571.28 69.306 68.971 210 2.28 050 194 100 SS 886 10.00 2.020 132 5.18 35 2 8 0. 5 R 6 B 6.01 K. 600 53.020 5.64 .14 c.p. 0 536 V01-58 13 K. 23.765 01.80 No 15.991 .78 65.76 53.000 PM 1.11 10. 54.322 68.971 073 2. 66.281 000 1. A1.758

6634

51.963

563

92.08

800

10.102

5v

100

.v

00

0

U

5

.00

.00

5

2

19.61

860

.gu

3

No.

50.00

Sp.

230

602

.Su

.ev

690

02.12

.e.2 e.2

201

.18

120

.22

VV.VV

24

0.

28 So.

57

av

.go

69.98

Q.

-59 -510 -510

10.07

19.11

057

.2

85°.51

200

-28°

.ç.

50.72

.2

600

.cu

102

00.

250

.2

.ch

20.50 V

204

20

00

oly

OSY

.51

000.00

c12

50.01

520.000

50

161.09

Noo V

.du

14

.ch

Nº.

.0

16.91

10.18

.0

2

686

cel

.6

10.12

000

6.

8

.cv

5000

.2

SS

eg.

613

·So.

16.02

.V.V.

S

127.10

.20

.02

Voy.

- CSS .00

00 050 100.07

080

.3

64.14

550

20.

665

.22

5

296

.00

0.

250

.08

808

00

128.00

.64

640

12 09:98

.2

12.13 160

.2

Sol 2

.20

Sos

020

\$

joy

.0

ST 259.11

2.

280

11:18

10.02 V802

.2

200

.2

02.00

080

819.24

.2

0.92

.2

680.67

.10

02

S.

826

0.

200

V28.14

8.

055.76

990

.0

570

.c.

1030

17. 17 89-17

.2

35

.20

250

V1001

.2

619

.20

810

068.91

8

.1

100000

.20

00

271.65

.0

61.00

So.

2000

.86

11 IV

Lit

202

500

.00

000

. So

000 12.00

.0

600

.00

200

20

168

150

The second

2

220.02

.6

CS I

.00

Si

.60

.07

200

.2

2007 .25

.....

20

000

18

2502

.c.V

isi a

.cs

10

S.

60.92

aves

219

.cr

SY

C19

.0

212.01

68V

.19

55

.6

S

.2

690

10: .020

So.

001.02

2.20

262

.10

162 2

.02

.60

250

032

29.

3.

39

0025

es.

.0.

S.

20 V

6.

860

2

1682

8.6

50

2.

17.08

016

v.

.0

.

100 V

Ser

120.01

she

\$.

169.24

.c.V

Wh

.60

.20

01001

100

.60

0000 c16

.0

818

N.

100.19

286

.07

6001

.05

160

·80

2

.c.

NOV.Y.

c.86

·00.

es es

.gv

11

186

2.

.8

025

.00

240

00.

23.765

51.322

3.

501

38.161

2

001

0.

176

30.

201

S.

34

15

1.

2005

55

24.00

573

0.

019

63.941

69.63

294

5.

10.50

02.008

501

1.321

08

0.

0 029

6.

2.514

282

12.07

01.

222

1.

198

10.

12.030

51.5

0.

3

23.323

198

32.029

30

600

17.653

123

D.

.00

83.155

1.

73.601

102

19.

8.

53.026

932

06.924

2

AA.TA

2

0.109

00

rad

Nº.

0

200

14.100

0

04.990

75.278

10

10.

52.625

30

16.00

150.

V.

502

04.033

263

00

3

S.

10.372

6.129

5.

19.50

26.039

672

2

5

15.11

20.13

100

32.11

20. 00.

16.000

20.148

89999 9999 9999

NOS

N

63.032

29.131

23.323

72.036

51.342

.1

SEV

62.08

.

01002

000

.02

206.26

.20

LYC

500 - 500 600-50

29

210

.02

25

.07

000

10.92 V.92

.~

.0.00 BP

.5

CY12

.2

224

889.11

. So

20

STE

950

.20

864

.02

.Su

- Si

671.00

096

820 c951.02 191.19 6 33 26.00

128.07

11.

SY

.8

052

.c.U

cos

~ 100 026301 5. 138

20.02

600 60.00

192.68

530

.¢

. 20

60.62

5

125.09

0.92

8.

69.

12 V

1.12

·00

e69.02

81.251

612.19

0.0

250 09

17

.0

-200°.77

18

2.28 VV

012

.80

20.05

1100

25.50

2.

250

2

555

926

13.291

20.

19.198

76.198

32.029

so.

12 000

6. 8MB

6

à

3

a

11.122

17.653

123

20.282

2.

1.605

302

15.020

6.

2.54

6.

101

81.01

7.321

2

2.291

So.

0.98

21.661

101

.0.

19.130

- Pis

34.60

089

38.131

\$3.

996.94

162.51

20

52.22

111.00

80.

208

NO5.24

.ch

89.99

1.21

161

00.15

240

57. 0.057

.08

520

250

.2

0.00

120

S. C.

.24

6500

0

5.12

. Su

020

- SS

200.00

V12:09

V

.00 2

2.2

162

851-1V

.15

000

.50

110.00

698

55.51 555

·v0

000

.60

· 10

2057

194

V

Vir oss 10.02

VA

6N . 6N

6.

6.07

24

823.61

0,98

8

101.000

100

· · · · · ·

220

8

692.95

009

.~

10.92

529.25

3 250

·%

150.10

60.63

25.25

200

444

0.

20

N.S.

cor

50

012

.2

050

083.155

69.593

73:01

00

00.00

15.00

201

23.121

0.709

NA.TA

72.098

200

13.65

3

61.859

35.013

28.004

12:518

S.

100

30.55

38.161

19.61

N1.111

24.384

01.574

089

3.

50.00

Po-

32.121

59.821

71.173

1.

10.1.10

57.322

12.

63.81

00.

250.00

.0

100

oly

.04

So.

602

w.

220

·G

000

121.09

0

.

.c9

c10

12

N'S

.6

.0. 277 2.0.

.8

50

1

252.28

.15

00.00

WW

SV 11 1000 EN

.28

512

120000

161

10.

2.

73.76

16.38

10.00

150.265

89.

5.101

P.

or.

5.5

80.

091

9.033

2.

or.

55

0.

84.97

00.019

050.52

328

30.

ngo.

21.00

SIA

2.

3 032

8.

593

30.

30

6.

22.

0.012

2.

2

37.942

51.381

N.

NI.

51.34

2.

31.

01.001

21.014

57.199

20.

ont

0

2.02.

5

1.74

.00

002

60.05

1100

000

Cy' STO

N

162 .8

·So.

6

.ch

80.00 .00

.02

18.51

0750

025.26

S

051-14

N.

1.003

21.599

00.736

82.623

5.

293

10.011

63.129

92.179

31.59

31.12

20.

50

\$.

00.

60.03

710

8.

29.00

Nº.

80.

0323

S.

de

5.

50.

1.

03.68

63:575

00.50

10

698

.0,

1054

....

60

0,92

So.

c200

.00

Nº Nº

230.10

176

.05

000

07.

2007

.

5

Jo.

107

v.

529.25

820

.2

0.

Su

6.

10'

.07

OLY

.cy

0,98

260

.00

1000

VY4

.Su

602

.cv

230.45

.20

050

5.91

0.

28

....

20

80.

721

36.01

g.

13.878

51-330

29.013

2.

076

81.375

221

51.651

3.

16.021

52.098

278

02.199

17.078

20.000

10.532

23.662

10.139

óp.

08.013

68.68

23.572

2.

5.

1.748

16.513

86.11

69.599

1222.082

67.301

14.661

10.

51.981

D'

127.52

102

25000

.02

100

8 200

3.

200

.2

SS

6.

148

ego. OLY

610

28

100 .0 NO. 867

.

20

1.

.2

200

.0

N.S.

.5

50.5

.0

640

6

.6

262

Va

8.

69.593

13.

10.102

50.00

60

19.430

sh.

10.989

82.3

Nº.

St

10

:01

81.201

P.

15.00

30.932

00.

924

0

22

AA.

5. 5. 60

:00?

er op

d'

2

12.

63.9AT

200

6.

2.50

10.50

201

91:01

7.321

0.088

d'

100

16.293

92.908

85.08

122

N

029

38.13

20.282

2.

1.605

15.020

67.217

2. 2. 8

198

0.

323

72.036

5

28.99

16.896

20.148

10.00

78.152

86.56

160

A.878

10 X.

77.355

31.121

3

2.

3

29.03

g.

10.00

057

2.

81.617

3

73.971

30.

20

28.00

3

60

3

25.5

0502

13.013

22.211

65.634

3.

18.971

S.

50

33

032

S.

20.00

15.11

20.

55.794

97.891

81.382

34.00

31.251

28.19

050

01.883

62:311

d'

92.992

10.086

39.76

7.579

d'

150

2

84.51

66.990

3

S.

32

6.129

67.672

151

153.138

75.293

22.682

05.151

66.519

8.519

2.

15.914

232

23.261

12.000

540

01.

26.529

11.924

22

0.

228

10.

53

64.555

83.519

230

13.051

1.000

16.20

N1.80

04.550

18.003

10.310

17.062

1.

20.814

81.91

50.0

8

83.104

77.910

551

81.859

292

8.050

39.774

13.991

691

91.255

82.894

12.003

35.31

23.164

13.003

29.608

250

85.05

5° 100

d'

25.54

17.121

15

17:00

v.

14.595

N. 1999

6.

80.11

61.615

Nº.

09.683

72.732

82.001

70.706

76.693

8.202

59.122

19.50

29.131

S.

NB

32.17

76.198

32.

323

1.1.

80.00 B

17.653

51.019

Problem Statement



Banks offer loans to customers.



Customers either repay or the default on their loan.



Want to create a model to predict if a customer will default

Executive Summary



Banks want to better be able to **predict** which customer **will default** on their loan.



5960 loans were utilized to build a predictive model to determine if a customer would **default** on their loan.



12 variables observed in creating the model.

Findings – Indicators of Default



Debt to Income (DEBTINC) ratio was the number one indicator





Number of **Derogatory (DEROG)** Credit Reports: DEROG >= 7 All default



Number of **Delinquent (DELINQ)** Credit Reports: DELINQ >= 6 All default

Chosen Model



Hyper Tuned Random Forest Model

2 80% Accuracy

Maximizing the recall – Minimizes the number of loans given to those who may default



80% Recall



Same performance on Training and Test Set (F1- score =62%)

Business Implications



Risk Mitigation - Reduce the number of customers who default

Maintain Profitability- The bank will save money by having fewer defaulted loans

Bank will be able to identify the **characteristics** of the customer profile which **lead to default**



Customer Education - Opportunity for bank to provide financial education to their customer



Regulatory Compliance, Resource Allocation, Reputation Management, Credit Score Improvement, Maintain Competitive Advantage, Portfolio Diversity

Implementation



Ensure the Debt-to-Income ratio of potential customers is low



Ensure the customer has fewer than 6 derogatory reports (DEROG < 6)



Ensure the customer has fewer than 5 delinquent credit reports (DELINQ <5)

Take greater care when offering a loan to customer who is self-employed



Model integrated or stand alone, Training of loan officers, Continuous monitoring and maintenance, Feedback from loan officers





1189 (20%) customers defaulted (1) on their loan

Data Overview

Summary Statistics

	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
coun	t 5960.000000	5442.000000	5848.000000	5445.000000	5252.000000	5380.000000	5652.000000	5450.000000	5738.000000	4693.000000
mear	18607.969799	73760.817200	101776.048741	8.922268	0.254570	0.449442	179.766275	1.186055	21.296096	33.779915
sto	11207.480417	44457.609458	57385.775334	7.573982	0.846047	1.127266	85.810092	1.728675	10.138933	8.601746
mir	1100.000000	2063.000000	8000.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.524499
25%	11100.000000	46276.000000	66075.500000	3.000000	0.000000	0.000000	115.116702	0.000000	15.000000	29.140031
50%	16300.000000	65019.000000	89235.500000	7.000000	0.000000	0.000000	173.466667	1.000000	20.000000	34.818262
75%	23300.000000	91488.000000	119824.250000	13.000000	0.000000	0.000000	231.562278	2.000000	26.000000	39.003141
max	89900.000000	399550.000000	855909.000000	41.000000	10.000000	15.000000	1168.233561	17.000000	71.000000	203.312149

Exploratory Data Analysis

Visualized the data

Univariate and Bivariate analysis

Greater debt to income ratio **does** seem to result in more defaults



The larger the loan the lower number of derogatory records.



Data Cleaning:

Correcting outliers Imputing missing data

11 columns had missing values

of missing values ranged from 112 to 1267 Numeric values replaced with the **median** Categorical values replaced with the **mode**







All variables had outliers

Outliers < Q1 were replaced with Lower Whisker Outliers > Q3 were replaced with Upper Whisker



Created and tuned models

Decision Tree Random Forest



Model is attempting to **find** those **that will default** (1) on their loan, which will be our True Positive (TP), and therefore non-defaulters (0) will be our True Negative(TN).



Model_Name	Train_f1	Train_recall	Test_f1	Test_recall	Test_precision	
d_tree_base	100	100	60	56	64	
d_tree_base	68	79	67	75	60	
random_forest	100	100	69	59	85	
random_forest	100	100	68	55	87	
random_forest	62	83	62	80	51	

Variables Indicating Default - Models



Debt to Income (DEBTINC) ratio was the number one indicator **Credit Line Age (CLAGE)** was the second indicator

Variables indicating Default – from EDA

Number of Derogatory Credit Reports DEROG >= 7 All default Number of Delinquent Credit Reports DELINQ >= 6 All default





Results – Decision Tree Model

Results

Tuned Decision Tree:

Blue (1) – Defaulter, Orange (0) – Non-defaulter

• First split is made on DEBTINC (Debt-toincome ratio) - high ability to predict defaulter (Higher DEBTINC indicates the applicant is more likely to default)

This is consistent with our observations in EDA

Next high priority splits are

- NINQ(Number of recent credit inquiries) -Where a higher number is more likely to default
- CLAGE (Age of the oldest credit line in months) - Where a lower credit age is more likely to defualt



Results - Hyper Random Forest

Goal: Reduce the False Negatives Predict a person will not default and they default

- 1000 - 1200 Not Eligible Not Eligible 288.00 1376.00 40.00 1128.00 - 1000 - 800 - 800 Actual Actual - 600 - 600 - 400 Eligible Eligible - 400 153.00 219.00 74.00 298.00 - 200 - 200 Not Eligible Eligible Not Eligible Eligible Predicted Predicted

Reduced the number of False Negatives

Model Comparison – Goal is to **maximize** the **recall**

Tuned Decision Tree Model

Hyper Tuned Random Forest Model

- Accuracy of 85%,
- an F1-score of 67%,
- a recall of 75%

- Accuracy of 80%,
- an F1-score of 62%,
- a recall of 80%

Problem Statement

Using a classification model, the bank should be able to predict which features or combination of features of a customer would likely default on their loan.





Accuracy

- Given our focus on optimizing recall while maintaining a reasonable level of accuracy, the Hyper Tuned Random Forest Model emerges as the preferred choice. Its higher recall rate ensures that we minimize the chances of missing positive cases, aligning well with our objective. The slight decrease in F1-score from 67% to 62% is an acceptable trade-off in light of achieving a more robust recall rate.
- Considering the importance of capturing true positive cases, we can confidently recommend the **Hyper Tuned Random Forest Model** as the more suitable option for our scenario.



• Debt to income ratio is a very powerful characterisitic in predicting defaulters.

The bank can use debt to income ratio as an initial indicator when evaluating a loan.

• Business/Customer Focus

Make customers with a high dept to income ratio aware of potential difficulties of paying off a loan.

Offer financial counseling on how to lower their debt-to-income ratio to qualify for future loans.

 Credit Age is another characteristic in predicting defaulters.

The amount of time someone has had credit the better the bank can gauge how well they will repay their loans.

• Business/Customer Focus

Make customers aware of how their credit age is a benefit for them.

Offer secure credit cards or small loans to assist in building credit.



- Number of Derogatory and Delinquent Credit Reports is another tool in predicting defaulters.
- Persons with greater than or equal to 7 Derogatory reports ALL defaulted on their loans.
- 2. Persons with greater than or equal to 6 Delinquent credit reports ALL defaulted on their loans.
- Business/Customer Focus

Make customers aware of how derogatory and delinquent credit reports affect their ability to secure a loan.

Offer financial counseling on how to prevent having these reports in the future and how to improve their current credit reports



Overall - For Future Modeling



More data should be used to create a stronger model.



More care should be taken when gathering data.

There were errors/ missing data that could have assisted with improving the overall model.



DEBTINC had the most missing values and had some possible inaccuracies in the data. This feature was found to be the MOST important feature for predicting if a person would default.

Appendices – Imbalance /Class Distribution

#check for imbalance

bad_counts = trainingData['BAD'].value_counts()

print(bad_counts)

BAD

0 3339 1 832 Name: count, dtype: int64

check class distrubution
class_distribution = trainingData['BAD'].value_counts(normalize=True)

print(class_distribution)

BAD

0 0.800527
1 0.199473
Name: proportion, dtype: float64

Class Distribution

- There is a class bias
- To address class imbalance, we need to use techniques like oversampling the minority class, undersampling the majority class, or using synthetic data generation techniques (e.g., SMOTE).

