.95

008.01

NO.38

15.67

.0,0

cycit!

.C,

25.70

97.8V

25.7%

111.89

V5.24

10.61

101.00

860

180,00

47.44

6020

J. CV

2000

रेश. दि

300.00

69.306

1.283

54.014

15.991

30.878

47.330

36.01

92.292

8.00

39.74

29.72.091

76.076

81.35

16.021

51.651

3.

52.098

70.532

92.199

23.662

10.130

050.255

85.05

26.354

17.121

2540

91.821

13.99h

91.255

82.89A

12.903

35.31

23.164

13.003

29.608

34.168

22.232

23.261

13.088

W1.05V

64.555

83.579

20.738

10.814

3.051

84.57

66.986

160.889

14.00

98.883

12.732

82.10A

10.00

16.693

8.202

59.182

15.914

11.819

OV. 220

18.003

10.316

17.062

204.01

109.67

60.00

Con Con

178.00

80.00

906.26

2005

N.S.

200

102.18

25.75

29.131

26.039

6.729

77.016

80.Ml

64. V2

7754

15.158

15.293

22.662

95.151

66.50

8.679

12.960

94.883

92.902

10.086

39.16

1.570

30.

1.50

3

67.672

81.91

55.04

97.891

81.382

34.918

31.251

28.119

62.347

22.766

NA.878

31.10

77.355

15.11

20.13

32.11

29.89

16.899

30.18

10.564

18.192

86.560

3.55

23.323

12.030

57.345

14.688

6.848

1.60%

NO.020

67.217

CAN

563.11

10

96.924

23.21

5.30

34.609

36.

0.

63.84

17.653

38.12

20.282

12.101

91.501

1.321

0.088

2.80.184

0602

21.664

16.293

22.21

65.634

V2.23

18.57A

NO. 989

2.544

10.565

92.908

85.08

69.593

13.601

81.201

Exploratory Data Analysis

Data Sourcing

 Campany already given the dataset contains the information about past loan applicants and whether they 'defaulted' or not.

 For this case study we need to focus only on provided dataset given by company.

Data Cleaning

01

Remove columns having only NAN Values 02

Remove columns having only zero values

03

Filter out variables which could not be used for defaulters

Data Cleaning Variable Type

Variable related to applicants.

Variable related to loan characteristics.

Variable related to customer behavior.
These variables generated after the loan is approved.

Selected variables

- 1. emp_length: Employment length could be directly relates to financial status of the borrower.
- 2. home_ownership: Home ownership could be directly relates to financial status of the borrower e.g. If borrower is already owning house or morgage or on rent.
- 3. annual_inc: Annual Income.
- 4. addr_state: Just kept it to analyse the state wise loan status.
- 5. dti: Debt to income ratio could play important role to identify defaulters.
- 6. loan_amnt: Loan Amount
- 7. funded_amnt: Funded Amount by Company
- 8. funded_amnt_inv: Funded Amount by Investors
- 9. term: Terms of loan
- 10. int_rate: Interest Rate
- 11. grade: Grades
- 12. sub grade: Sub Grades
- 13. verification_status: Loan verification status
- 14. loan_status: Indicating defaulters. Possible values Charged Off (Defaulter), Fully Paid, Current
- 15. purpose: Purpose of loan

Data Cleaning

Strip the % sign from interest rate column and convert to float

As per data dictonary possible values for home ownership are RENT, OWN, MORTGAGE and OTHER. But in data set NONE is also present. Hence moving NONE to OTHER

Loan status marked **Current** are neither **Fully Paid** nor **Charged Off(defaulted)**, So we can get rid of rows having loan status as **Current**.

Tagging the other two values (Fully Paid and Charged Off) as 0 or 1 to make your analysis simple and clean

Remove NAN values from Employment length

Univariate Analysis

Unordered Categorical Variables

Home ownership: There are many people belong to mortgage category.

Verification Status: More than 16000 loans are not verified.

Loan Status: More than 5000 loans have Charged Off out of approximately 37000.

Purpose: People are taking loan mostly to do debt consolidation.

Address State: People from state with code CA, NY, FL, TX and NJ are the top 5 for having loans.

Ordered Categorical Variables

Employment Length: Most people taking loan having employment length more than 10 Years or less than 1 year.

Loan Terms: More than 25K loan are having terms of 36 months.

Grade: Most people avails loan at lower interest rates having grade A, B and C.

Sub Grade: Most people avails loan at lower interest rates having grade A1-A5, B1-B5 and C1-C5.

Quantitative Variables

Annual Income: Too many outliers present in the dataset. Filter out the 90th percentile from dataset. Mean annual income is around 58K.

DTI: Count of loan is very less when dti is greater than 25.

Loan Amount: Count of loan is maximum for 10K loan amount.

Funded Amount Investors: Count of loan is maximum for 5K and 10K amount invested by investors.

Interest Rate: People are intrested in loan for interest rate around 7 and 12.

Segmentation



Annual income of defaulters is relatively lesser than the annual income of fully paid borrowers. mean of annual income is relatively higher for fully paid.



Debt to income ratio is relatively higher for defaulters. 75th and 50th quantile of dti is relatively higher for defaulters.

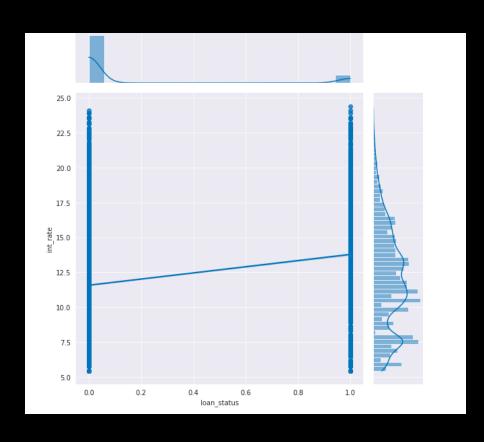


Loan amount are relatively higher for defaulters. 75th and 50th quantile of loan amount is relatively higher for defaulters.



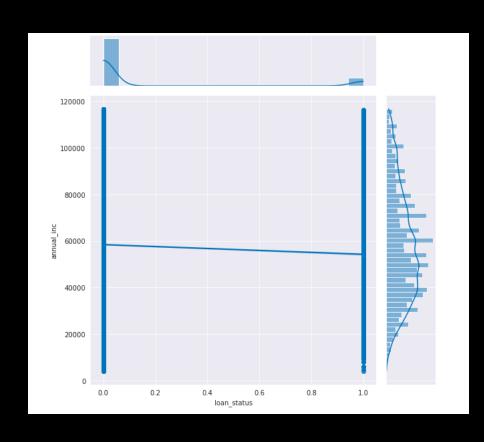
Amount invested by investors is relatively higher for defaulters. 75th and 50th quantile of funded amount by investors is relatively higher for defaulters.

Bivariate Analysis



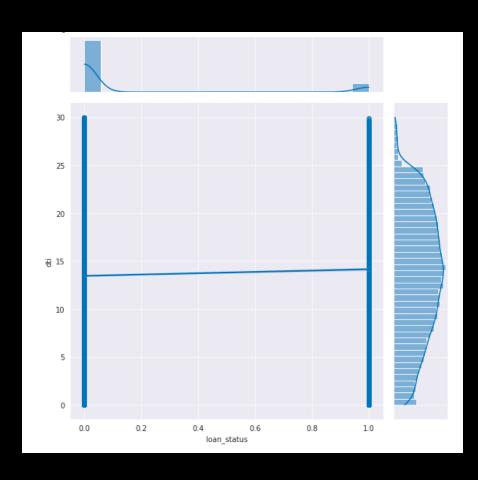
Interest Rate vs Loan Status

Interest rates for defaulters is slightly higher compare to fully paid borrowers.



Annual Income vs Loan Status

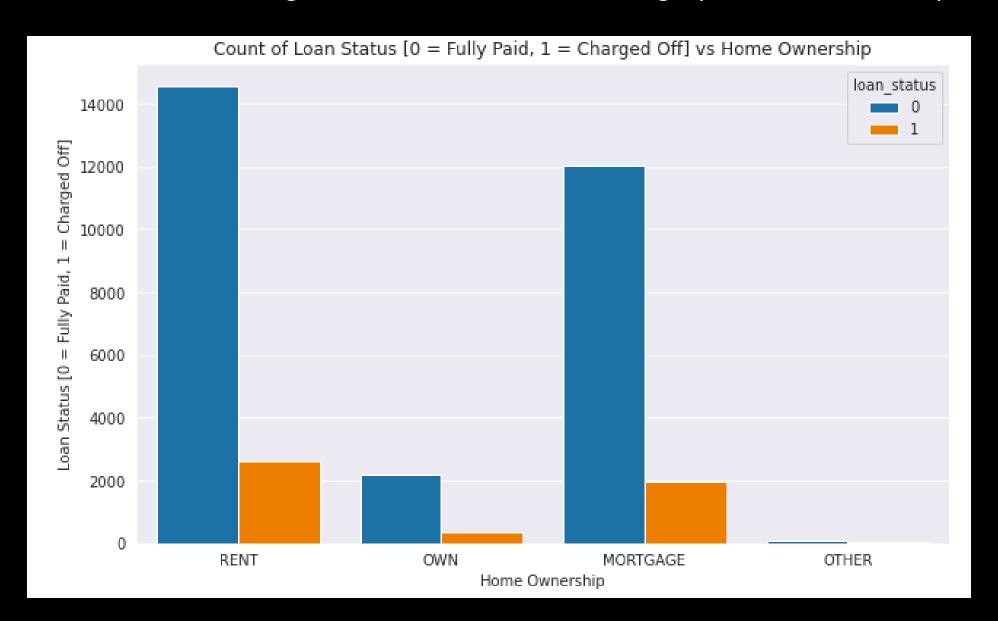
Annual income of defaulters are slightly lesser compare to fully paid borrowers.



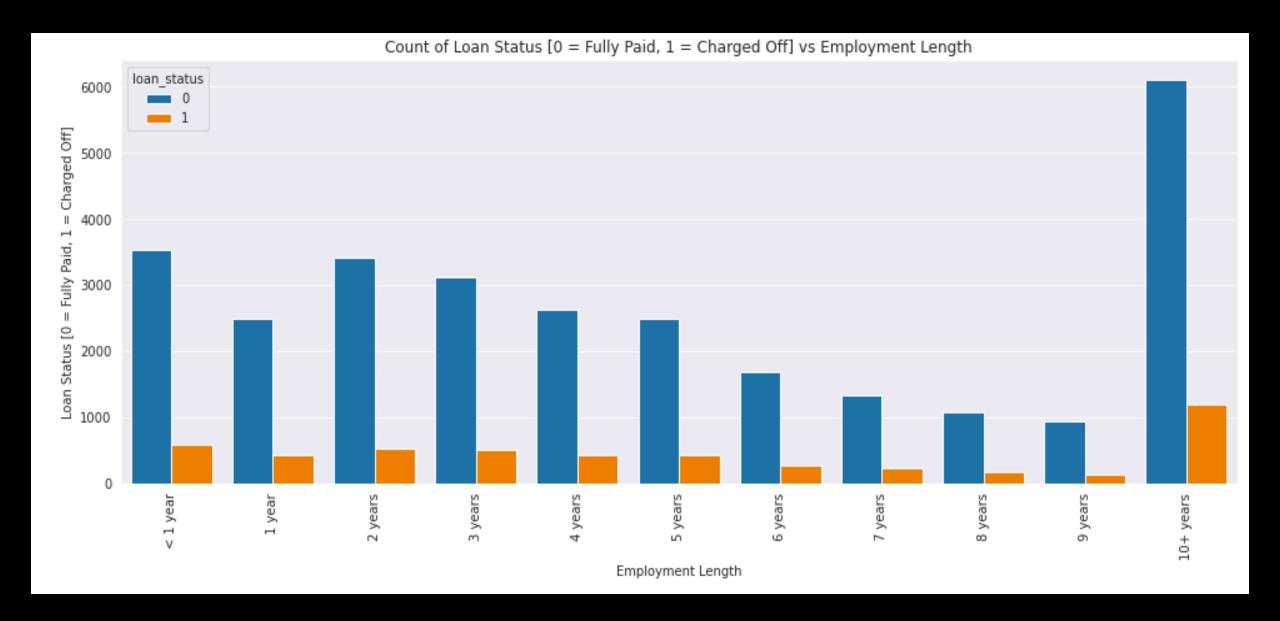
DTI vs Loan Status

DTI ratio is slightly more for defaulters.

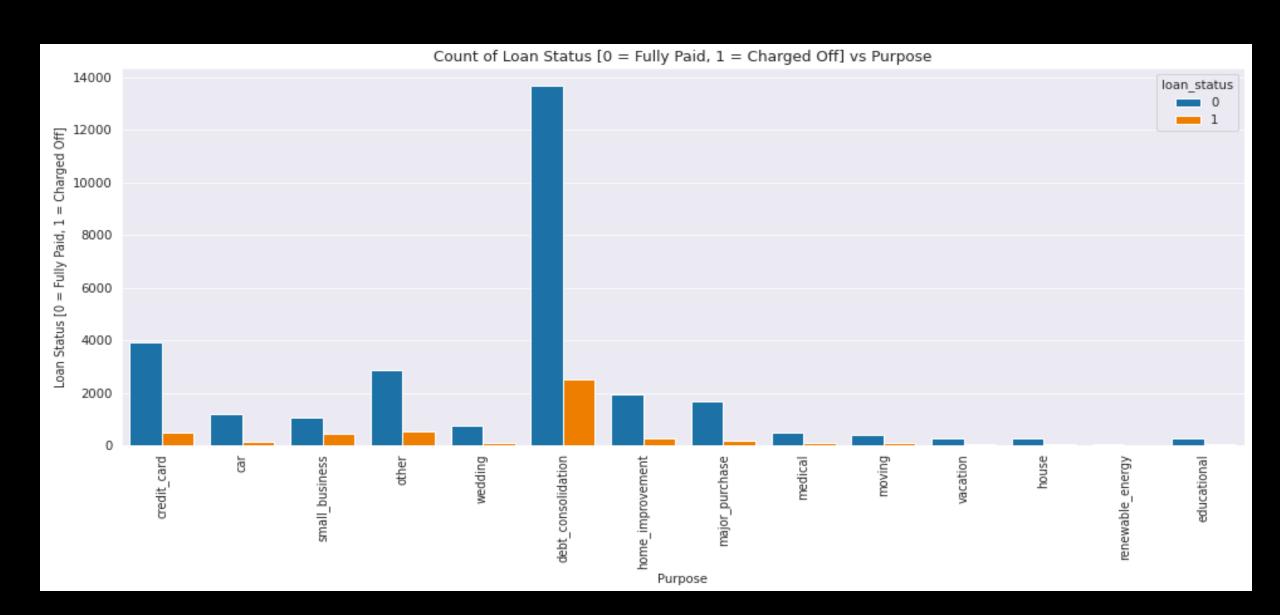
Most defaulter belongs to RENT or MORTGAGE category of home ownership.



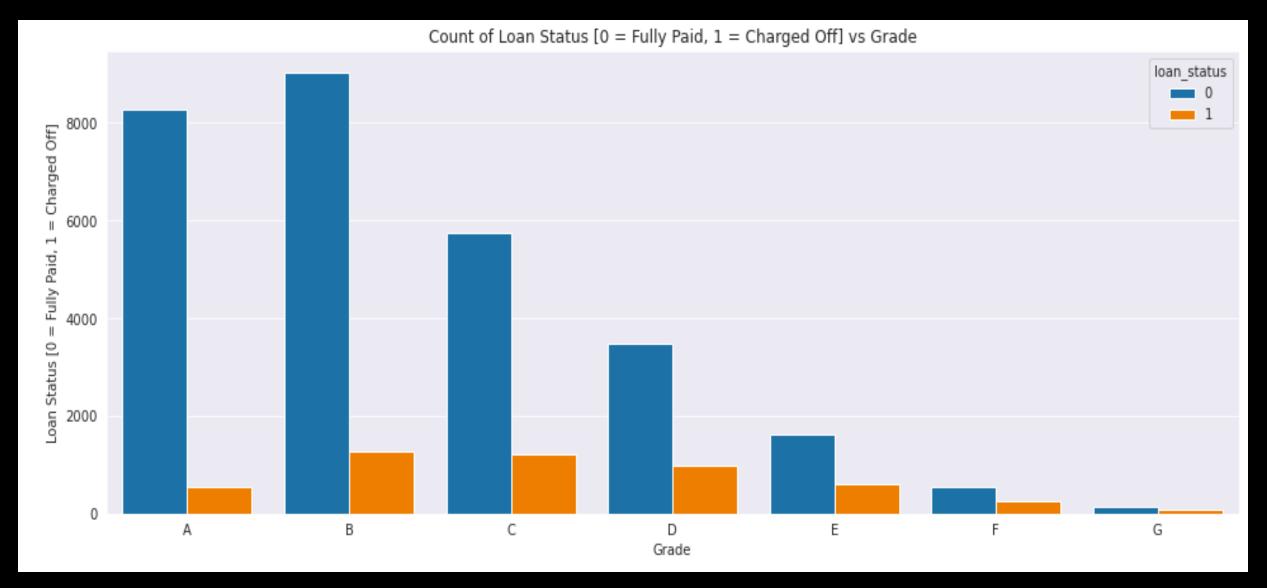
More than 1000 defaulters are having employment length greater than 10+ years.



Number of defaulter are more than 3K for debt consolidation purpose.



Ratio of defaulters after grade B is quite heigher. Tendency of defaulters are more from group C to G.



Recommendations

- More than 4.5K defaulters are from MORTGAGE and RENT category of home ownership. It will be great to check the DTI and purpose for these categories before approval.
- More than 16K loans are not verified. It will be benifial to the comapny if they can reduce the number by strenthening the verification process.
- People are taking loan mostly for debt consolidation (one of the category in pupose), Approximate 3K loans are defaulters in this category. The total number is more than ~17K out of ~37K. This is a huge number. Company should think of their business approch. The number of loan for other purpose e.g. house, renewable energy and education is very low.
- More than ~1K borrowers become defaulters out of ~7K borrowers for 10+ years category of employment length. The ratio is 1 out of 7. It will be great to check the DTI and purpose for those application before approval.
- Approximately 13K loans belongs to category of less than year or more than 10 years. Please do proper verification before approval.
- There are more defaulters belongs to state CA, TX, FL, NY anf NJ as there are more loans from these state. Please do strenthen
 the process of verification in these states.
- DTI ratio should not be more than 13.5 to avoid defaulters.
- Average annual income should be greater than 58K (USD) to avoid defaulters.