Exploratory Data Analysis

A Quick Overview of the EDA Framework on a sample Loan Defaulter Dataset

SMART SOLUTIONS

EDA Process

1. **Data Understanding ➔ Interpreting Data Dictionary**
2. **Setting an Objective for EDA (***Identify Target Label***)**
3. **Segregate Features by Data Types**
4. **Preprocessing**
    a) **Handling Null Values, Redundant Features**
    b) **Feature Generation through Feature Engineering**
5. **For the set of Features for each Data Type, perform suitable EDA (***1 Feature Vs. Target Label***):**
    a) **Correlation (***Detect Multi-Collinearity among Continuous Features***)**
    b) **Biserial Correlation (***Detect Association between the Target Label and a Continuous Feature***)**
    c) **Histogram (***Identify Influential Level from a Non Boolean Categorical Feature against the Target Label***)**
    d) **Quadrant Analysis (***Detect Association between the Target Label and a Boolean Feature***)**
6. **Selected Features**
7. **Insights through Unsupervised Learning (***Dimensionality Reduction, Anomaly Detection & Clustering***)**
8. **Dimensionality Reduction on Selected Features (***Reduce the size of dataset while retaining key information***)**
9. **Anomaly Detection (***Identify a minority set of data points that behaves differently from the majority***)**
10. **Clustering using DBSCAN & HDBSCAN (***Identify generated clusters that have higher than average defaulter rate***)**
11. **Final Feature Set (***To input into supervised learning models to predict defaulters***)**

SMART SOLUTIONS

1. Meaning of an Observation

- **An observation refers a unique customer (*Identified via "Unique_ID" feature*)**
- **Each other customer is characterized by 41 features (*columns*)**
- **There are in total 233,154 customers (*rows*) in the dataset**

1. Interpreting The Features

Feature No.	Feature Name	Description
1	UniqueID	Identifier for customers
2	loan_default	Payment default in the first EMI on due date
3	disbursed_amount	Amount of Loan disbursed
4	asset_cost	Cost of the Asset
5	ltv	Loan to Value of the asset
6	branch_id	Branch where the loan was disbursed
7	supplier_id	Vehicle Dealer where the loan was disbursed
8	manufacturer_id	Vehicle manufacturer(Hero, Honda, TVS etc.)
9	Current_pincode ID	Current pincode of the customer
10	Date.of.Birth	Date of birth of the customer
11	Employment.Type	Employment Type of the customer (Salaried/Self Employed)
12	DisbursalDate	Date of disbursement
13	State_ID	State of disbursement
14	Employee_code_ID	Employee of the organization who logged the disbursement
15	MobileNo_Avl_Flag	if Mobile no. was shared by the customer then flagged as 1
16	Aadhar_flag	if aadhar was shared by the customer then flagged as 1
17	PAN_flag	if pan was shared by the customer then flagged as 1
18	VoterID_flag	if voter was shared by the customer then flagged as 1
19	Driving_flag	if DL was shared by the customer then flagged as 1
20	Passport_flag	if passport was shared by the customer then flagged as 1
21	PERFORM_CNS.SCORE	Bureau Score
22	PERFORM_CNS.SCORE.DESCRIPTION	Bureau score description

SMART SOLUTIONS

1. Interpreting The Features

Feature No.	Feature Name	Description	
23	PRI.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement	Primary accounts are those which the customer has taken for his personal use
24	PRI.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement	
25	PRI.OVERDUE.ACCTS	count of default accounts at the time of disbursement	
26	PRI.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement	
27	PRI.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement	
28	PRI.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement	
29	SEC.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement	Secondary accounts are those which the customer act as a co-applicant or gaurantor
30	SEC.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement	
31	SEC.OVERDUE.ACCTS	count of default accounts at the time of disbursement	
32	SEC.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement	
33	SEC.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement	
34	SEC.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement	
35	PRIMARY.INSTAL.AMT	EMI Amount of the primary loan	
36	SEC.INSTAL.AMT	EMI Amount of the secondary loan	
37	NEW.ACCTS.IN.LAST.SIX.MONTHS	New loans taken by the customer in last 6 months before the disbursment	
38	DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	Loans defaulted in the last 6 months	
39	AVERAGE.ACCT.AGE	Average loan tenure	
40	CREDIT.HISTORY.LENGTH	Time since first loan	
41	NO.OF_INQUIRIES	Enquries done by the customer for loans	

SMART SOLUTIONS

2. Set an Objective for EDA

“Identify a set of Features, which are approximately independent of one another, explains significantly the differences between an individual who defaulted a loan and an individual who did not default a loan”

Hence, “**Loan_Default**” Feature will be the **target label**, as signal to which features are significant.

3. Segregate Features by Data Types

Continuous Features (# : 21)
disbursed_amount
asset_cost
ltv
PERFORM_CNS.SCORE
PRI.NO.OF.ACCTS
PRI.ACTIVE.ACCTS
PRI.OVERDUE.ACCTS
PRI.CURRENT.BALANCE
PRI.SANCTIONED.AMOUNT
PRI.DISBURSED.AMOUNT
SEC.NO.OF.ACCTS
SEC.ACTIVE.ACCTS
SEC.OVERDUE.ACCTS
SEC.CURRENT.BALANCE
SEC.SANCTIONED.AMOUNT
SEC.DISBURSED.AMOUNT
PRIMARY.INSTAL.AMT
SEC.INSTAL.AMT
NEW.ACCTS.IN.LAST.SIX.MONTHS
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS
NO.OF_INQUIRIES

Non Boolean Categorical Feature (# : 7) Type: Identifiers
UniqueID
branch_id
supplier_id
manufacturer_id
Current_pincode_ID
State_ID
Employee_code_ID

Non Boolean Categorical Feature (# : 6) Type: Demographics
Date.of.Birth
Employment.Type
DisbursalDate
PERFORM_CNS.SCORE.DESCRIPTION
AVERAGE.ACCT.AGE
CREDIT.HISTORY.LENGTH

Boolean Categorical Feature (# : 7)
MobileNo_Avl_Flag
Aadhar_flag
PAN_flag
VoterID_flag
Driving_flag
Passport_flag
loan_default (Target Label)

SMART SOLUTIONS

4.a) Replace Null Values

```
df.isnull().sum()
```

```
UniqueID                              0
disbursed_amount                      0
asset_cost                            0
ltv                                   0
branch_id                             0
supplier_id                           0
manufacturer_id                       0
Current_pincode_ID                    0
Date.of.Birth                         0
Employment.Type                    7661
DisbursalDate                         0
State_ID                              0
Employee_code_ID                      0
MobileNo_Avl_Flag                     0
Aadhar_flag                           0
PAN_flag                              0
VoterID_flag                          0
Driving_flag                          0
Passport_flag                         0
PERFORM_CNS.SCORE                     0
PERFORM_CNS.SCORE.DESCRIPTION         0
PRI.NO.OF.ACCTS                       0
PRI.ACTIVE.ACCTS                      0
PRI.OVERDUE.ACCTS                     0
PRI.CURRENT.BALANCE                   0
PRI.SANCTIONED.AMOUNT                 0
PRI.DISBURSED.AMOUNT                  0
SEC.NO.OF.ACCTS                       0
SEC.ACTIVE.ACCTS                      0
SEC.OVERDUE.ACCTS                     0
SEC.CURRENT.BALANCE                   0
SEC.SANCTIONED.AMOUNT                 0
SEC.DISBURSED.AMOUNT                  0
PRIMARY.INSTAL.AMT                    0
SEC.INSTAL.AMT                        0
NEW.ACCTS.IN.LAST.SIX.MONTHS          0
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS   0
AVERAGE.ACCT.AGE                      0
CREDIT.HISTORY.LENGTH                 0
NO.OF_INQUIRIES                       0
loan_default                          0
dtype: int64
```

Before

```
Self employed    127635
Salaried          97858
Name: Employment.Type, dtype: int64
```

After

```
Self employed    127635
Salaried          97858
No Record          7661
Name: Employment.Type, dtype: int64
```

SMART SOLUTIONS

4.a) Remove Redundant Features

	count	mean	std	min	25%	50%	75%	max
MobileNo_Avl_Flag	233154.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0

‘MobileNo_Avl_Flag’ feature is reduntant as it has standard deviation of 0.

This means every customer has shared their Mobile No.

Hence, it should be removed from the analysis.

4.b) Feature Engineering

#new_feature: age (in years) = year(disbursal - dob)

	DisbursalDate	Date.of.Birth	age
0	2018-03-08	1984-01-01	34
1	2018-09-26	1985-07-31	33
2	2018-01-08	1985-08-24	32
3	2018-10-26	1993-12-30	25
4	2018-09-26	1977-09-12	41
5	2018-09-19	1990-08-09	28
6	2018-09-23	1988-01-06	31
7	2018-09-16	1989-04-10	29
8	2018-05-09	1991-11-15	26
9	2018-09-16	2068-01-06	0

Negative age are converted to 0

4.b) Feature Engineering

#feature_formatting: changed dtype from ‘object’ to ‘int64’

- AVERAGE.ACCT.AGE (in Months)
- CREDIT.HISTORY.LENGTH (in Months)

```
AVERAGE.ACCT.AGE                object
CREDIT.HISTORY.LENGTH           object
```

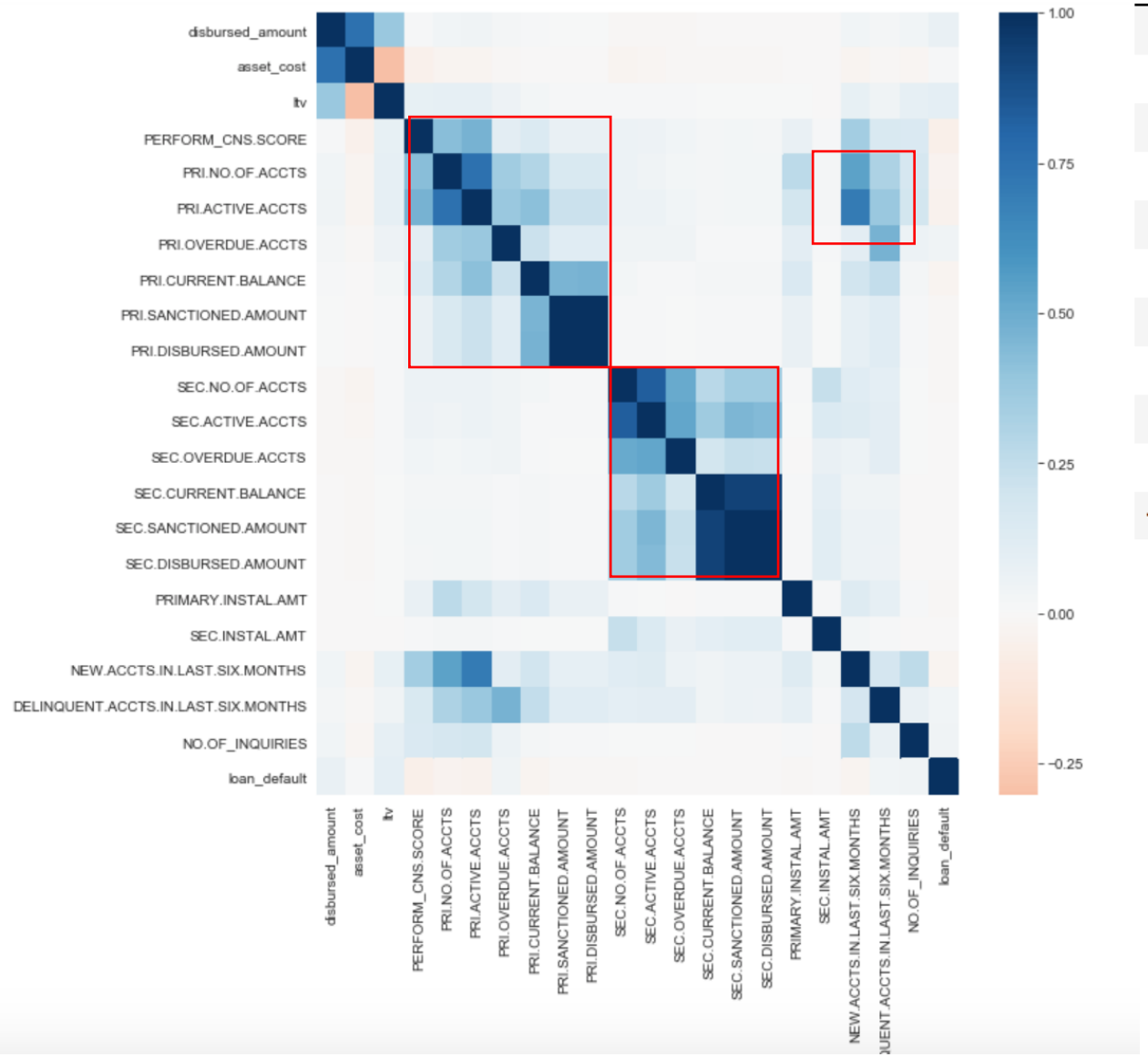
```
AVERAGE.ACCT.AGE                int64
CREDIT.HISTORY.LENGTH           int64
```

AVERAGE.ACCT.AGE	CREDIT.HISTORY.LENGTH
0yrs 0mon	0yrs 0mon
1yrs 11mon	1yrs 11mon
0yrs 0mon	0yrs 0mon
0yrs 8mon	1yrs 3mon
0yrs 0mon	0yrs 0mon

	AVERAGE.ACCT.AGE	CREDIT.HISTORY.LENGTH
0	0	0
1	23	23
2	0	0
3	8	15
4	0	0

SMART SOLUTIONS

5.a) Correlation

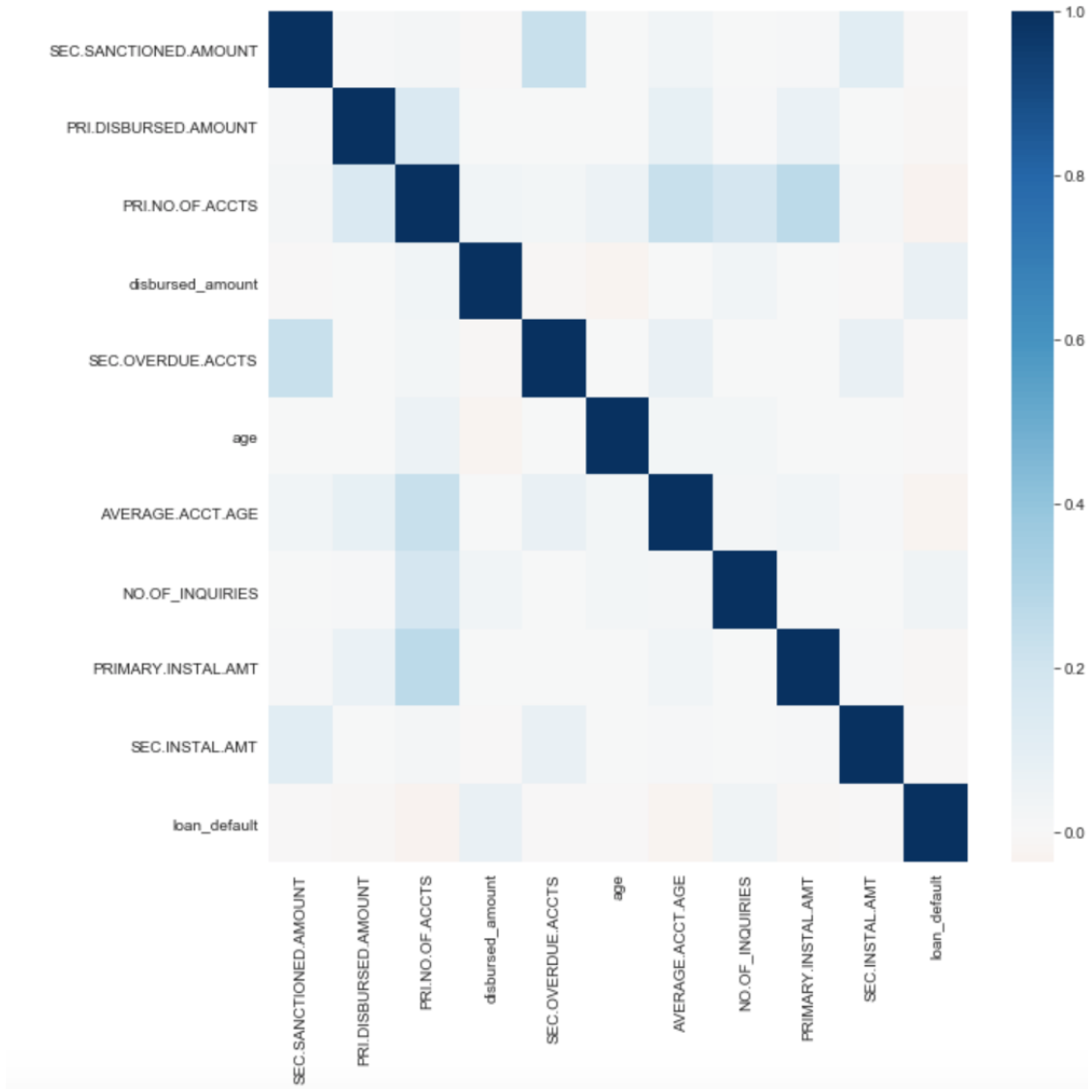


	var1	var2	correl
0	SEC.SANCTIONED.AMOUNT	SEC.DISBURSED.AMOUNT	0.999646
1	PRI.DISBURSED.AMOUNT	PRI.SANCTIONED.AMOUNT	0.998747
2	SEC.CURRENT.BALANCE	SEC.DISBURSED.AMOUNT	0.929995
3	SEC.CURRENT.BALANCE	SEC.SANCTIONED.AMOUNT	0.929196
4	SEC.NO.OF.ACCTS	SEC.ACTIVE.ACCTS	0.828727
5	PRI.NO.OF.ACCTS	PRI.ACTIVE.ACCTS	0.753565
6	disbursed_amount	asset_cost	0.752668
7	PRI.ACTIVE.ACCTS	NEW.ACCTS.IN.LAST.SIX.MONTHS	0.702943
8	NEW.ACCTS.IN.LAST.SIX.MONTHS	PRI.NO.OF.ACCTS	0.537622
9	SEC.OVERDUE.ACCTS	SEC.ACTIVE.ACCTS	0.526209
10	SEC.OVERDUE.ACCTS	SEC.NO.OF.ACCTS	0.510394

High Multi-Collinearity detected among continuous features

SMART SOLUTIONS

5.a) Correlation



Selected Continuous Features
SEC.SANCTIONED.AMOUNT
PRI.DISBURSED.AMOUNT
PRI.NO.OF.ACCTS
disbursed_amount
SEC.OVERDUE.ACCTS
age
AVERAGE.ACCT.AGE
NO.OF_INQUIRIES
PRIMARY.INSTAL.AMT
SEC.INSTAL.AMT

Selected a set of features which are approximately **independent** of one another (-0.3 <Correlation<0.3)

SMART SOLUTIONS

5.b) Biserial Correlation

	count	mean	std	min	25%	50%	75%	max
SEC.SANCTIONED.AMOUNT	233154.0	7295.923347	1.831560e+05	0.0	0.0	0.0	0.0	3.000000e+07
PRI.DISBURSED.AMOUNT	233154.0	218065.898655	2.377744e+06	0.0	0.0	0.0	60800.0	1.000000e+09
PRI.NO.OF.ACCTS	233154.0	2.440636	5.217233e+00	0.0	0.0	0.0	3.0	4.530000e+02
disbursed_amount	233154.0	54356.993528	1.297131e+04	13320.0	47145.0	53803.0	60413.0	9.905720e+05
SEC.OVERDUE.ACCTS	233154.0	0.007244	1.110789e-01	0.0	0.0	0.0	0.0	8.000000e+00
age	233154.0	29.866908	1.191642e+01	0.0	24.0	30.0	38.0	5.000000e+01
AVERAGE.ACCT.AGE	233154.0	8.915764	1.510642e+01	0.0	0.0	0.0	13.0	3.690000e+02
NO.OF_INQUIRIES	233154.0	0.206615	7.064977e-01	0.0	0.0	0.0	0.0	3.600000e+01
PRIMARY.INSTAL.AMT	233154.0	13105.481720	1.513679e+05	0.0	0.0	0.0	1999.0	2.564281e+07
SEC.INSTAL.AMT	233154.0	323.268449	1.555369e+04	0.0	0.0	0.0	0.0	4.170901e+06
loan_default	233154.0	0.217071	4.122523e-01	0.0	0.0	0.0	0.0	1.000000e+00

5.b) Biserial Correlation

	count	mean	std	min	25%	50%	75%	max
SEC.SANCTIONED.AMOUNT	233154.0	7295.923347	1.831560e+05	0.0	0.0	0.0	0.0	3.000000e+07
PRI.DISBURSED.AMOUNT	233154.0	218065.898655	2.377744e+06	0.0	0.0	0.0	60800.0	1.000000e+09
PRI.NO.OF.ACCTS	233154.0	2.440636	5.217233e+00	0.0	0.0	0.0	3.0	4.530000e+02
disbursed_amount	233154.0	54356.993528	1.297131e+04	13320.0	47145.0	53803.0	60413.0	9.905720e+05
SEC.OVERDUE.ACCTS	233154.0	0.007244	1.110789e-01	0.0	0.0	0.0	0.0	8.000000e+00
age	233154.0	29.866908	1.191642e+01	0.0	24.0	30.0	38.0	5.000000e+01
AVERAGE.ACCT.AGE	233154.0	8.915764	1.510642e+01	0.0	0.0	0.0	13.0	3.690000e+02
NO.OF_INQUIRIES	233154.0	0.206615	7.064977e-01	0.0	0.0	0.0	0.0	3.600000e+01
PRIMARY.INSTAL.AMT	233154.0	13105.481720	1.513679e+05	0.0	0.0	0.0	1999.0	2.564281e+07
SEC.INSTAL.AMT	233154.0	323.268449	1.555369e+04	0.0	0.0	0.0	0.0	4.170901e+06
loan_default	233154.0	0.217071	4.122523e-01	0.0	0.0	0.0	0.0	1.000000e+00

Selected features from 5.a)

SMART SOLUTIONS

5.b) Biserial Correlation

	count	mean	std	min	25%	50%	75%	max
SEC.SANCTIONED.AMOUNT	233154.0	7295.923347	1.831560e+05	0.0	0.0	0.0	0.0	3.000000e+07
PRI.DISBURSED.AMOUNT	233154.0	218065.898655	2.377744e+06	0.0	0.0	0.0	60800.0	1.000000e+09
PRI.NO.OF.ACCTS	233154.0	2.440636	5.217233e+00	0.0	0.0	0.0	3.0	4.530000e+02
disbursed_amount	233154.0	54356.993528	1.297131e+04	13320.0	47145.0	53803.0	60413.0	9.905720e+05
SEC.OVERDUE.ACCTS	233154.0	0.007244	1.110789e-01	0.0	0.0	0.0	0.0	8.000000e+00
age	233154.0	29.866908	1.191642e+01	0.0	24.0	30.0	38.0	5.000000e+01
AVERAGE.ACCT.AGE	233154.0	8.915764	1.510642e+01	0.0	0.0	0.0	13.0	3.690000e+02
NO.OF_INQUIRIES	233154.0	0.206615	7.064977e-01	0.0	0.0	0.0	0.0	3.600000e+01
PRIMARY.INSTAL.AMT	233154.0	13105.481720	1.513679e+05	0.0	0.0	0.0	1999.0	2.564281e+07
SEC.INSTAL.AMT	233154.0	323.268449	1.555369e+04	0.0	0.0	0.0	0.0	4.170901e+06
loan_default	233154.0	0.217071	4.122523e-01	0.0	0.0	0.0	0.0	1.000000e+00

Statistics

- **Count**: Total # records
- **Mean**: Sum(values) in records /Count
- **Std**: Standard Deviation, the average distance between values in records and the Mean

- **Min**: Minimum value found across all records
- **25%**: 25th Percentile
- **50%**: 50th Percentile (Median)
- **75%**: 75th Percentile
- **Max**: Maximum value found across all records

SMART SOLUTIONS

5.b) Biserial Correlation

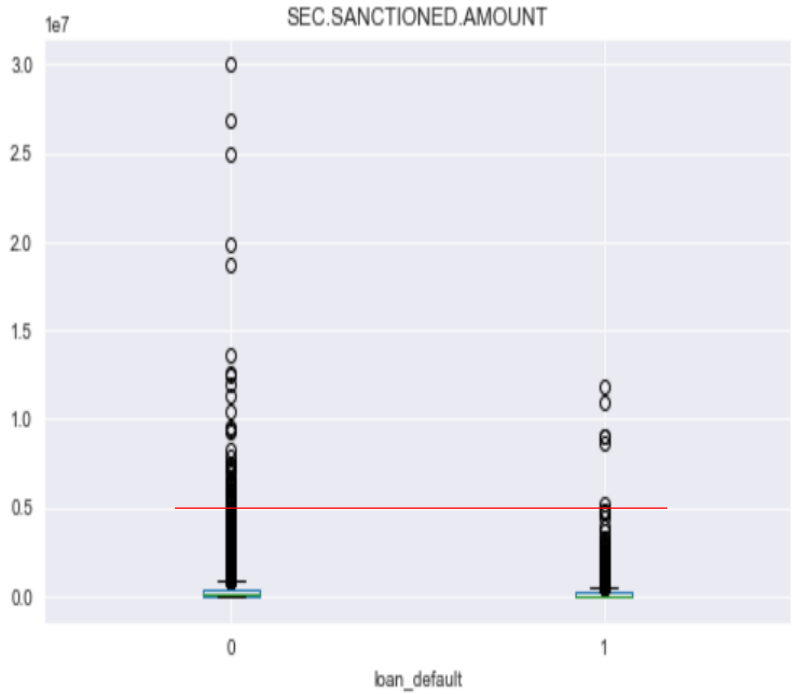
	count	mean	std	min	25%	50%	75%	max
SEC.SANCTIONED.AMOUNT	233154.0	7295.923347	1.831560e+05	0.0	0.0	0.0	0.0	3.000000e+07
PRI.DISBURSED.AMOUNT	233154.0	218065.898655	2.377744e+06	0.0	0.0	0.0	60800.0	1.000000e+09
PRI.NO.OF.ACCTS	233154.0	2.440636	5.217233e+00	0.0	0.0	0.0	3.0	4.530000e+02
disbursed_amount	233154.0	54356.993528	1.297131e+04	13320.0	47145.0	53803.0	60413.0	9.905720e+05
SEC.OVERDUE.ACCTS	233154.0	0.007244	1.110789e-01	0.0	0.0	0.0	0.0	8.000000e+00
age	233154.0	29.866908	1.191642e+01	0.0	24.0	30.0	38.0	5.000000e+01
AVERAGE.ACCT.AGE	233154.0	8.915764	1.510642e+01	0.0	0.0	0.0	13.0	3.690000e+02
NO.OF_INQUIRIES	233154.0	0.206615	7.064977e-01	0.0	0.0	0.0	0.0	3.600000e+01
PRIMARY.INSTAL.AMT	233154.0	13105.481720	1.513679e+05	0.0	0.0	0.0	1999.0	2.564281e+07
SEC.INSTAL.AMT	233154.0	323.268449	1.555369e+04	0.0	0.0	0.0	0.0	4.170901e+06
loan_default	233154.0	0.217071	4.122523e-01	0.0	0.0	0.0	0.0	1.000000e+00

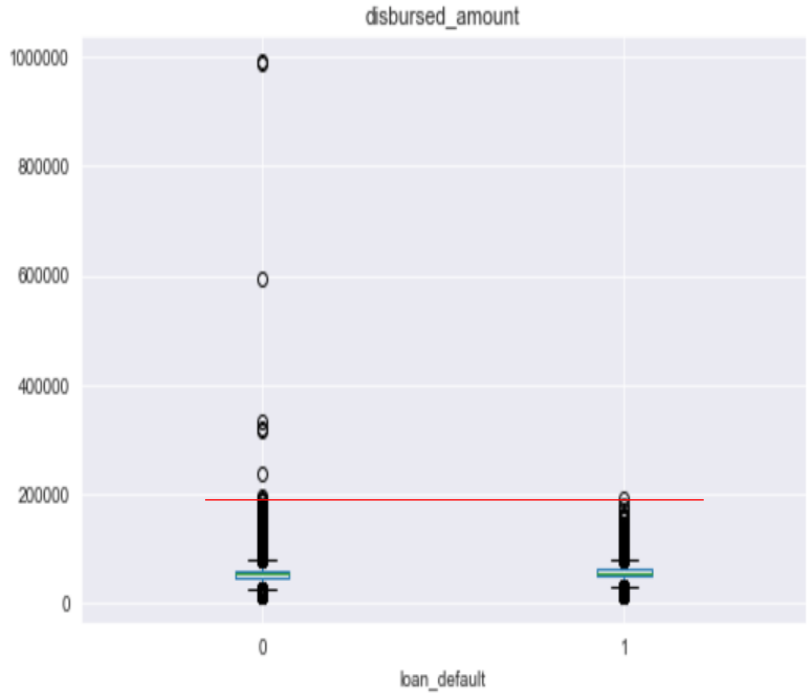
Most features have high occurrence of values close to 0 (red) with high standard deviations (green)

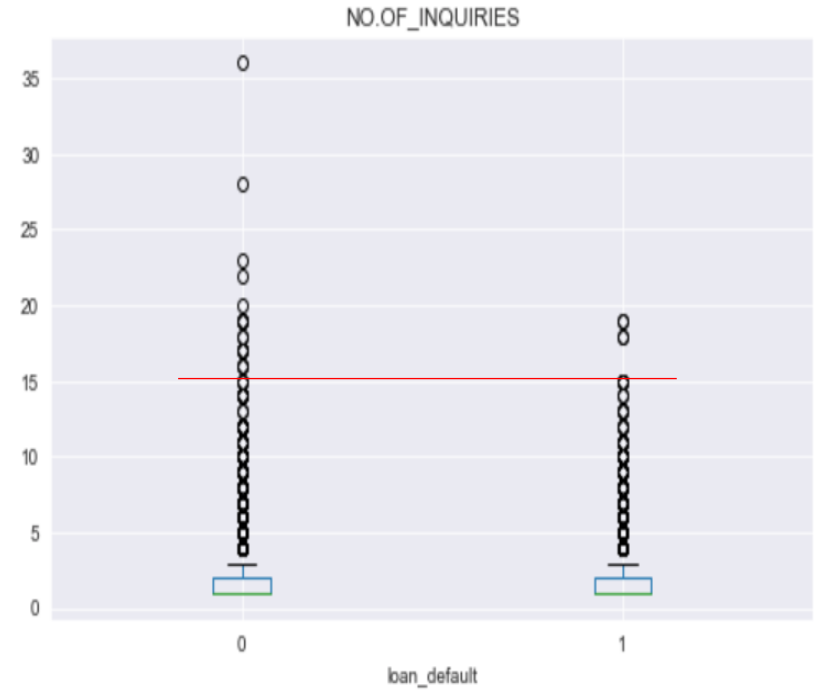
➔ Remove 0 values from features

SMART SOLUTIONS

5.b) Biserial Correlation





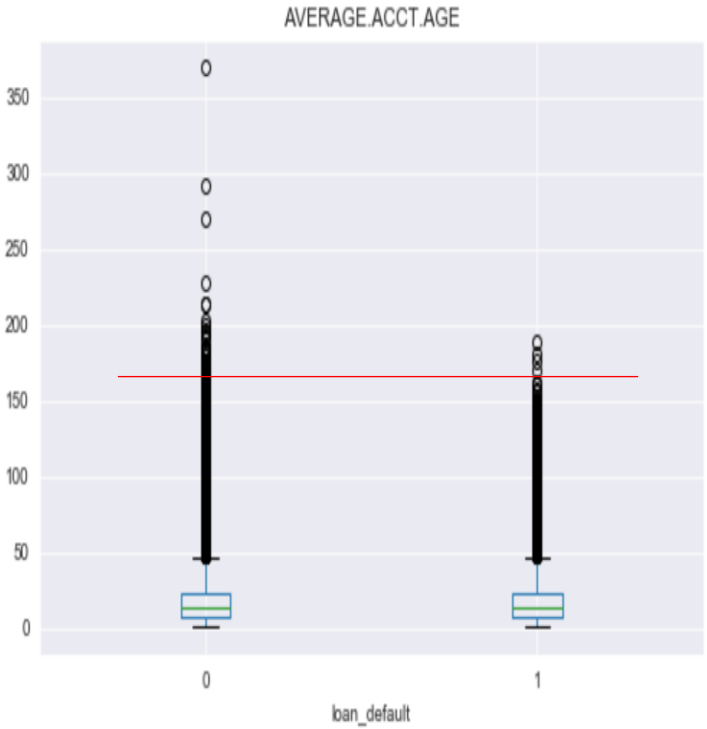


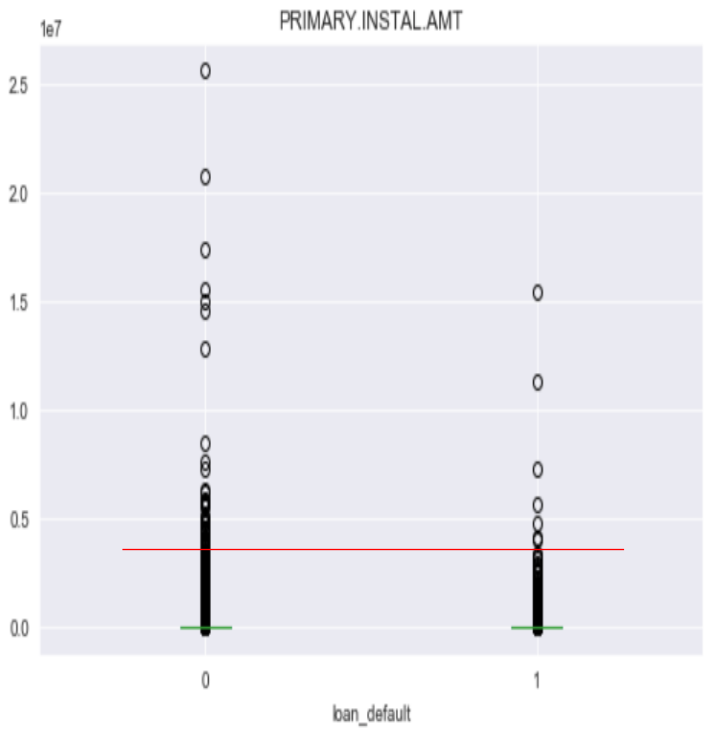
For above features, distinction is observed for data above 75th percentile

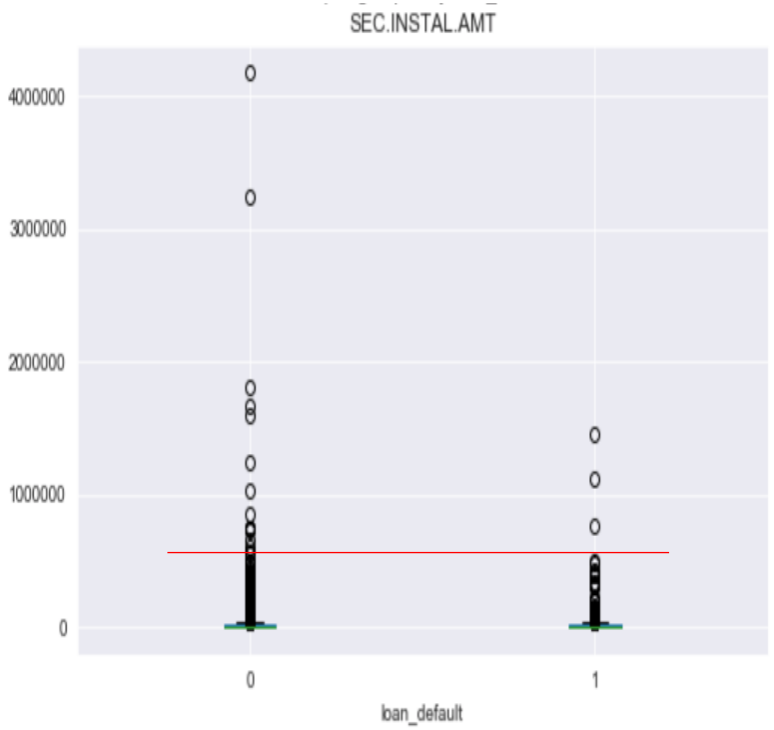
➔ Discover upper bound for defaulters

SMART SOLUTIONS

5.b) Biserial Correlation





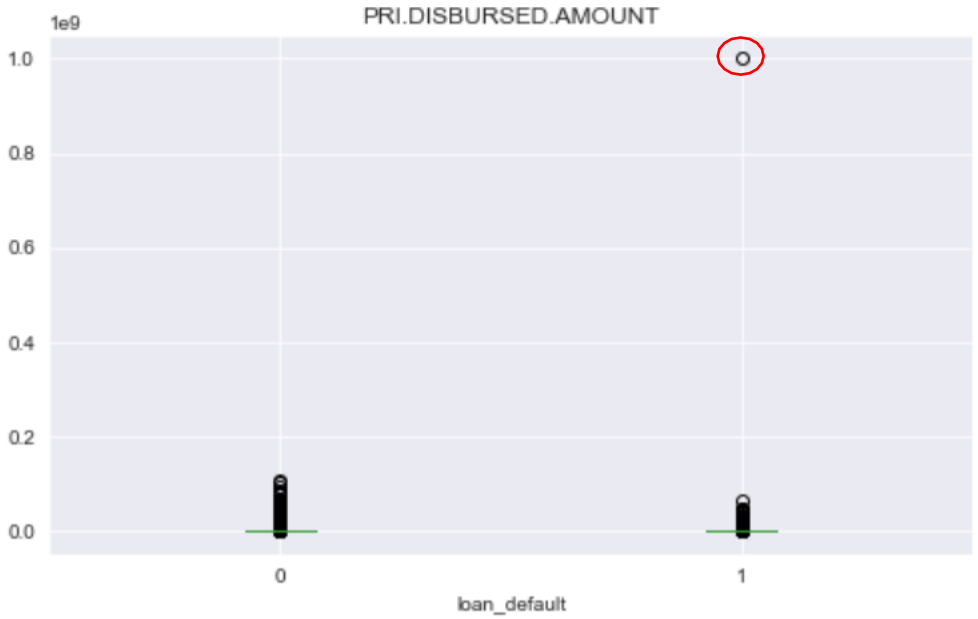


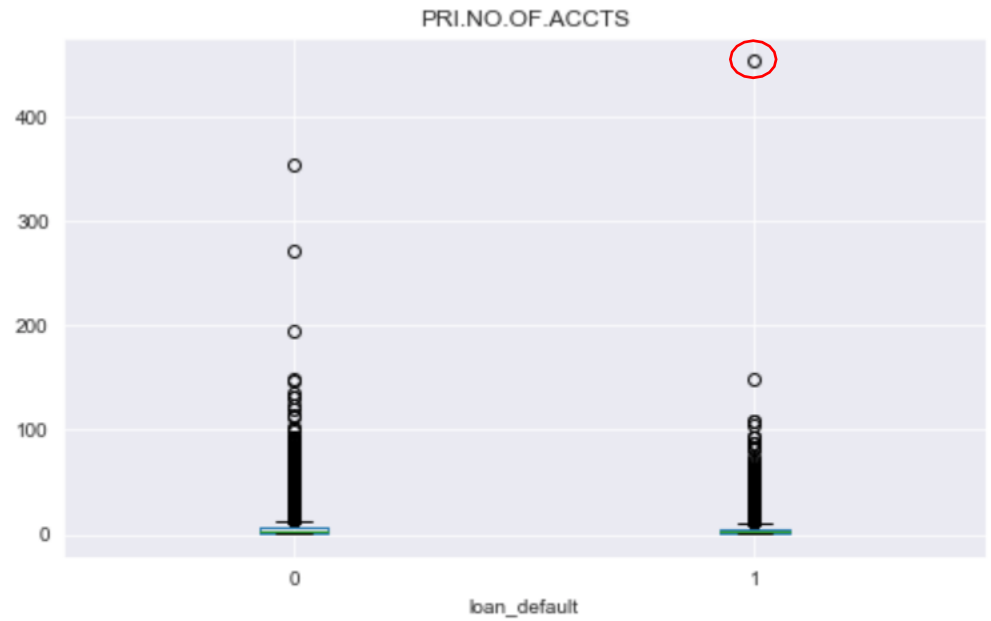
For above features, distinction is observed for data above 75th percentile

➔ Discover upper bound for defaulters

SMART SOLUTIONS

5.b) Biserial Correlation



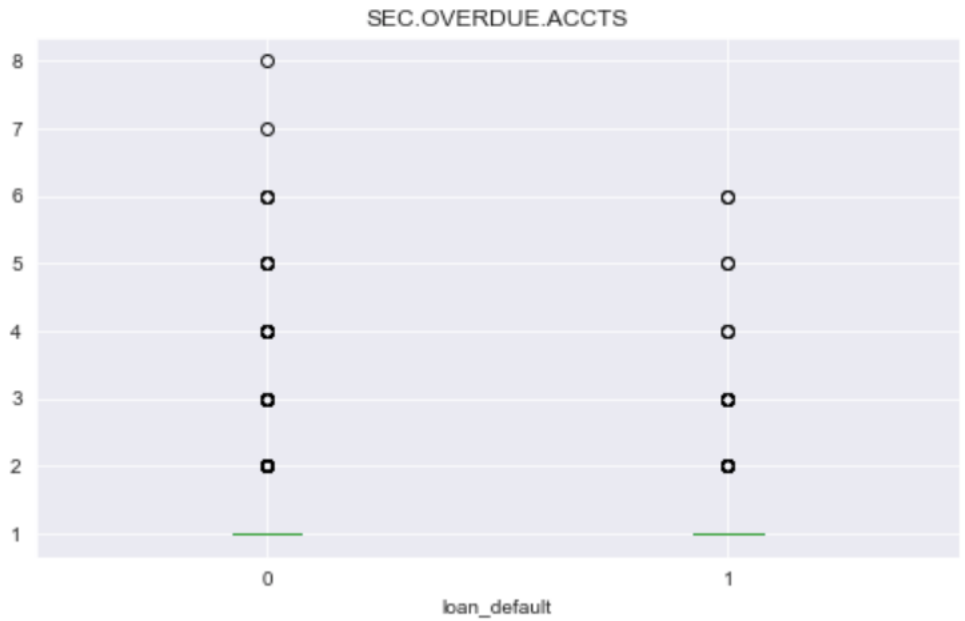


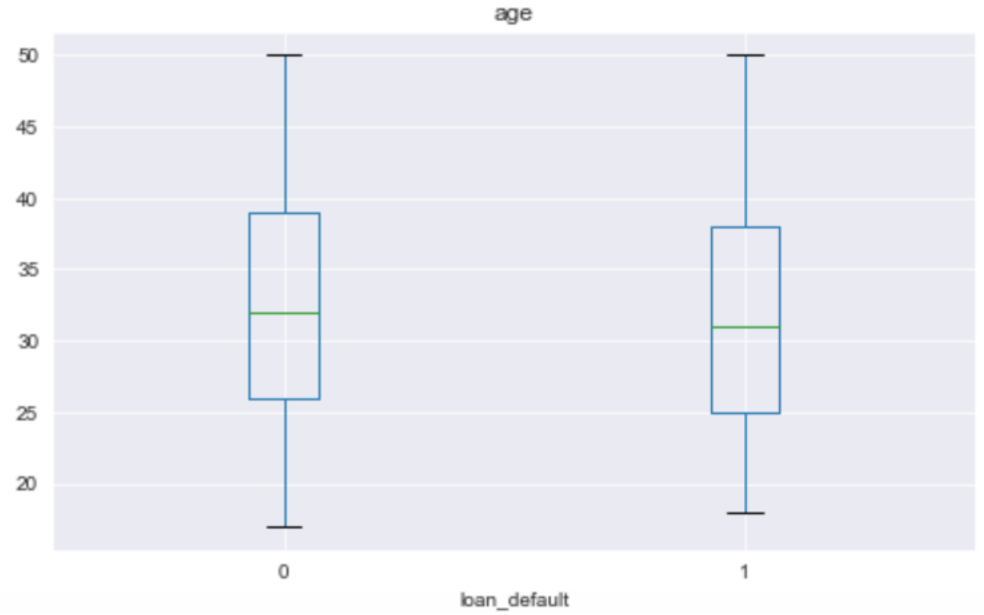
For above features, distinction is observed for data above 75th percentile

➔ Extreme outliers detected for defaulters

SMART SOLUTIONS

5.b) Biserial Correlation





For above features, no clear distinctions are observed

➔ To be removed from analysis

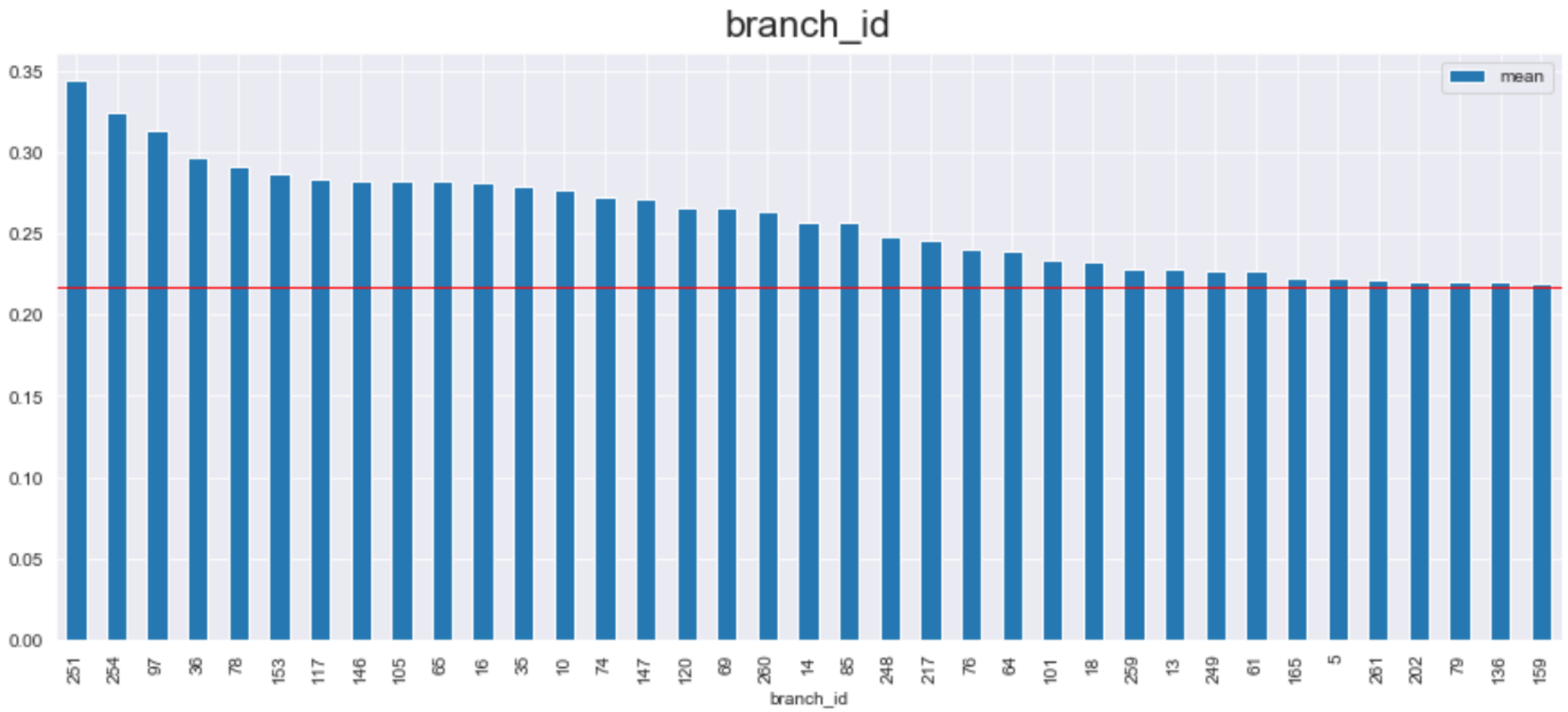
SMART SOLUTIONS

5.c) Histogram

21.7% is the defaulter rate in the entire dataset.

- The value is generated by taking mean of the Target Label, “loan_default”
- In the histogram,
  - **Y-axis**: Defaulter rate
  - **X-axis**: Levels of a categorical feature
  - **Red Horizontal Line**: Threshold set at **21.7%**
- Through histogram, we can capture the set of levels that have significantly high defaulter rate (i.e. above red horizontal line)
  ➔ Only these levels will be kept for analysis

5.c) Histogram

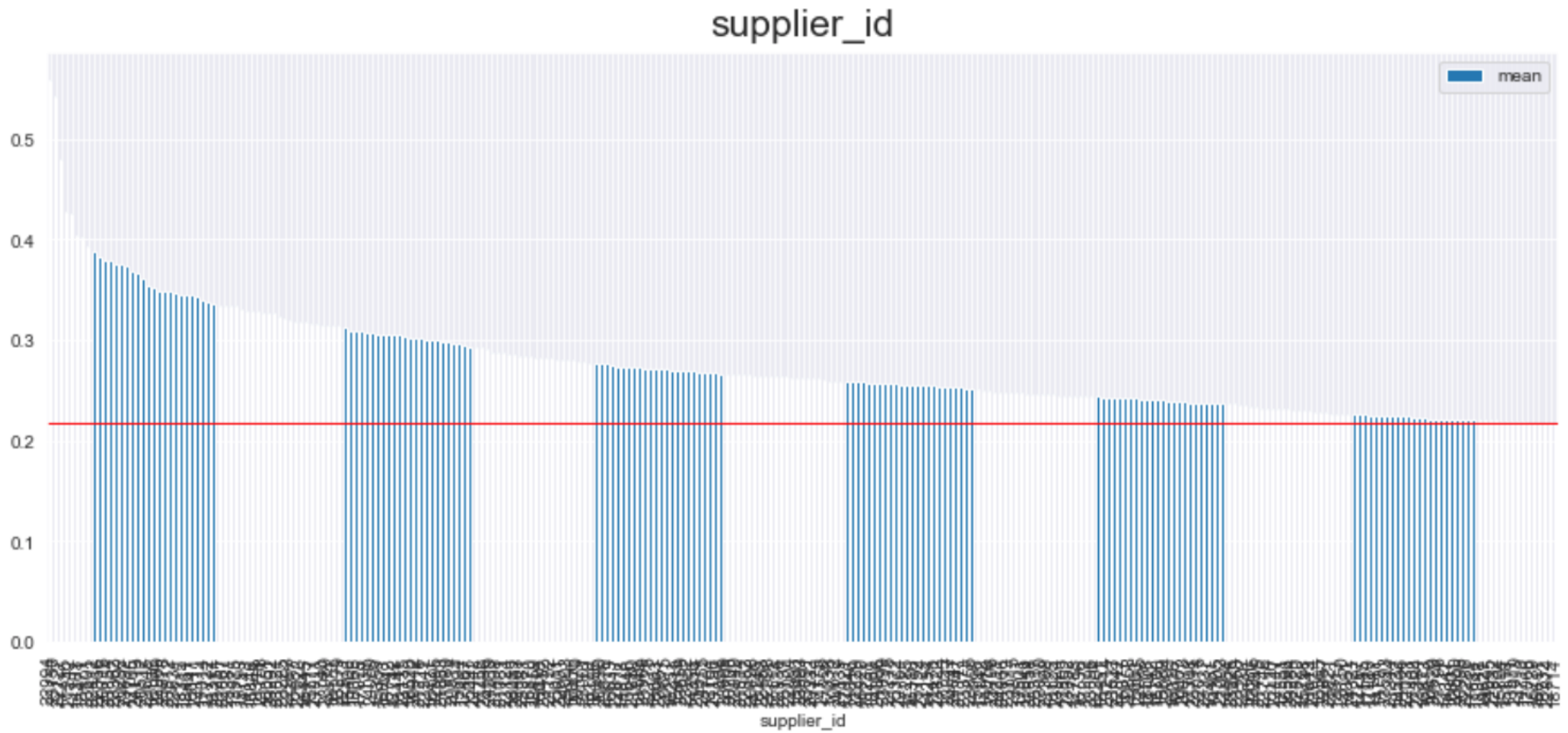


Levels above the threshold: 37

Level with the highest defaulter rate: branch_id 251 (~34%)

SMART SOLUTIONS

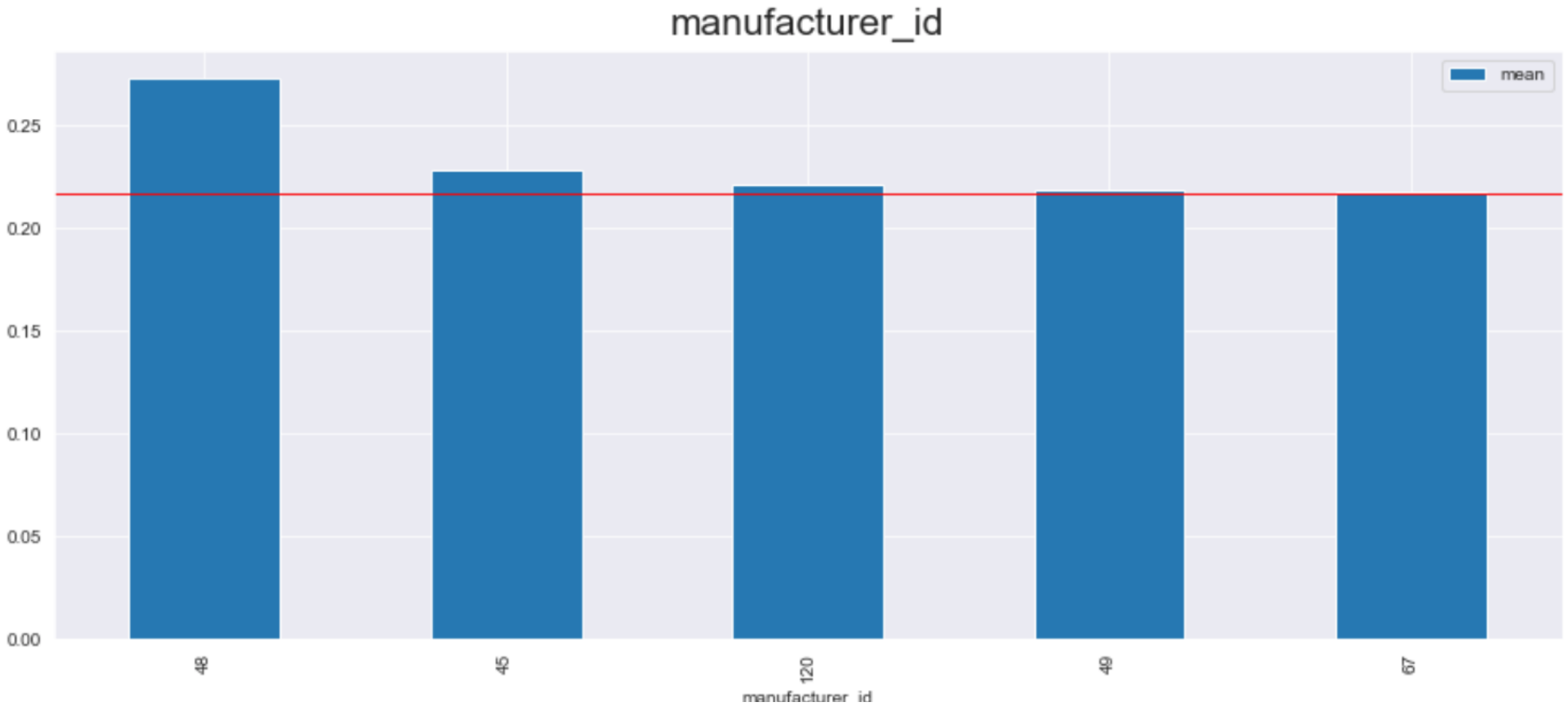
5.c) Histogram



Levels above the threshold: 277

Level with the highest defaulter rate: supplier_id 22994 (~56%)

5.c) Histogram

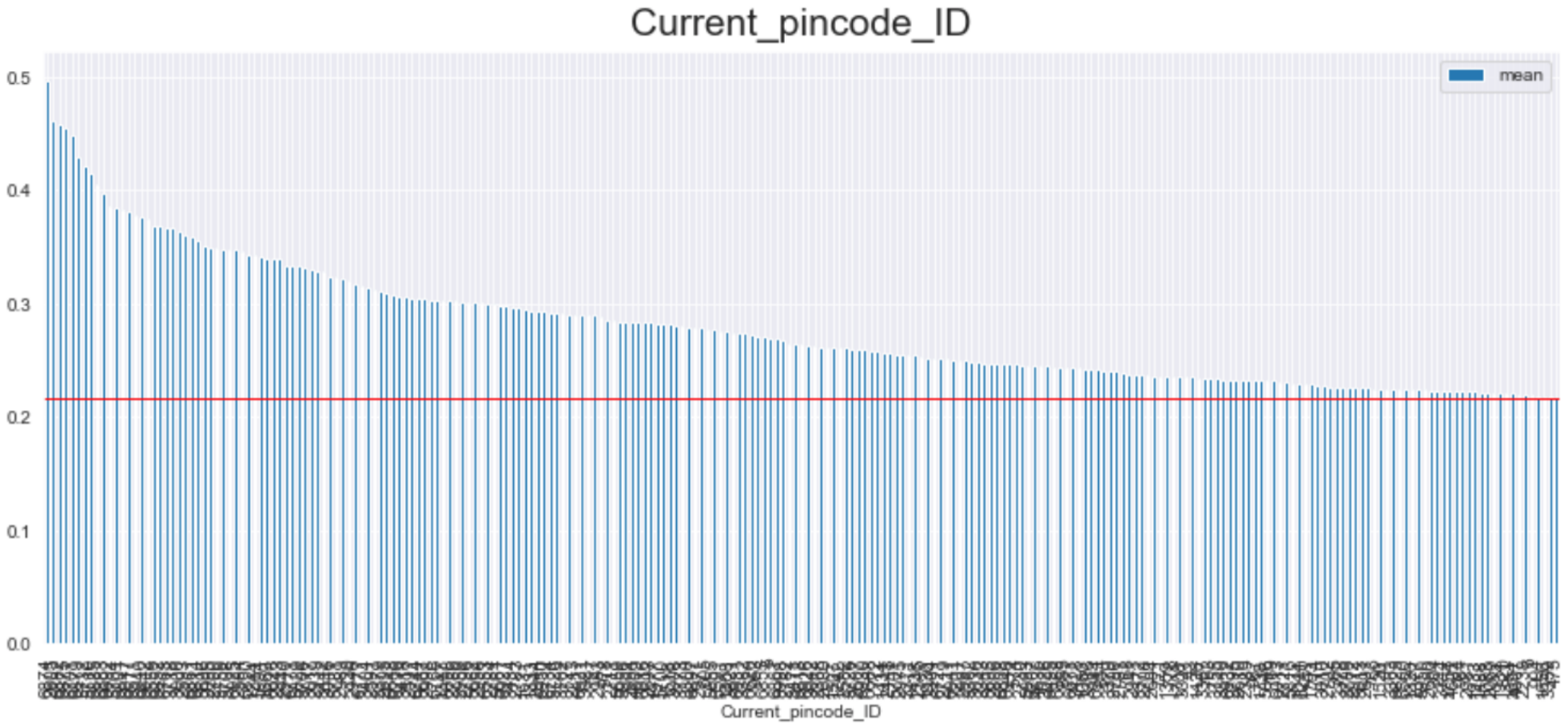


Levels above the threshold: 5

Level with the highest defaulter rate: manufacturer_id 48 (~27%)

SMART SOLUTIONS

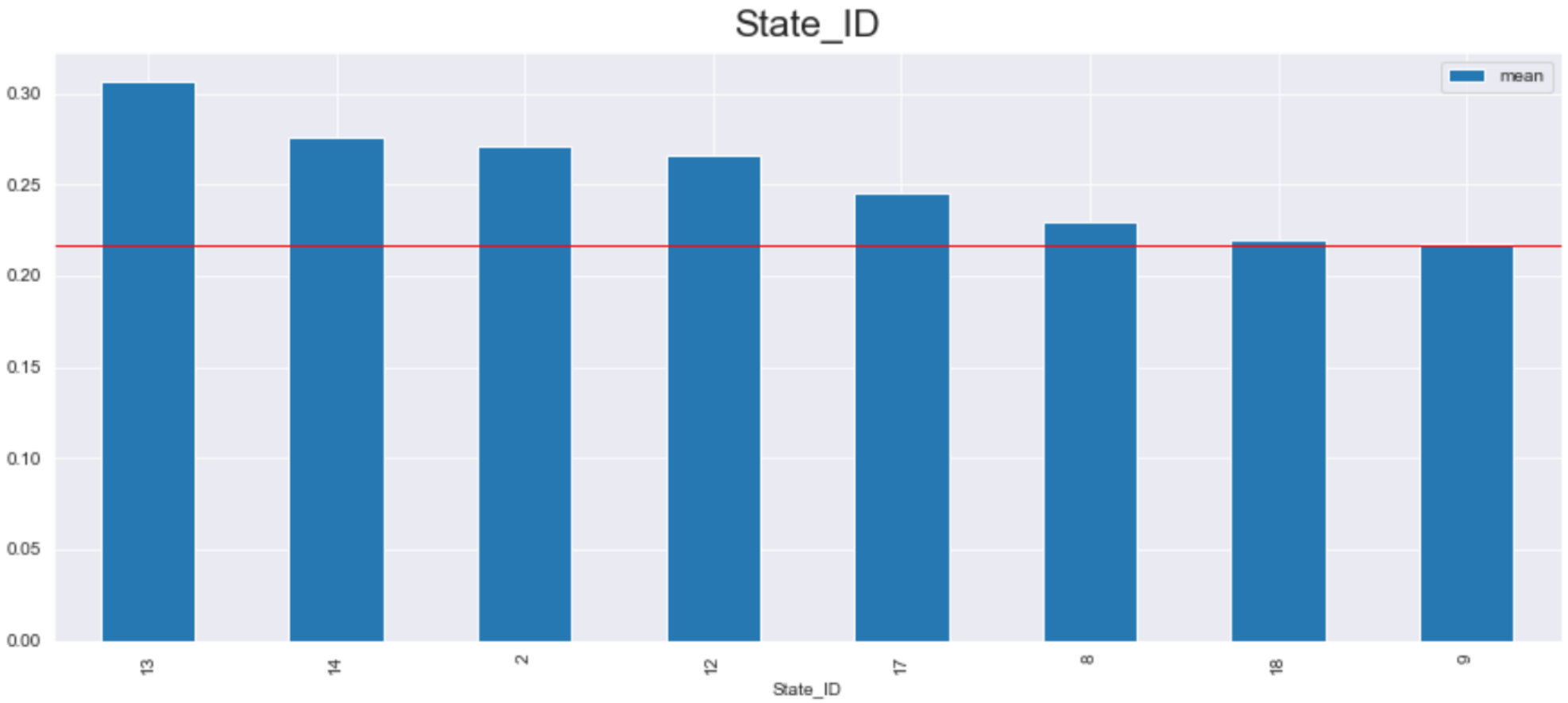
5.c) Histogram



Levels above the threshold: 241

Level with the highest defaulter rate: Current_pincode_id 6874 (~50%)

5.c) Histogram

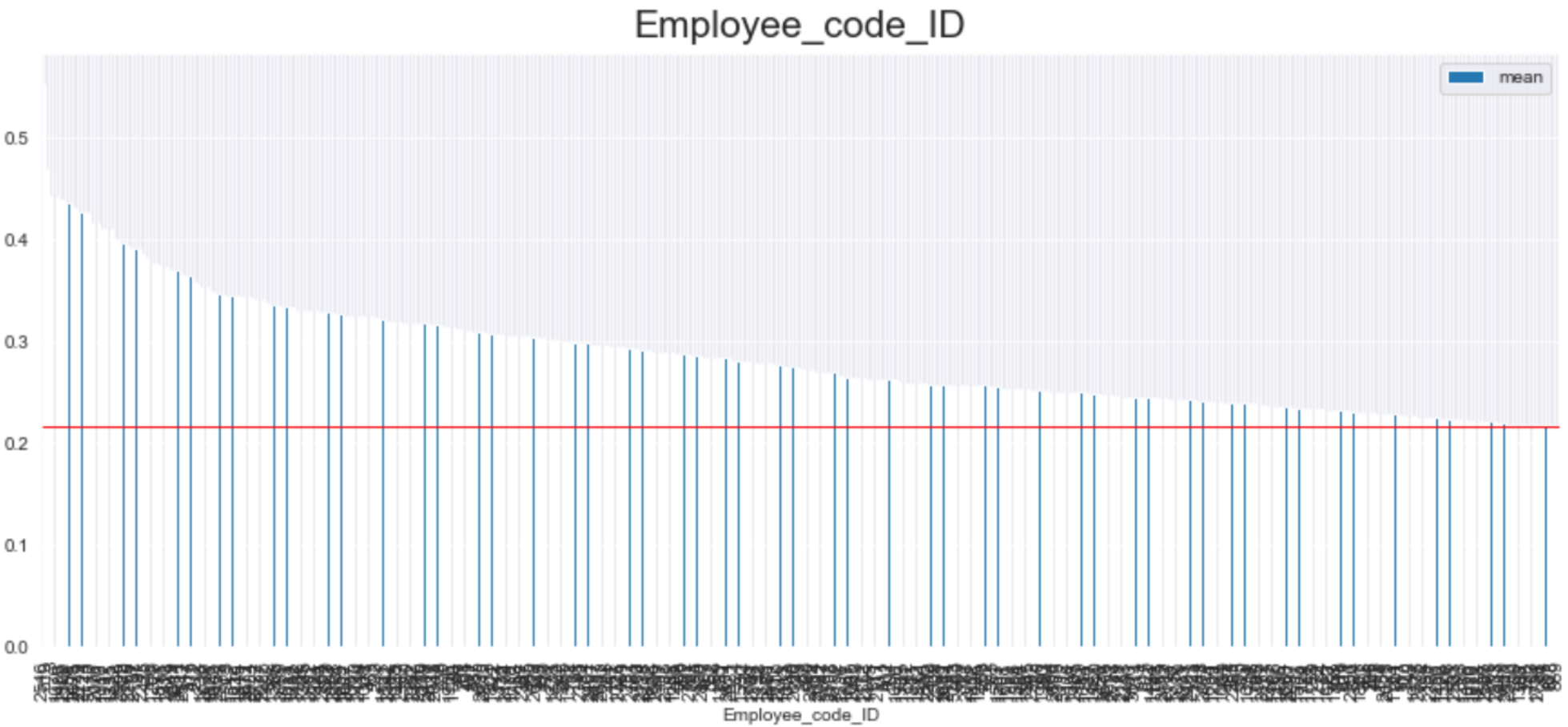


Levels above the threshold: 8

Level with the highest defaulter rate: State_ID 13 (~31%)

SMART SOLUTIONS

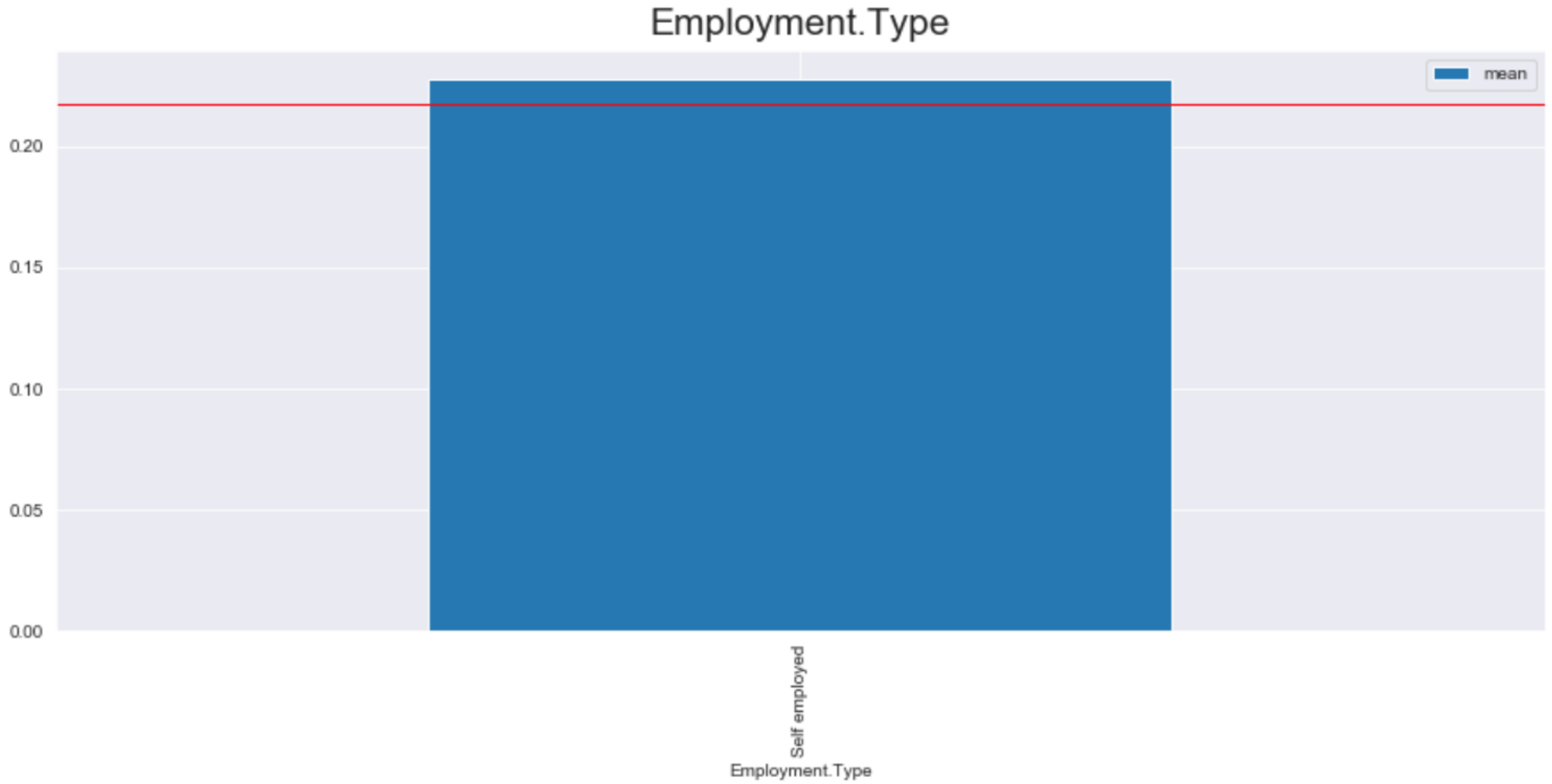
5.c) Histogram



Levels above the threshold: 363

Level with the highest defaulter rate: Employee_code_ID 2546 (~55%)

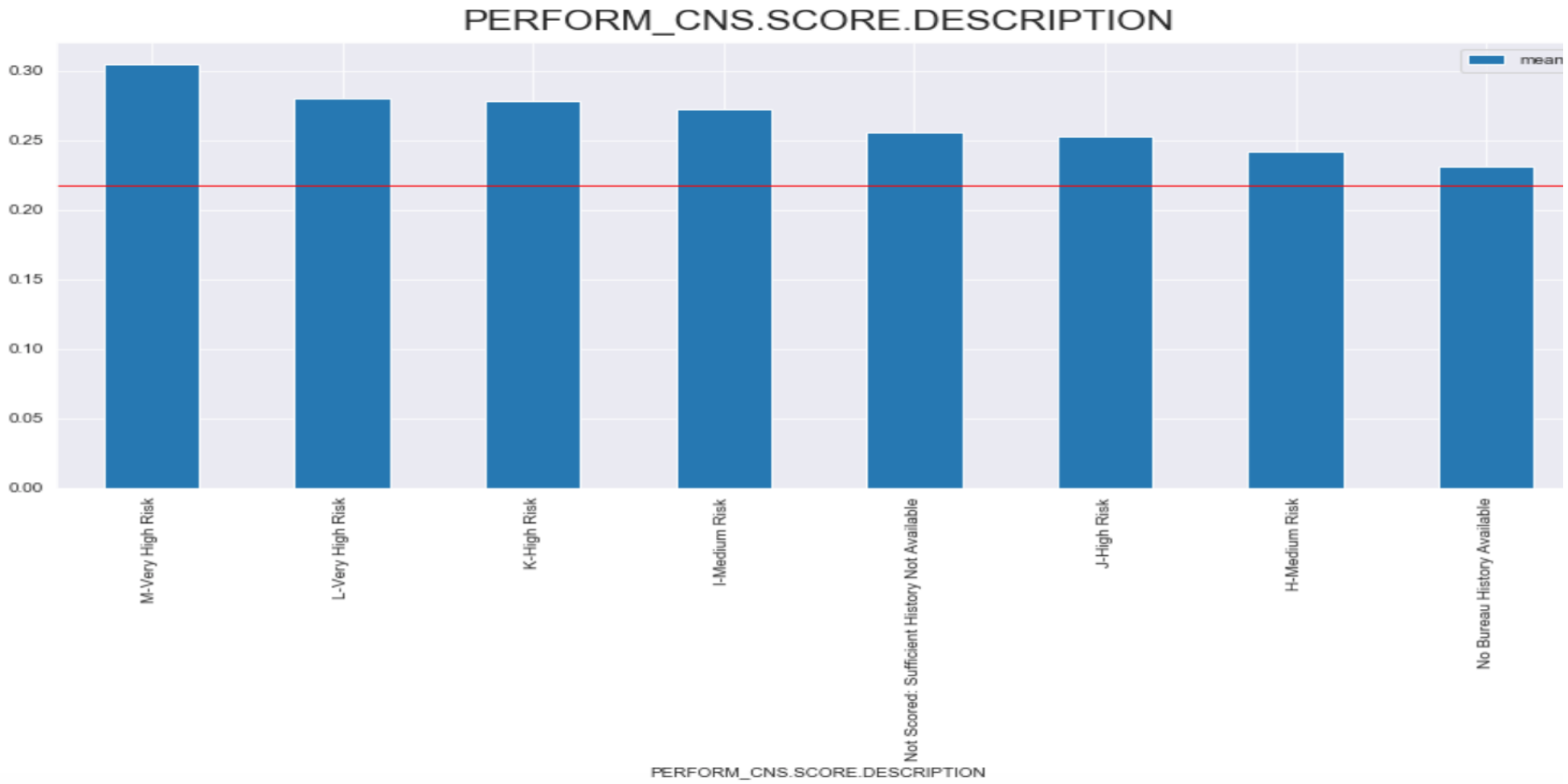
5.c) Histogram



Levels above the threshold: 1

Level with the highest defaulter rate: Self employed (~23%)

5.c) Histogram



Levels above the threshold: 8

Level with the highest defaulter rate: M-Very High Risk (~30%)

SMART SOLUTIONS

5.d) Quadrant Analysis

Metric Definition :

- **Recall**: $\frac{\text{\# customers did not provide information(i.e 0 for a boolean flag feature)}}{\text{\# customers in a selected level for the target group}}$
- **Precision**: $\frac{\text{\# customers defaulted (i.e. 1 for loan\_default feature)}}{\text{\# customers in a selected level for the boolean flag feature group}}$

5.d) Quadrant Analysis

Aadhar_flag

Loan Default	0	1	Total	Recall
0	27684	154859	182543	15.17%
1	9546	41065	50611	18.86%
Total	37230	195924	233154	
Precision	25.64%	20.96%		

Aadhar_flag

Loan Default	0	1	Total
0	11.87%	66.42%	78.29%
1	4.09%	17.61%	21.71%
Total	15.97%	84.03%	100.00%

PAN_flag

Loan Default	0	1	Total	Recall
0	168799	13744	182543	92.47%
1	46734	3877	50611	92.34%
Total	215533	17621	233154	
Precision	21.68%	22.00%		

PAN_flag

Loan Default	0	1	Total
0	72.40%	5.89%	78.29%
1	20.04%	1.66%	21.71%
Total	92.44%	7.56%	100.00%

VoterID_flag

Loan Default	0	1	Total	Recall
0	157565	24978	182543	86.32%
1	41795	8816	50611	82.58%
Total	199360	33794	233154	
Precision	20.96%	26.09%		

VoterID_flag

Loan Default	0	1	Total
0	67.58%	10.71%	78.29%
1	17.93%	3.78%	21.71%
Total	85.51%	14.49%	100.00%

Driving_flag

Loan Default	0	1	Total	Recall
0	178216	4327	182543	97.63%
1	49519	1092	50611	97.84%
Total	227735	5419	233154	
Precision	21.74%	20.15%		

Driving_flag

Loan Default	0	1	Total
0	76.44%	1.86%	78.29%
1	21.24%	0.47%	21.71%
Total	97.68%	2.32%	100.00%

Passport_flag

Loan Default	0	1	Total	Recall
0	182121	422	182543	99.77%
1	50537	74	50611	99.85%
Total	232658	496	233154	
Precision	21.72%	14.92%		

Passport_flag

Loan Default	0	1	Total
0	78.11%	0.18%	78.29%
1	21.68%	0.03%	21.71%
Total	99.79%	0.21%	100.00%

SMART SOLUTIONS

5.d) Quadrant Analysis

Aadhar_flag

Loan Default	0	1	Total	Recall
0	27684	154859	182543	15.17%
1	9546	41065	50611	18.86%
Total	37230	195924	233154	
Precision	25.64%	20.96%		

Aadhar_flag

Loan Default	0	1	Total
0	11.87%	66.42%	78.29%
1	4.09%	17.61%	21.71%
Total	15.97%	84.03%	100.00%

PAN_flag

Loan Default	0	1	Total	Recall
0	168799	13744	182543	92.47%
1	46734	3877	50611	92.34%
Total	215533	17621	233154	
Precision	21.68%	22.00%		

PAN_flag

Loan Default	0	1	Total
0	72.40%	5.89%	78.29%
1	20.04%	1.66%	21.71%
Total	92.44%	7.56%	100.00%

VoterID_flag

Loan Default	0	1	Total	Recall
0	157565	24978	182543	86.32%
1	41795	8816	50611	82.58%
Total	199360	33794	233154	
Precision	20.96%	26.09%		

VoterID_flag

Loan Default	0	1	Total
0	67.58%	10.71%	78.29%
1	17.93%	3.78%	21.71%
Total	85.51%	14.49%	100.00%

Driving_flag

Loan Default	0	1	Total	Recall
0	178216	4327	182543	97.63%
1	49519	1092	50611	97.84%
Total	227735	5419	233154	
Precision	21.74%	20.15%		

Driving_flag

Loan Default	0	1	Total
0	76.44%	1.86%	78.29%
1	21.24%	0.47%	21.71%
Total	97.68%	2.32%	100.00%

Passport_flag

Loan Default	0	1	Total	Recall
0	182121	422	182543	99.77%
1	50537	74	50611	99.85%
Total	232658	496	233154	
Precision	21.72%	14.92%		

Passport_flag

Loan Default	0	1	Total
0	78.11%	0.18%	78.29%
1	21.68%	0.03%	21.71%
Total	99.79%	0.21%	100.00%

SMART SOLUTIONS

5.d) Quadrant Analysis

Aadhar_flag

Loan Default	0	1	Total	Recall
0	27684	154859	182543	15.17%
1	9546	41065	50611	18.86%
Total	37230	195924	233154	
Precision	25.64%	20.96%		

Aadhar_flag

Loan Default	0	1	Total
0	11.87%	66.42%	78.29%
1	4.09%	17.61%	21.71%
Total	15.97%	84.03%	100.00%

Selected features

PAN_flag

Loan Default	0	1	Total	Recall
0	168799	13744	182543	92.47%
1	46734	3877	50611	92.34%
Total	215533	17621	233154	
Precision	21.68%	22.00%		

PAN_flag

Loan Default	0	1	Total
0	72.40%	5.89%	78.29%
1	20.04%	1.66%	21.71%
Total	92.44%	7.56%	100.00%

VoterID_flag

Loan Default	0	1	Total	Recall
0	157565	24978	182543	86.32%
1	41795	8816	50611	82.58%
Total	199360	33794	233154	
Precision	20.96%	26.09%		

VoterID_flag

Loan Default	0	1	Total
0	67.58%	10.71%	78.29%
1	17.93%	3.78%	21.71%
Total	85.51%	14.49%	100.00%

Driving_flag

Loan Default	0	1	Total	Recall
0	178216	4327	182543	97.63%
1	49519	1092	50611	97.84%
Total	227735	5419	233154	
Precision	21.74%	20.15%		

Driving_flag

Loan Default	0	1	Total
0	76.44%	1.86%	78.29%
1	21.24%	0.47%	21.71%
Total	97.68%	2.32%	100.00%

Passport_flag

Loan Default	0	1	Total	Recall
0	182121	422	182543	99.77%
1	50537	74	50611	99.85%
Total	232658	496	233154	
Precision	21.72%	14.92%		

Passport_flag

Loan Default	0	1	Total
0	78.11%	0.18%	78.29%
1	21.68%	0.03%	21.71%
Total	99.79%	0.21%	100.00%

SMART SOLUTIONS

6. Features Selected After EDA

Continuous Features (# : 8)
SEC.SANCTIONED.AMOUNT
PRI.DISBURSED.AMOUNT
PRI.NO.OF.ACCTS
disbursed_amount
AVERAGE.ACCT.AGE
NO.OF_INQUIRIES
PRIMARY.INSTAL.AMT
SEC.INSTAL.AMT

Boolean Categorical Features (# : 942)
Aadhar_flag
VoterID_flag
branch_id (37 levels)
supplier_id (277 levels)
manufacturer_id (5 levels)
Current_pincode_id (241 levels)
State_ID (8 levels)
Employee_code_ID (363 levels)
Employment.Type (1 level; Self Employed)
Perform_CNS.SCORE.Description (8 levels)

In total, **950 features** selected for further evaluation through unsupervised learning

SMART SOLUTIONS

7. Insights through Unsupervised Learning

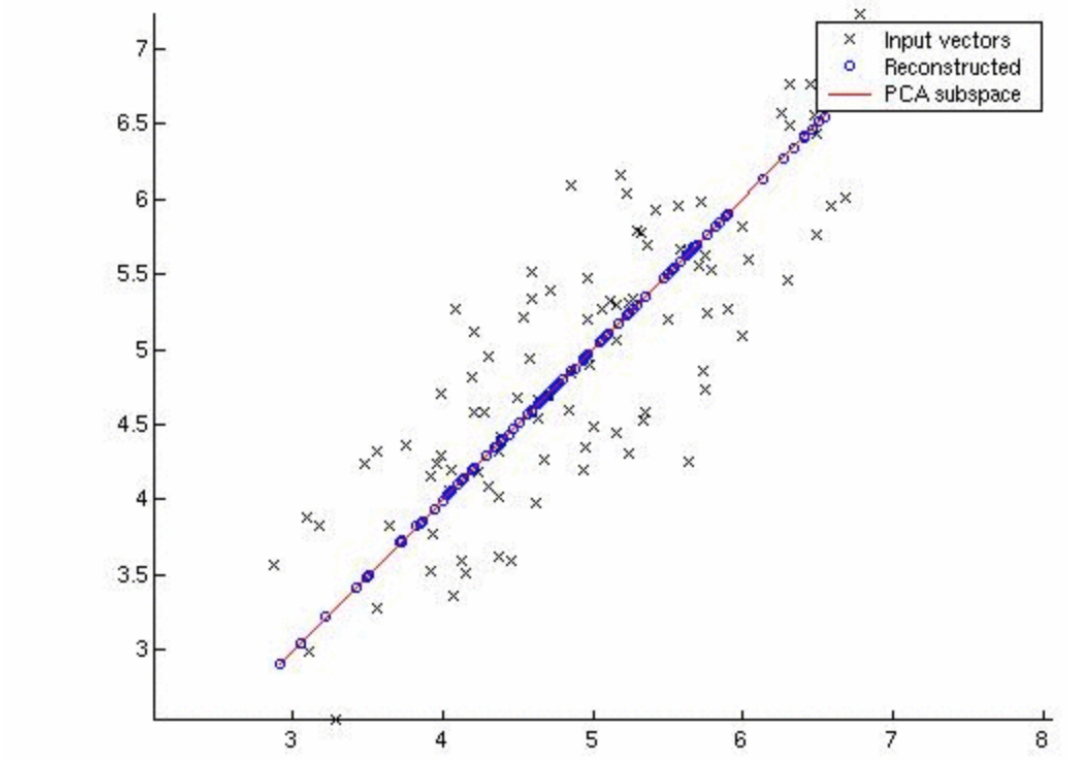
Apply unsupervised machine learning techniques on selected features to discover insights from multi-dimensional space

8. **Perform Dimensionality Reduction to reduce noise in features**
9. **Detect anomalies through Autoencoders technique**
10. **Discover meaningful groups using Clustering techniques (i.e. DBSCAN, HDBSCAN)**

8. Dimensionality Reduction

Dimensionality reduction is an unsupervised machine learning technique of **retaining only the important information** from a large set of input features and representing them through a smaller set of newly generated features.

For example, this picture shows a 2D dataset being mapped to one dimension:



8. Dimensionality Reduction

For boolean categorical features with more than 1 level, PCA (Principal Component Analysis) was applied to reduce the number of levels per feature to just 1.

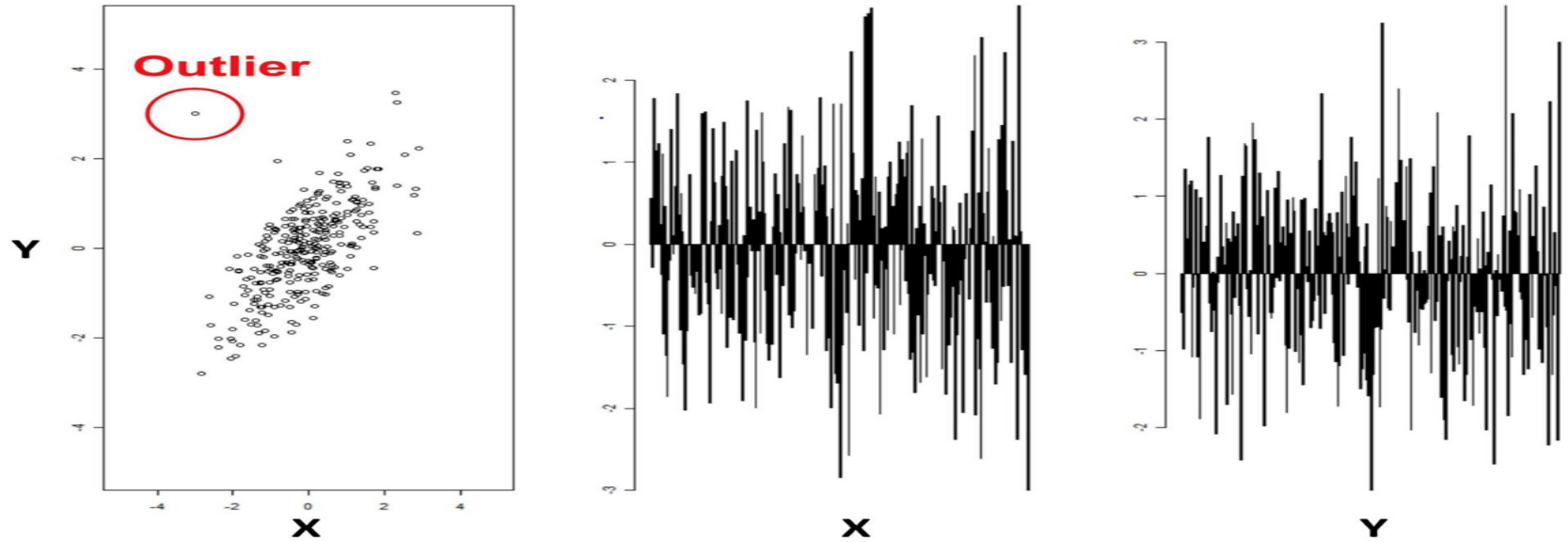
Before PCA (# : 939, data type: Boolean categorical)
branch_id (37 levels)
supplier_id (277 levels)
manufacturer_id (5 levels)
Current_pincode_id (241 levels)
State_ID (8 levels)
Employee_code_ID (363 levels)
Perform_CNS.SCORE.Description (8 levels)

After PCA (# : 7, data type: Continuous)
branch_id
supplier_id
manufacturer_id
Current_pincode_id
State_ID
Employee_code_ID
Perform_CNS.SCORE.Description

This means that the total number of features has **reduced from 950 to just 18**.

SMART SOLUTIONS

9. Anomaly Detection



From 2D plot (left), it is easy to visually identify anomalies located outside the typical distribution. From 1D plots (Right), the anomalies cannot be located investigating a variable independently. Hence, the combination of the variables makes identification of the anomalies easier. However, for multi-dimensional feature space, advanced machine learning techniques like **Autoencoders** are necessary for anomaly detection as it is not possible to do so visually.

9. Anomaly Detection

An autoencoder is a neural network that learns efficient data encodings in an unsupervised manner. The objective is to **generate** from this encoding a **representation** close to its original input.

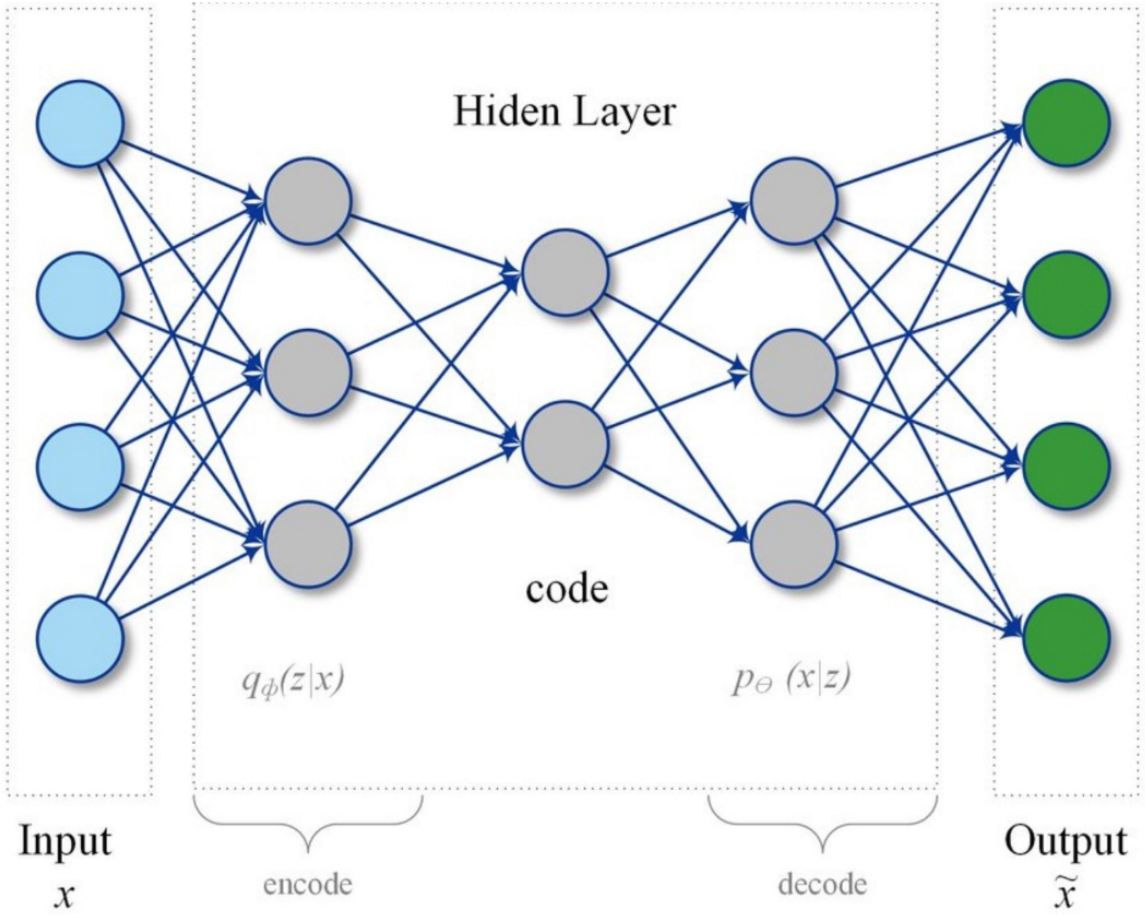


Figure 2: Autoencoder network

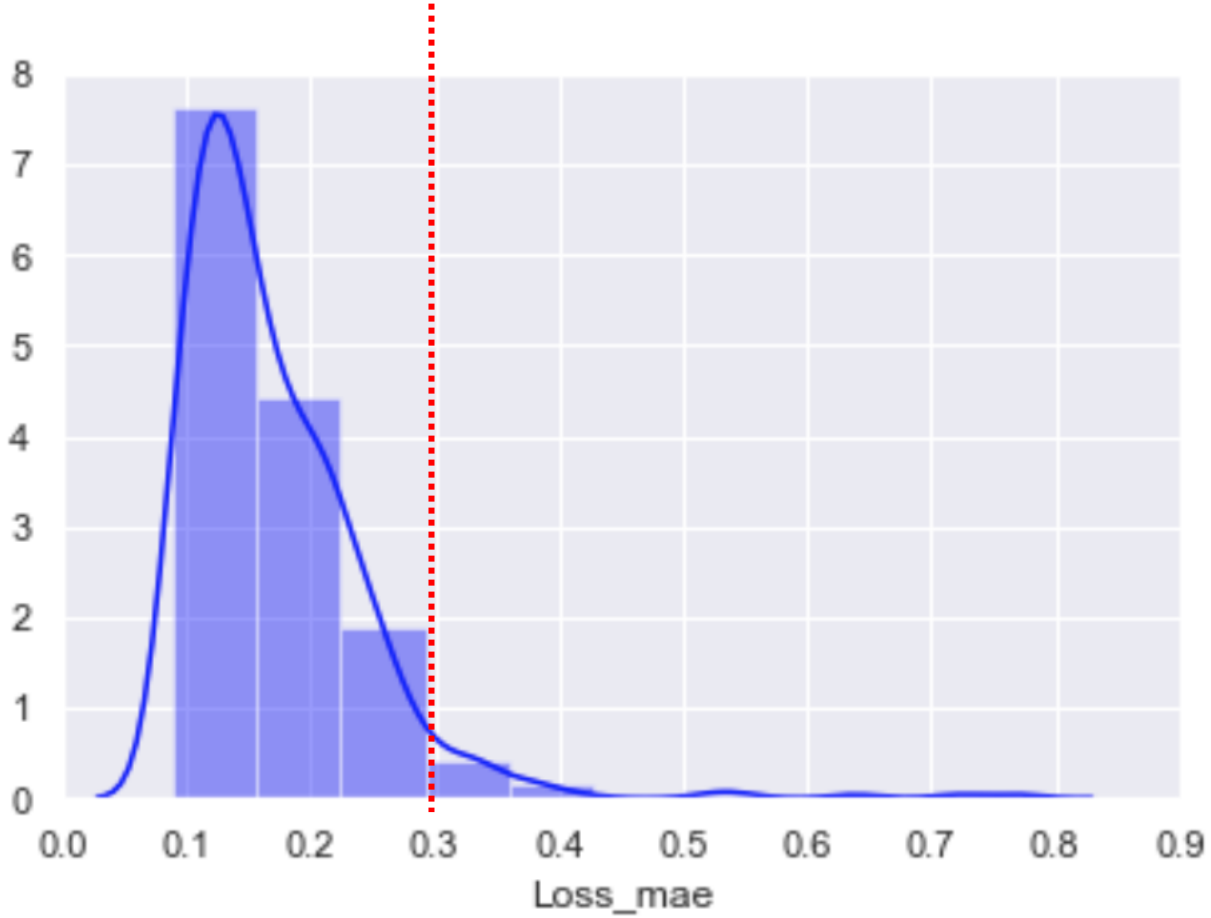
9. Anomaly Detection

The error rate, loss_mae, is then measured between input data and its representation generated by autoencoder.

A threshold is set (red) based on distribution of the error rate.

Data points with error rate higher than the threshold are labelled as anomalies

Consequently, this set of labels is stored as a new feature



9. Anomaly Detection

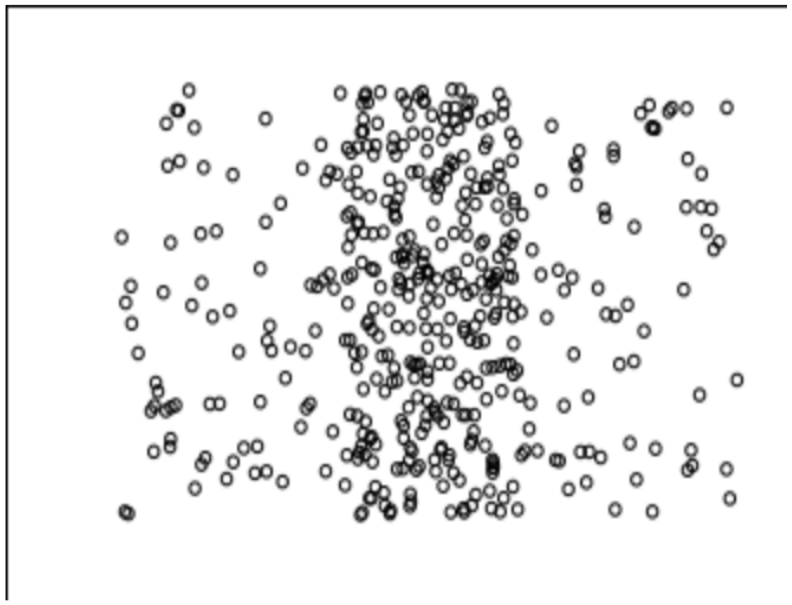
		Anamoly			
		0	1	Total	Recall
Loan Default	0	149761	32782	182543	17.96%
	1	39954	10657	50611	21.06%
	Total	189715	43439	233154	
	Precision	21.06%	24.53%		

		Anamoly		
		0	1	Total
Loan Default	0	64.23%	14.06%	78.29%
	1	17.14%	4.57%	21.71%
	Total	81.37%	18.63%	100.00%

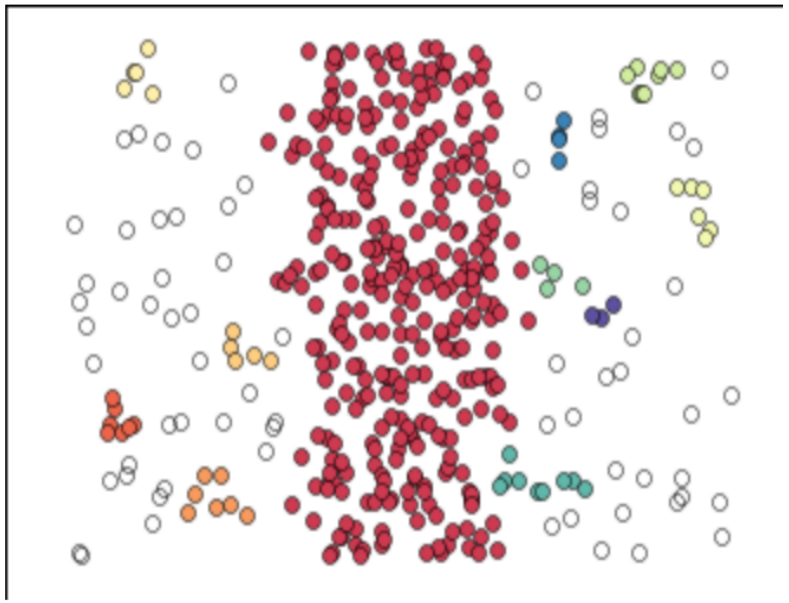
Comparing overlap between defaulters and anomalous data, it can be deduced that **Anomaly as feature does affect the target label**, Loan Default, significantly as there is discrepancy in proportion within both precision set and recall set.

SMART SOLUTIONS

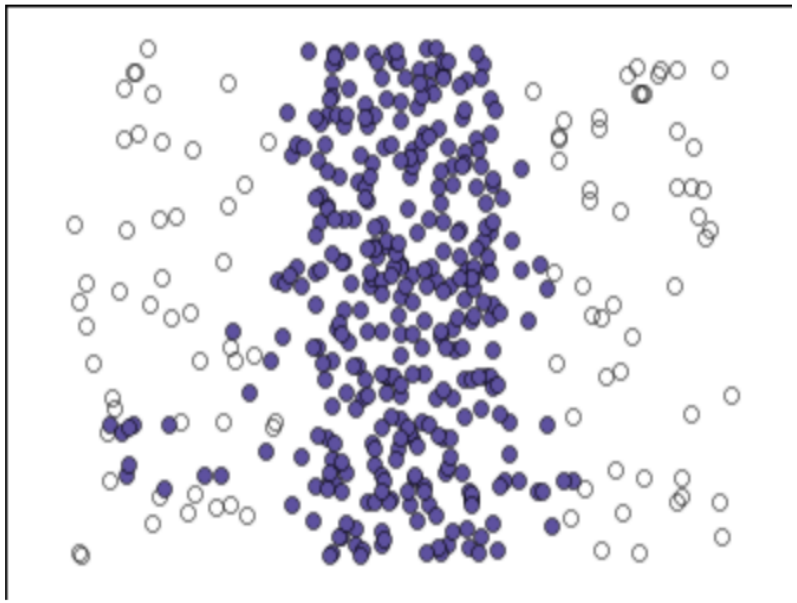
10. Clustering: DBSCAN & HDBSCAN



Density Bars



Density Bars with DBScan Applied



Density Bars with HDBScan Applied

DBSCAN & HDBSCAN are both density based clustering techniques. These are better at detecting non linear patterns (i.e. circle within a circle). HDBSCAN, experimentally, tend to perform better than DBSCAN as it can separate clusters from noise better and takes shorter time to compute.

10. Clustering: DBSCAN Clusters Vs. Loan Default

Levels above the threshold: 64

Level with the highest defaulter rate: db_clusters_182 (~60%)

SMART SOLUTIONS

10. Clustering: HDBSCAN Clusters Vs. Loan Default

Levels above the threshold: 79

Level with the highest defaulter rate: hdb_clusters_277 (~43%)

SMART SOLUTIONS

10. Clustering: DBSCAN & HDBSCAN

Comparing overlap between defaulters and DBSCAN or HDBSCAN generated labels, it can be deduced that **both set of labels as features does affect the target label**, Loan Default, as there are clusters with significantly higher than average defaulter rate found in both.

Characteristics of these clusters can be further explored by looking into averages of other features beside loan default (i.e. disbursed_amount).

Since these are binary categorical features with high dimensionality (i.e. 64 clusters for DBSCAN, 79 clusters for HDBSCAN), PCA can be applied to generate 1 representative cluster for each of the clustering methodologies.

11. Final Feature Set

Continuous Features (# : 17)
SEC.SANCTIONED.AMOUNT
PRI.DISBURSED.AMOUNT
PRI.NO.OF.ACCTS
disbursed_amount
AVERAGE.ACCT.AGE
NO.OF_INQUIRIES
PRIMARY.INSTAL.AMT
SEC.INSTAL.AMT
branch_id_PCA
supplier_id_PCA
manufacturer_id_PCA
Current_pincode_id_PCA
State_ID_PCA
Employee_code_ID_PCA
Perform_CNS.SCORE.Description_PCA
DBSCAN_clusters_PCA
HDBSCAN_clusters_PCA

Boolean Categorical Features (# : 4)
Aadhar_flag
VoterID_flag
Employment.Type (1 level; Self Employed)
anomaly

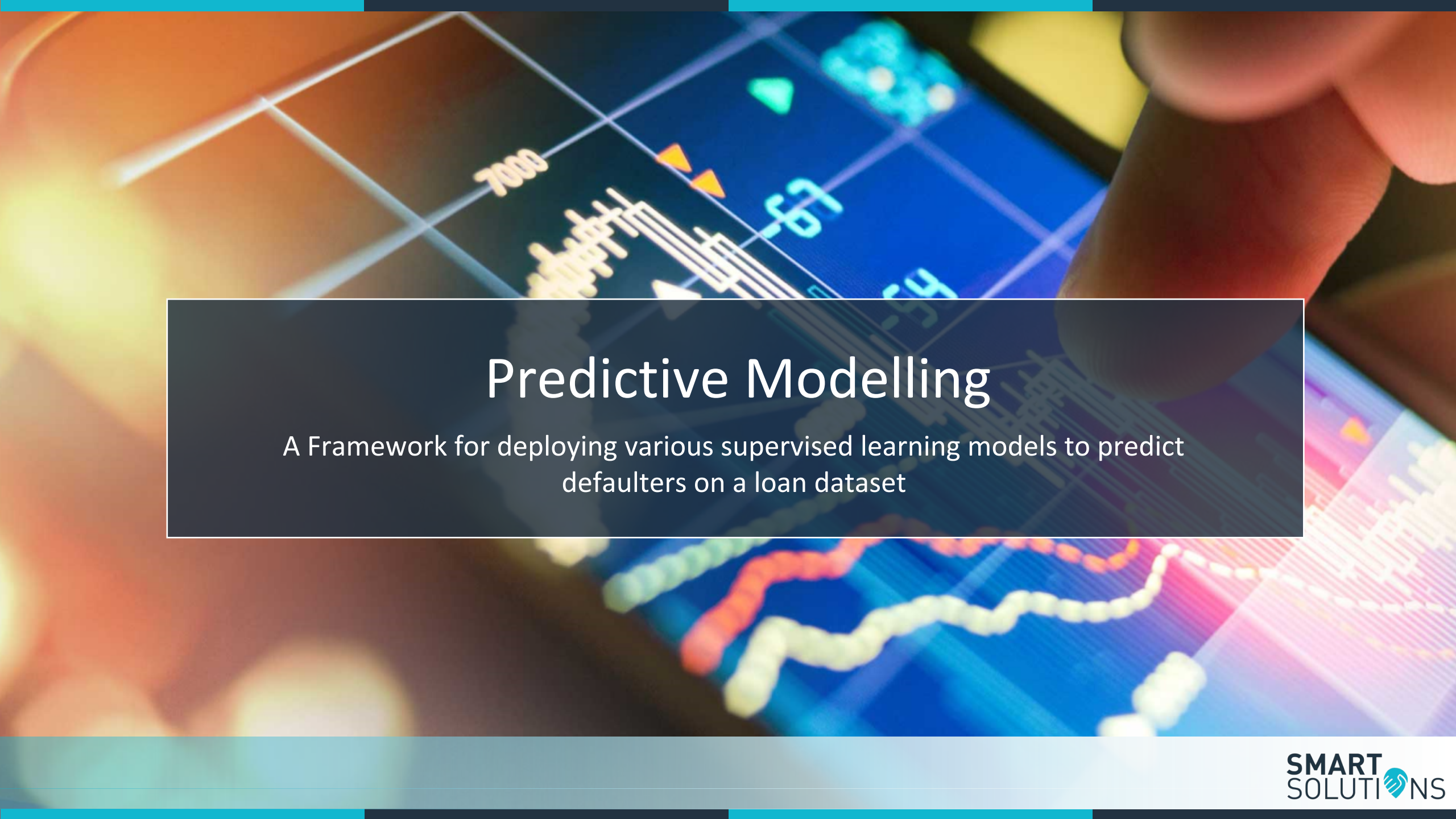
In total, **21 features** that meets the objective criterion selected for further analysis after EDA

Strengths of these features to accurately predict defaulters through application of supervised learning model (*Next Step after EDA*)

SMART SOLUTIONS

SMART SOLUTIONS

SMART SOLUTIONS



Predictive Modelling

A Framework for deploying various supervised learning models to predict defaulters on a loan dataset

SMART SOLUTIONS

Predictive Modelling Framework

1. Review of EDA
2. Handling Imbalanced Dataset
3. Metrics for Evaluating Model Performance (*Precision, Recall, F1 and ROC*)
4. Supervised Learning Models
    a) Logistic Regression
    b) Support Vector Machine
    c) Decision Trees & Random Forest
    d) K – Nearest Neighbors
    e) Naïve Bayes
5. Ensemble Learning Models
    a) Bagging
    b) Boosting
    c) Stacking
6. Conclusion

1. Review of EDA

- **21 features were selected after EDA on the basis:**
  - **Pairwise correlations among features are insignificant ➔ Features are approximately independent of one another**
  - **Each feature showed significant effect on the target label, “Loan Default”**
- **Features, in general, were left skewed ➔ Not normally distributed**
- **Anomalies alone were not enough to distinguish between defaulters and non-defaulters**

2. Handling Imbalanced Dataset

- *Recall*: 21.7% (50,611 out of 233,154 customers) are defaulters in the dataset ➔ Imbalanced Dataset
- To balance the dataset, randomly selected equal proportion of non defaulters as defaulters (another 50,611 customers who did not default)
- In total, 43.4% (101,222 out of 233,154 customers) were selected for modelling with equal ratio between defaulter and non defaulter customers

3. Metrics for Evaluating Model Performance

```
X_train.shape, y_train.shape, X_test.shape, y_test.shape
```

```
((75916, 21), (75916,), (25306, 21), (25306,))
```

Data Split

```
print(y_train.value_counts(),'\n', y_test.value_counts())
```

```
1    37958
0    37958
Name: loan_default, dtype: int64
 1    12653
0    12653
Name: loan_default, dtype: int64
```

50% defaulter rate across both train and test set

Train Test Split:

- 75% data to Train a model
- 25% data to score the trained model

3. Metrics for Evaluating Model Performance

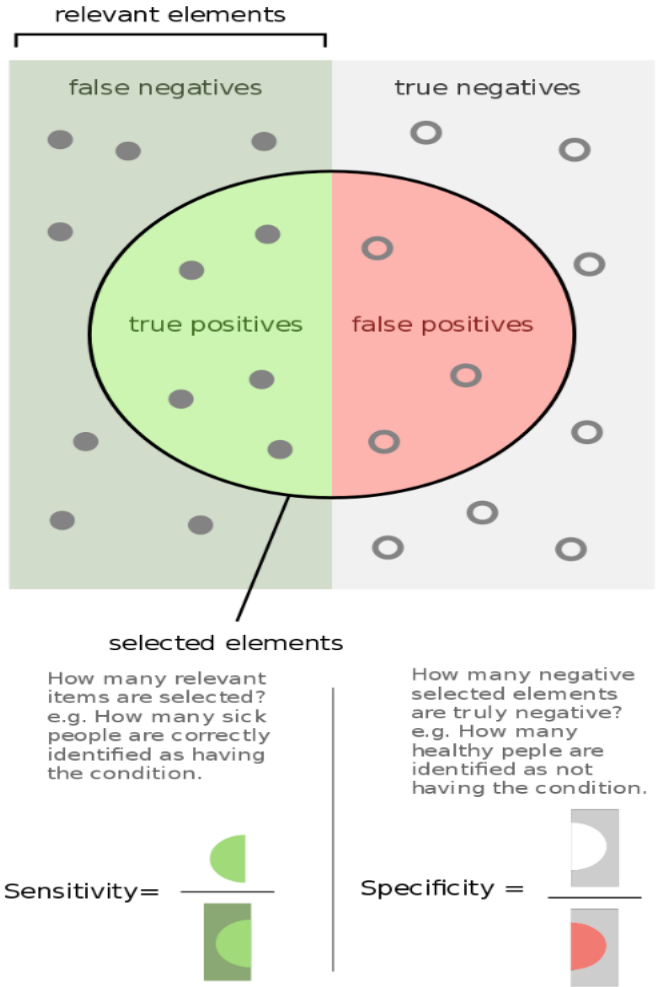
Metric Definition :

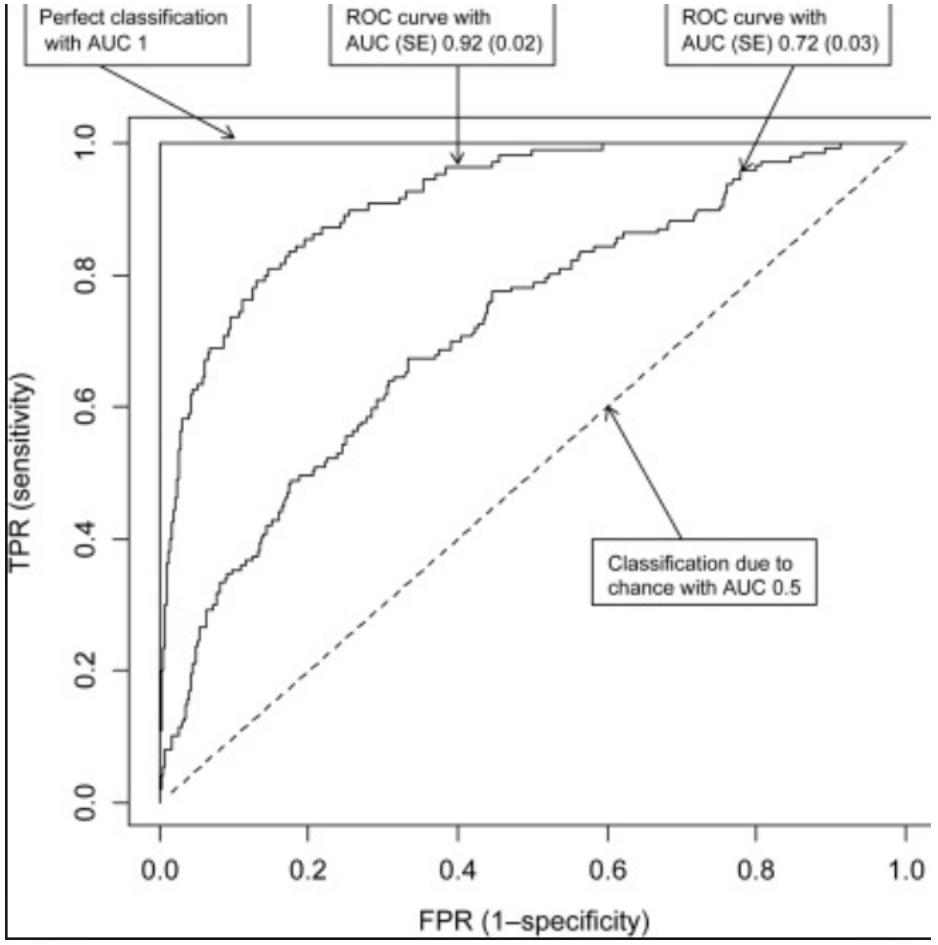
- **Recall (Sensitivity)** : $\frac{\text{\# defaulters correctly detected}}{\text{\# actual defaulters}}$
- **Precision**: $\frac{\text{\# defaulters correctly detected}}{\text{\# predicted defaulters}}$
- **F1**: $\frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$

F1 is more reliable metrics than Recall or precision alone as it penalizes any low value for precision or recall. Hence, **high F1 suggests a strong model**.

Accuracy is not chosen as a metric since the focus is only to accurately detect defaulters, not non defaulters.

3. Metrics for Evaluating Model Performance





AUC (Area Under Curve) for the ROC plot (left) is another reliable metric to identify a good model

Closer AUC is to 1, stronger the model

SMART SOLUTIONS

4. Supervised Learning

"Unlike unsupervised learning model which helps to generate suitable labels on unlabeled dataset, supervised learning model learns from given labels and tries to ***replicate the learned labeling pattern*** *on a separate dataset"*

4. Supervised Learning

Known Labels

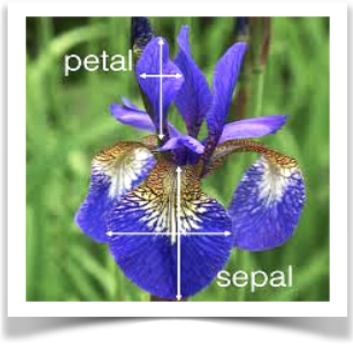
Model-building

MACHINE LEARNING

Model

Setosa
Versicolor
Virginica

Predicted Labels



Known Features

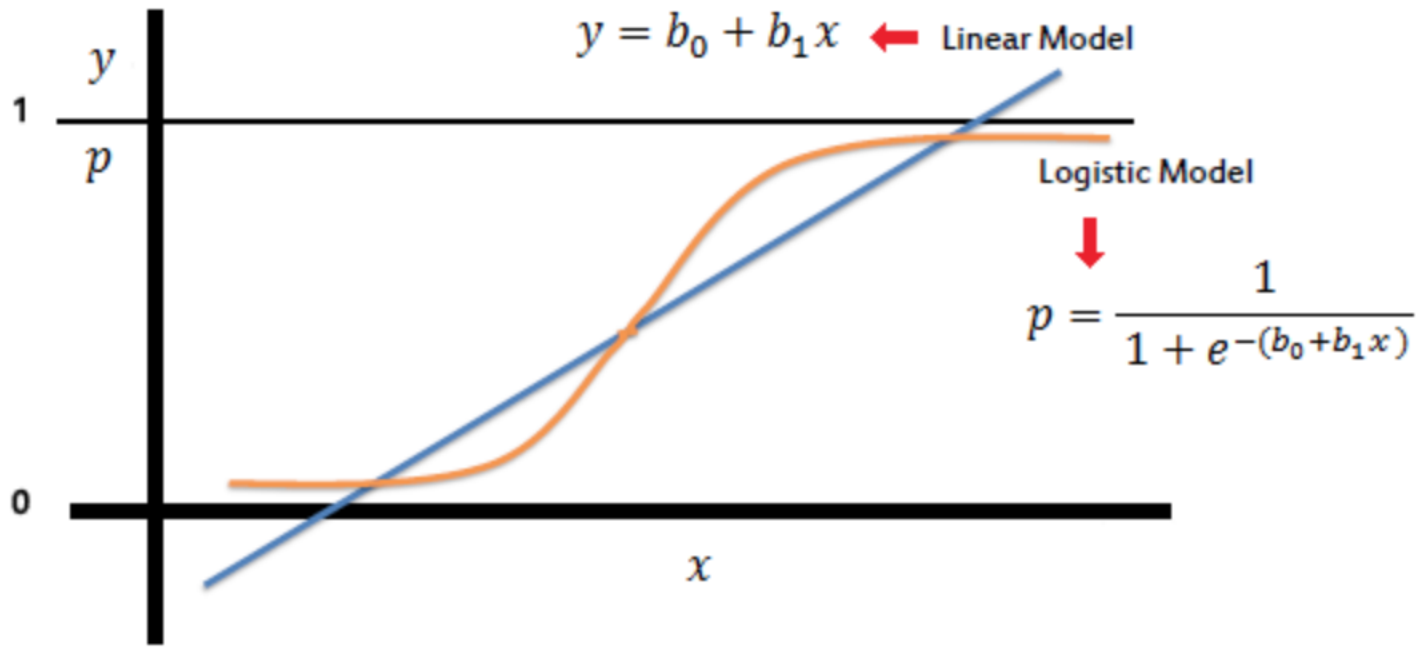
Sepal Length
Sepal Width
Petal Length
Petal Width

New Data

Sepal Length
Sepal Width
Petal Length
Petal Width

SMART SOLUTIONS

4.a) Logistic Regression (Concept)



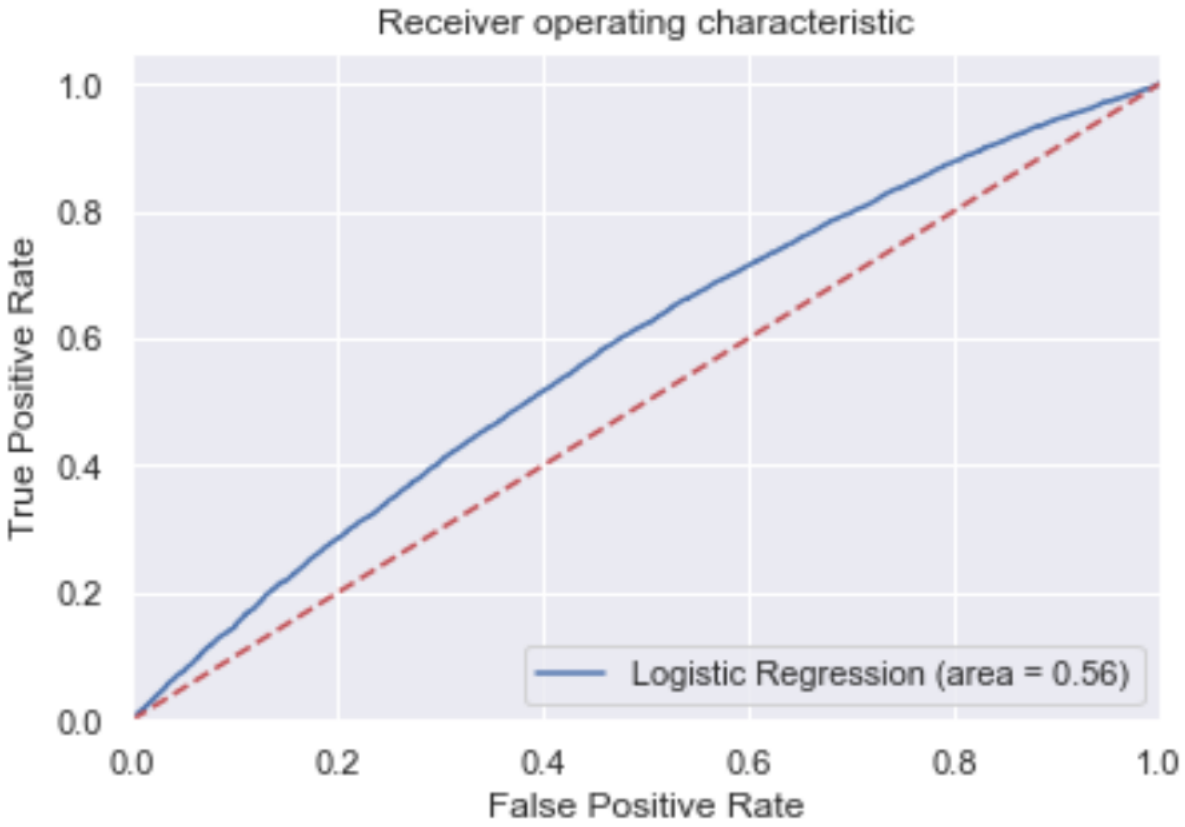
Logistic Regression is a generalized linear model whose **output is bounded by (0,1)**, making it suitable for classification tasks

Besides, it is simple, interpretable and scalable on big data

Hence, usually used as benchmark model to validate performance of more complex model

SMART SOLUTIONS

4.a) Logistic Regression (Result)

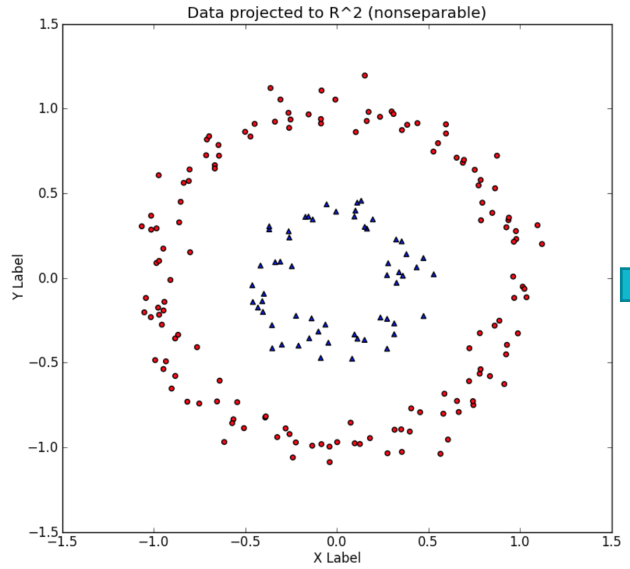


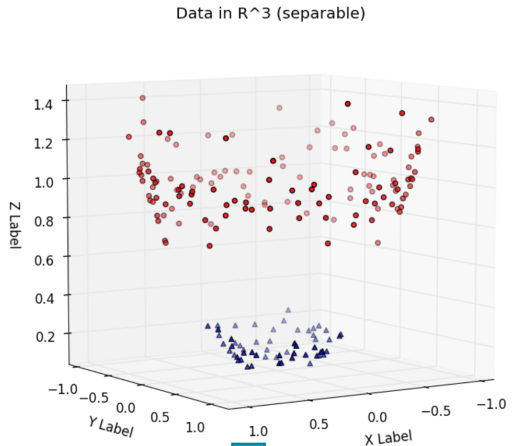
	precision	recall	f1-score
0	0.56	0.55	0.55
1	0.56	0.58	0.57

As a baseline model, it generated a decent F1 and AUC scores at 0.57 and 0.56 respectively

SMART SOLUTIONS

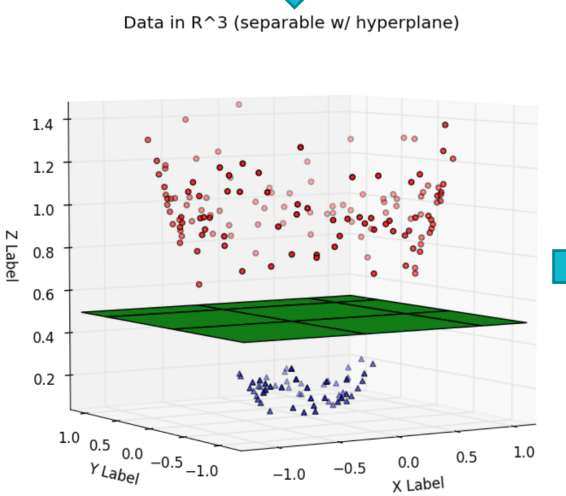
4.b) Support Vector Machine (Concept)

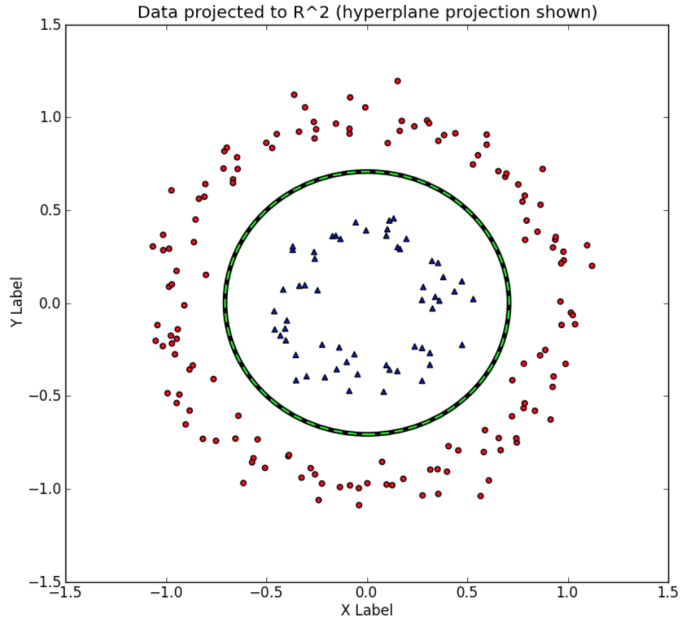




Model that transforms data that is linearly non separable to linearly separable by **adding a Support Vector**, Label Z, which can be interpreted as sqrt(X^2+Y^2) here.

Then, a hyperplane (**green**) is used **separate labels in higher dimension** before transforming back to original dimensionality.





SMART SOLUTIONS

4.b) Support Vector Machine (Result)

	precision	recall	f1-score
0	0.55	0.64	0.59
1	0.57	0.47	0.52

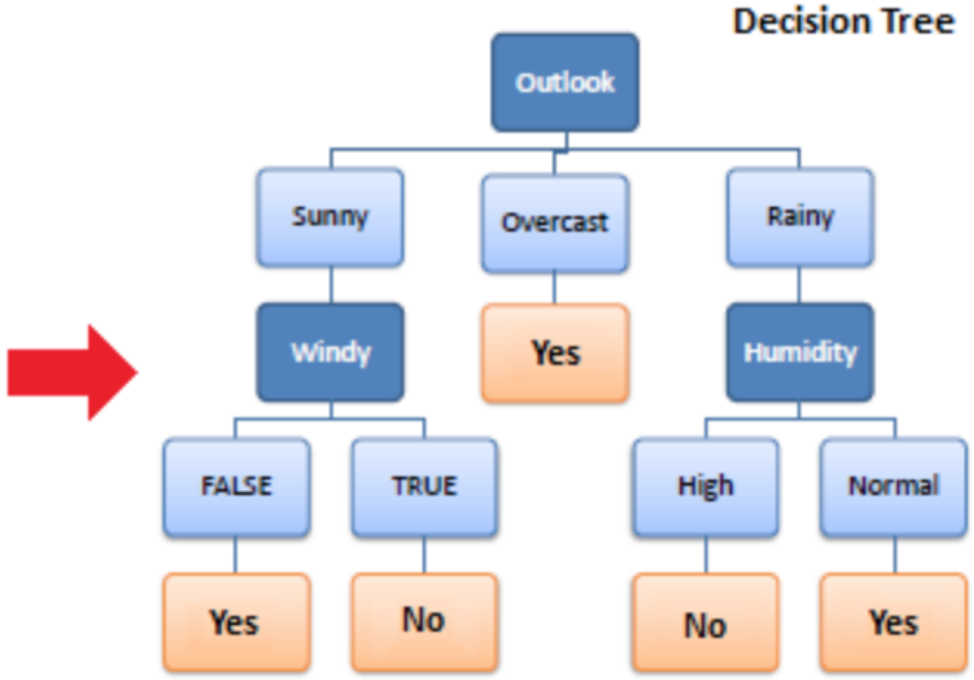
Compared to baseline, F1 has dropped significantly while AUC remains the same.

In addition, the model is computationally expensive and takes very long to run.

4.c) Decision Tree (Concept)

Predictors | Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



A model that learns **a decision structure** from training data and apply this rule based decision making to predict on test data. The **learning** involves 2 optimization process to address the following questions:

1. Which feature to select first?
2. When to stop growing the tree?

SMART SOLUTIONS

4.c) Decision Tree (Concept)

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

1. **Which feature to select first?**

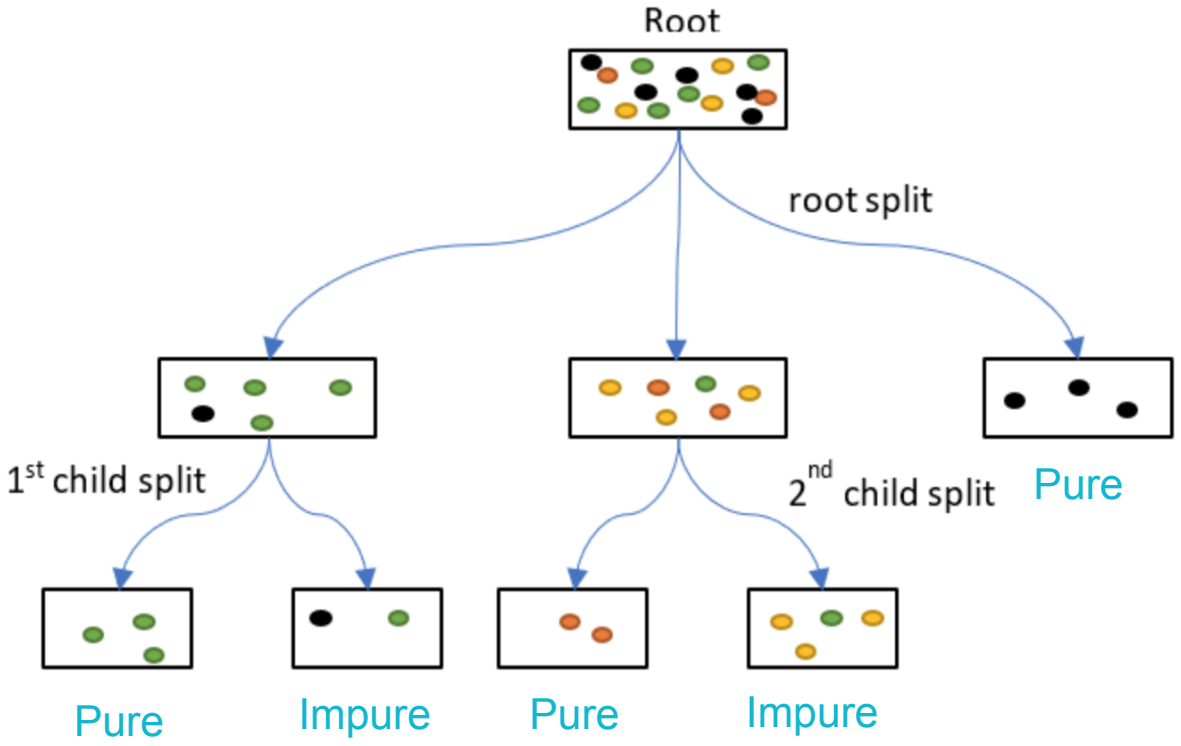
Feature that gives the highest **information gain**.

"*Information gain* is a measurement of how much increase in certainty is achieved from a position of uncertainty"

For example, having "*outlook*" feature as the 1st feature to split on makes it *easier* to derive pattern from this data.

SMART SOLUTIONS

4.c) Decision Tree (Concept)



However, other stopping criterion can be applied like **min # of samples** (5 samples) or **max. tree depth** (2 split levels).

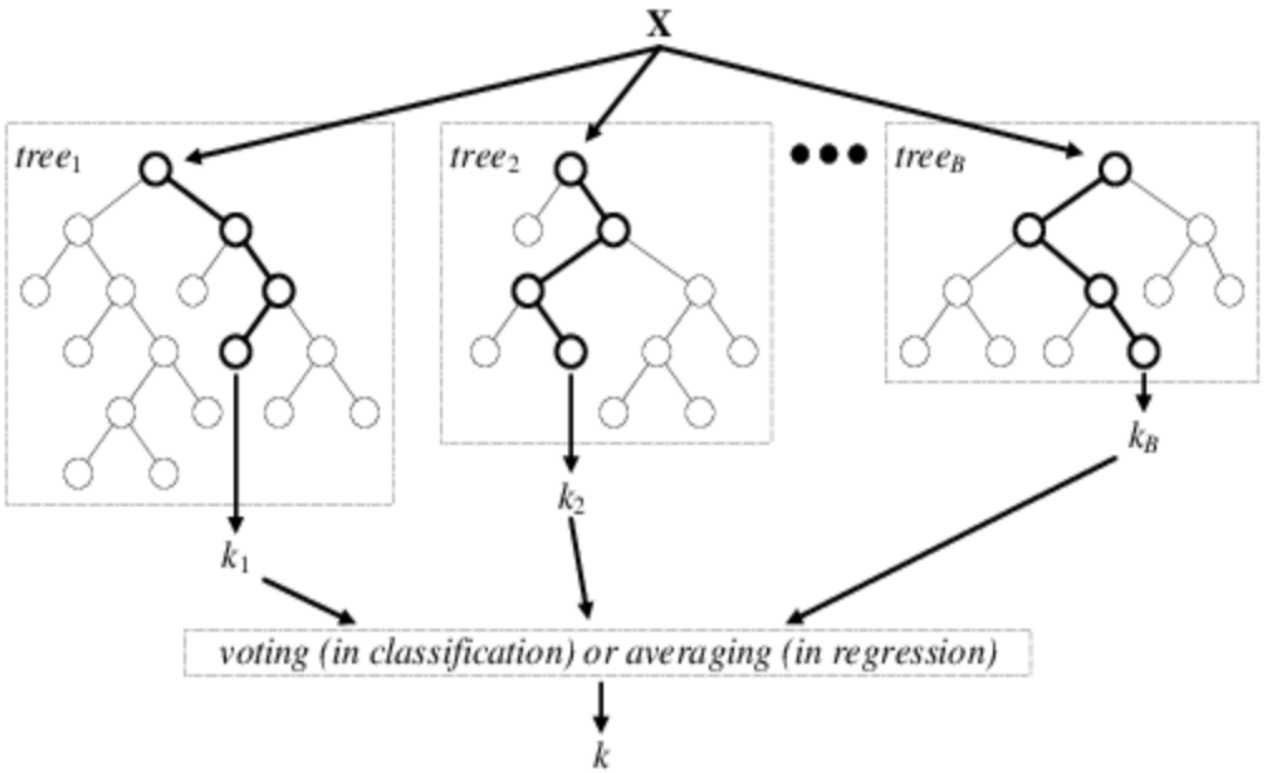
2. When to stop growing the tree?

When each resulting node after the split is **pure/homogeneous.**

"A node is usually considered pure/homogeneous if it contains 90%-100% of only one type of target label"

For example, if a node contains 9 out of 10 its data points where the person played golf, the node is pure and need no further splitting.

4.c) Random Forest (Concept)



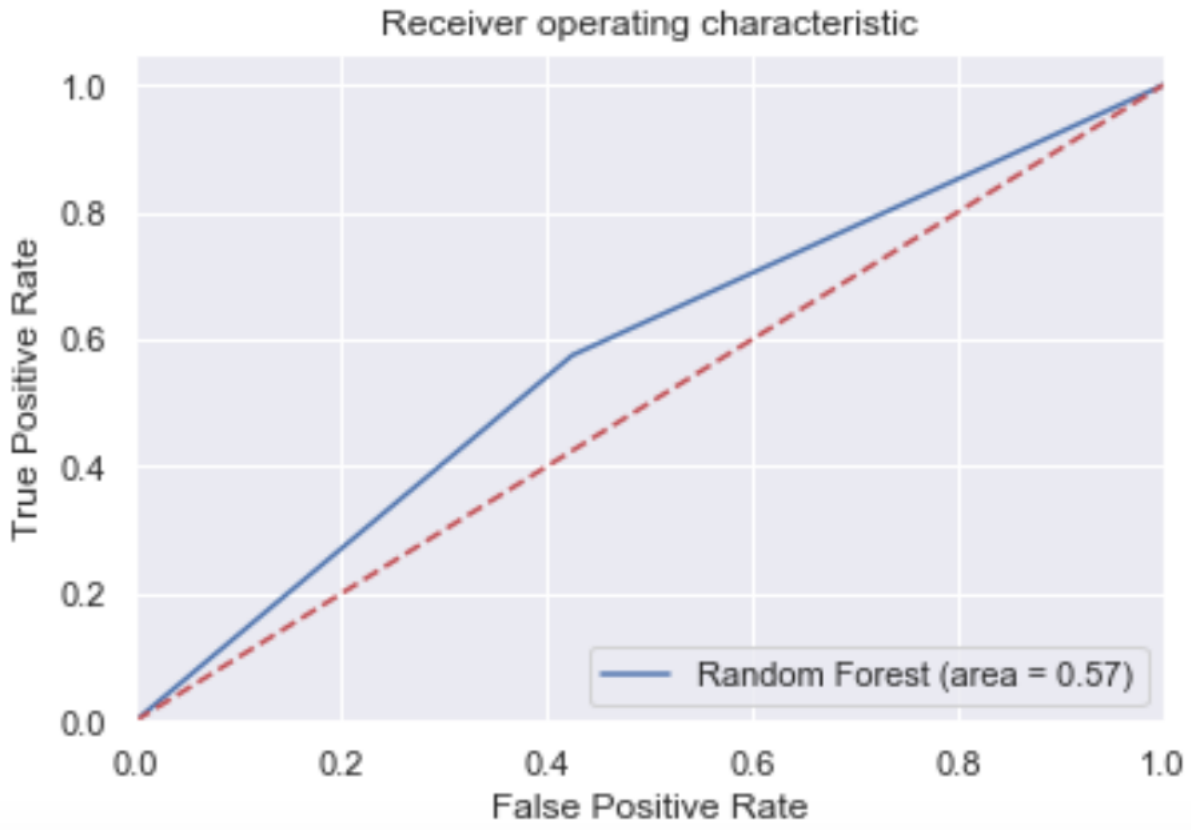
A model that **builds multiple unique decision trees** based on sampling a subset of the entire training data.

Then, it collects predictions on test data from all its trees and report the **most occurring predicted label as the final predicted label**.

Why Random Forest *better* than a Decision Tree?

- More robust to unpredictable samples (i.e. anomalies)
- **Idea**: Diversified Portfolio ➔ Less Risk

4.c) Random Forest (Result)



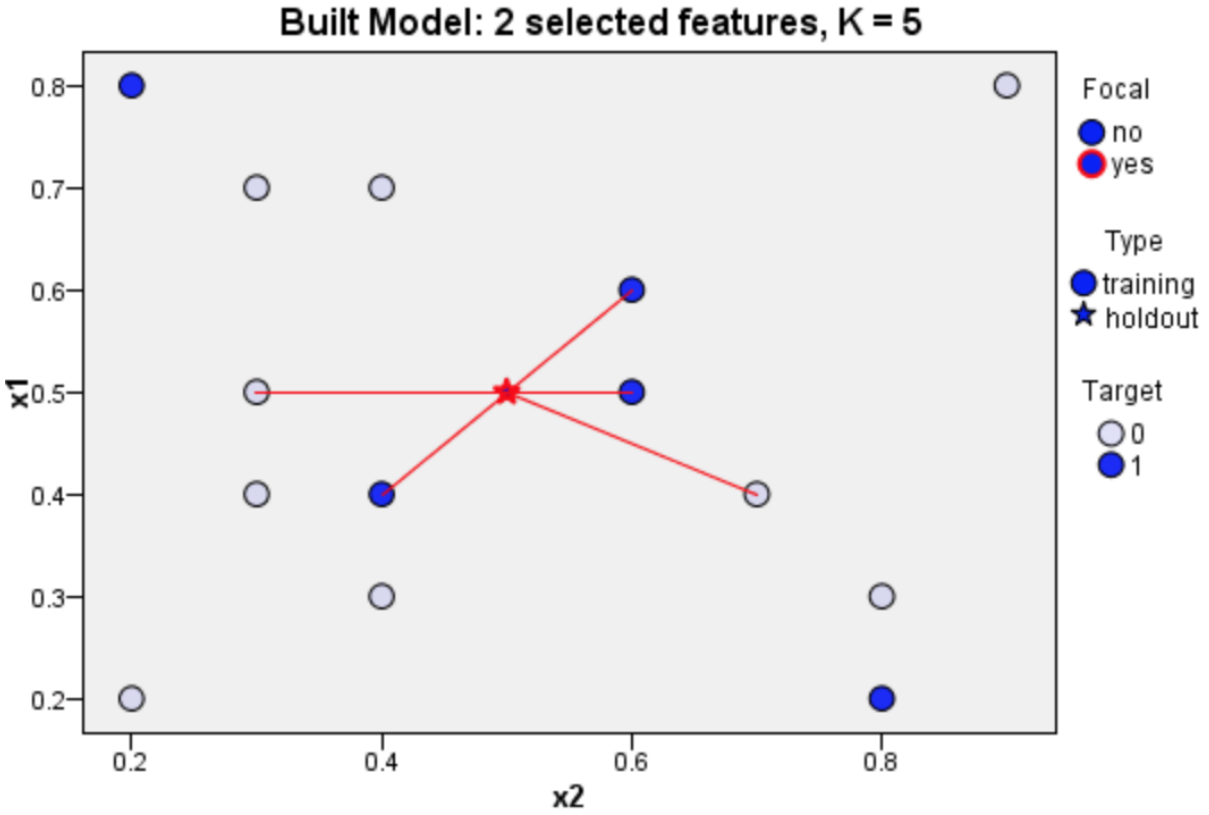
	precision	recall	f1-score
0	0.57	0.58	0.58
1	0.57	0.57	0.57

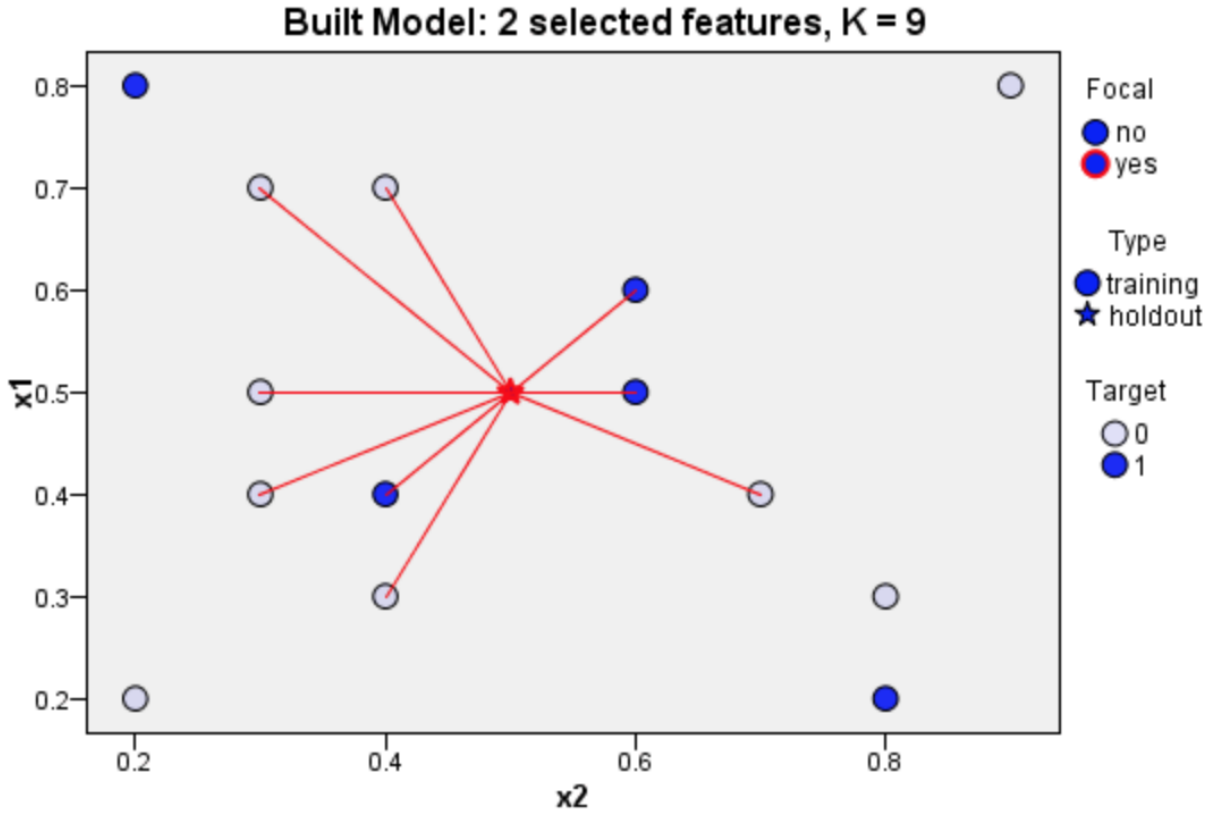
Forest with 200 trees produced the best results from tested set {50, 100, 200, 400}.

This model performs slightly better than baseline since AUC is slightly higher while F1 remains the same.

SMART SOLUTIONS

4.d) K-Nearest Neighbors (Result)

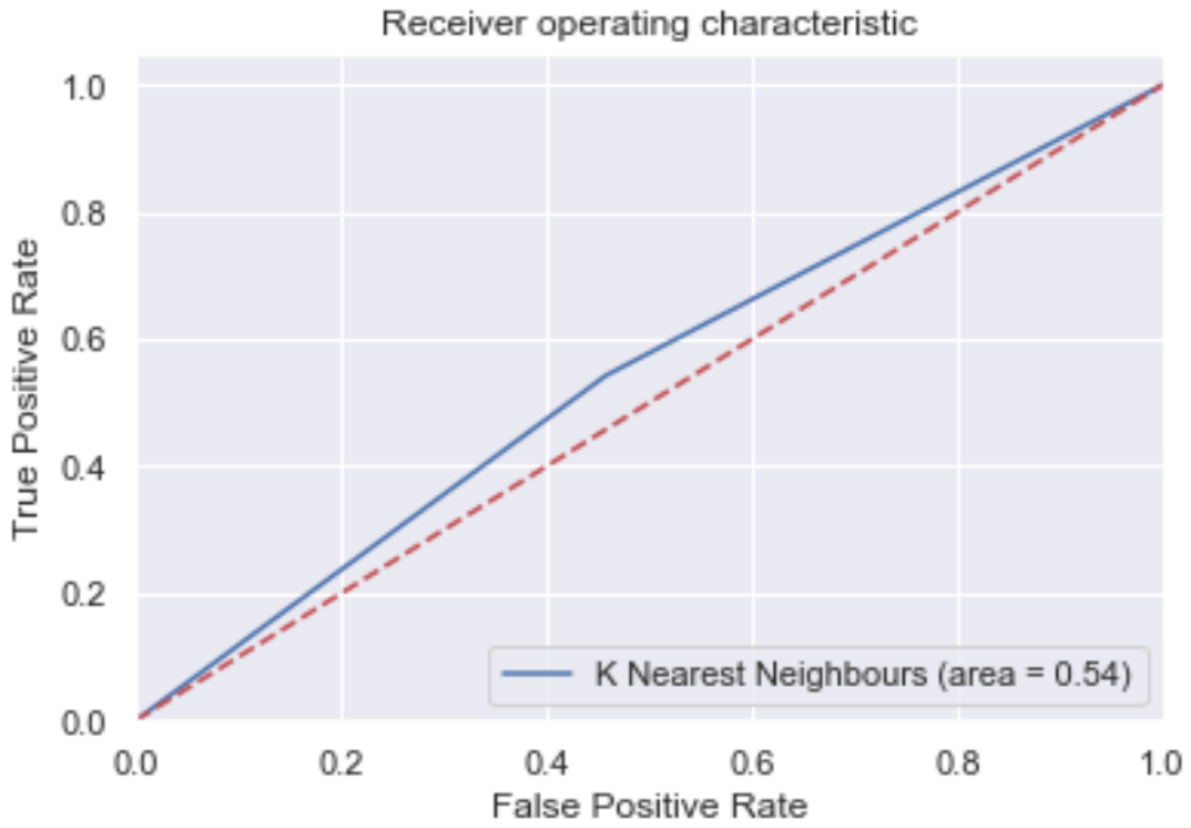




A simple model that specifies a label for a test data as the label which is the **majority among its nearest neighbors** of labeled trained data. How many neighbors to consider is used defined.

SMART SOLUTIONS

4.d) K-Nearest Neighbors (Result)



	precision	recall	f1-score
0	0.54	0.54	0.54
1	0.54	0.54	0.54

K= 9 (i.e. 9 nearest neighbors) was best possible KNN model generated from a K set {1,25}.

Yet, this model turns out to be not suitable as it performs worse than baseline for both F1 and AUC

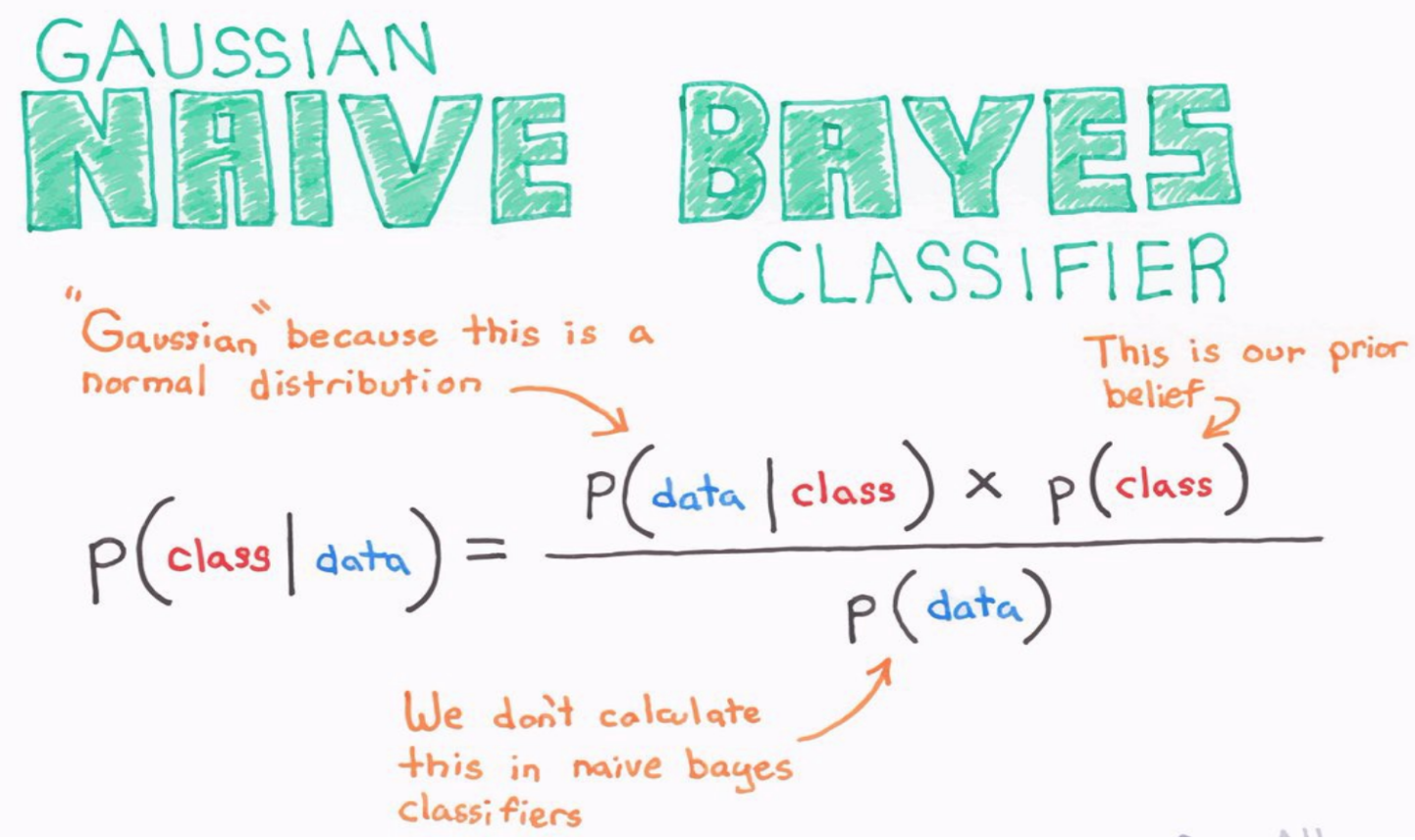
SMART SOLUTIONS

4.e) Naïve Bayes (Concept)

		Play Golf		Statistics			
	Train Data	No	Yes	Total	P(Weather)	P(No|Weather)	P(Yes|Weather)
Weather	**Rainy**	15	1	16	**0.39**	**0.94**	**0.06**
	Sunny	5	20	25	**0.61**	**0.2**	**0.8**
Statistics	Total	20	21	41			
	P(Golf)	**0.49**	**0.51**				
	P(Rainy|Golf)	**0.75**	**0.05**				
	P(Sunny|Golf)	**0.25**	**0.95**				

Given training data with 1 feature (weather) and a target label (Play Golf), can we predict if Golf was played given a test data on the weather? For example, if the weather is Sunny, what will be our prediction be?

4.e) Naïve Bayes (Concept)



Using the Naïve Bayes formula (left), construct the probabilities of each predicted class (i.e. play golf = Yes, Play golf = No)

$$P(\text{Play Golf = Yes} \mid \text{Sunny}) = \frac{P(\text{Sunny} \mid \text{Played Golf}) * P(\text{Played Golf})}{P(\text{Sunny})}$$

$$P(\text{Play Golf = No} \mid \text{Sunny}) = \frac{P(\text{Sunny} \mid \text{Not Played Golf}) * P(\text{Not Played Golf})}{P(\text{Sunny})}$$

SMART SOLUTIONS

4.e) Naïve Bayes (Concept)

		Play Golf		Statistics			
	Train Data	No	Yes	Total	P(Weather)	P(No|Weather)	P(Yes|Weather)
Weather	**Rainy**	15	1	16	**0.39**	**0.94**	**0.06**
	Sunny	5	20	25	**0.61**	**0.2**	**0.8**
Statistics	Total	20	21	41			
	P(Golf)	**0.49**	**0.51**				
	P(Rainy|Golf)	**0.75**	**0.05**				
	P(Sunny|Golf)	**0.25**	**0.95**				

$$P(\text{Play Golf = Yes} \mid \text{Sunny}) = \frac{P(\text{Sunny}\mid \text{Played Golf}) * P(\text{Played Golf})}{P(\text{Sunny})} = \frac{0.95*0.51}{0.61} = 0.8$$

$$P(\text{Play Golf = No} \mid \text{Sunny}) = \frac{P(\text{Sunny}\mid \text{Not Played Golf}) * P(\text{Not Played Golf})}{P(\text{Sunny})} = \frac{0.25*0.49}{0.61} = 0.2$$

SMART SOLUTIONS

4.e) Naïve Bayes (Concept)

		Play Golf		Statistics			
	Train Data	No	Yes	Total	P(Weather)	P(No|Weather)	P(Yes|Weather)
Weather	**Rainy**	15	1	16	**0.39**	**0.94**	**0.06**
	Sunny	5	20	25	**0.61**	**0.2**	**0.8**
Statistics	Total	20	21	41			
	P(Golf)	**0.49**	**0.51**				
	P(Rainy|Golf)	**0.75**	**0.05**				
	P(Sunny|Golf)	**0.25**	**0.95**				

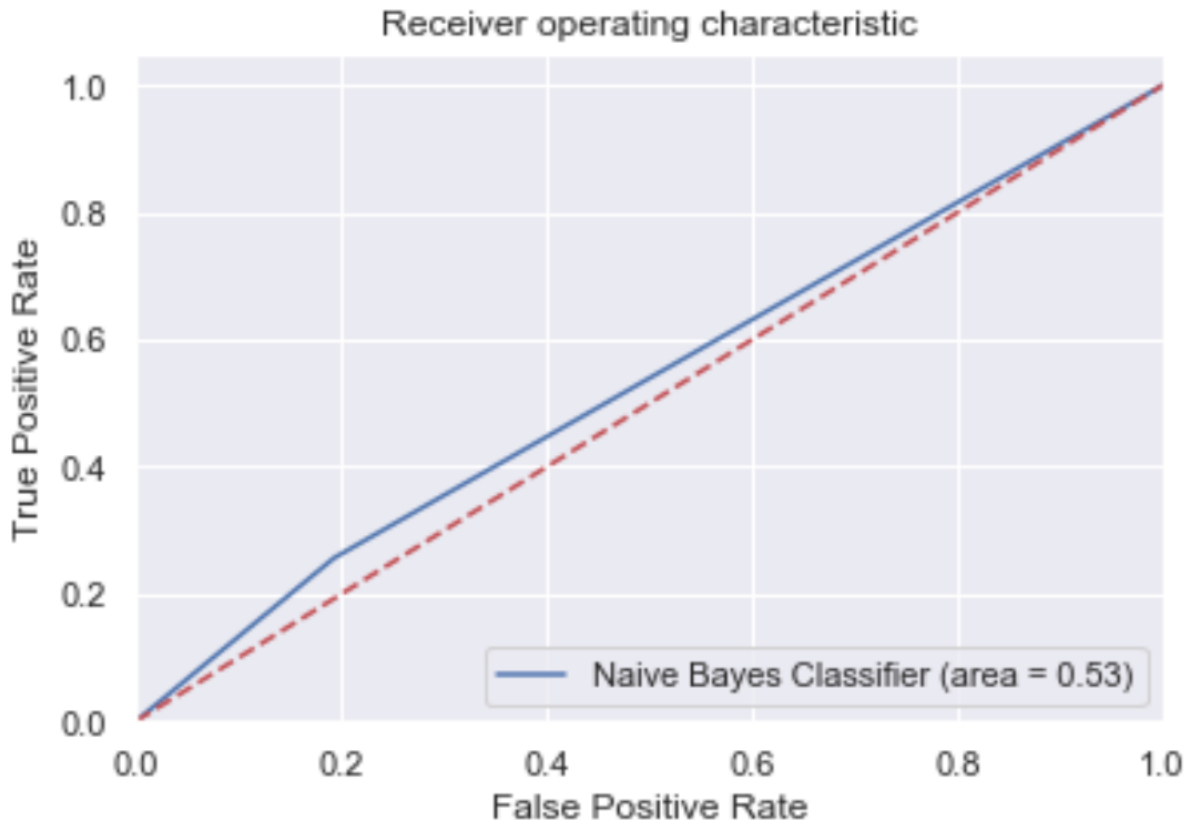
$$P(\text{Play Golf = Yes} \mid \text{Sunny}) = \frac{P(\text{Sunny}\mid \text{Played Golf}) * P(\text{Played Golf})}{P(\text{Sunny})} = \frac{0.95*0.51}{0.61} = 0.8$$

Predict = Will Play Golf!

$$P(\text{Play Golf = No} \mid \text{Sunny}) = \frac{P(\text{Sunny}\mid \text{Not Played Golf}) * P(\text{Not Played Golf})}{P(\text{Sunny})} = \frac{0.25*0.49}{0.61} = 0.2$$

SMART SOLUTIONS

4.e) Naïve Bayes (Result)



	precision	recall	f1-score
0	0.52	0.81	0.63
1	0.57	0.26	0.35

This model is not suitable as it performs worse than baseline for both F1 and AUC

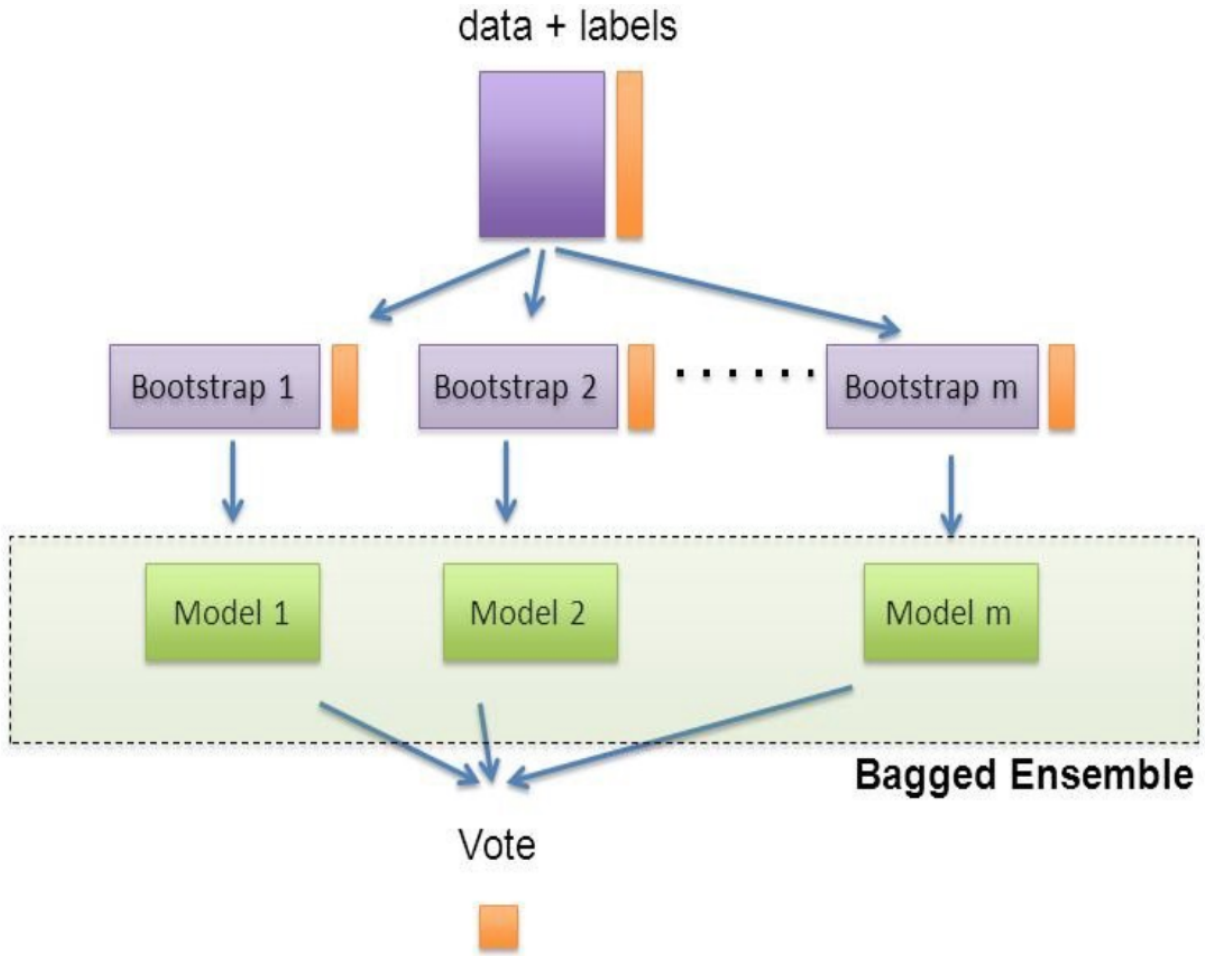
5. Ensemble Learning Models

Combining the predictive efforts of multiple models to reduce variance or bias and to power prediction capabilities is called ***Ensemble Learning****.*

There are 3 main types of Ensemble Learners that exist:

1. *Bagging (Bootstrap Aggregating)*
2. *Boosting*
3. *Stacking*

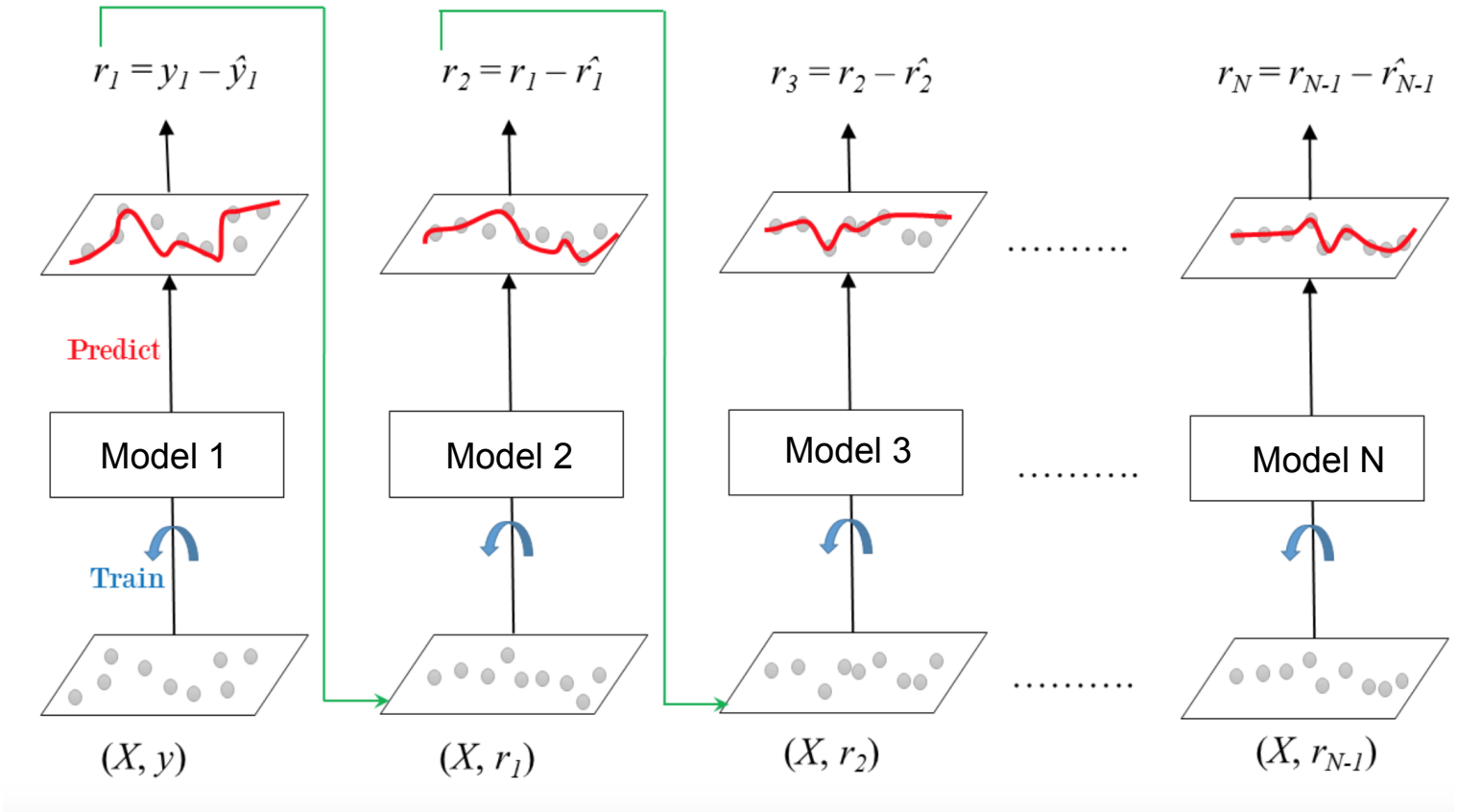
5.a) Bagging (Concept)



A technique combining multiple models in a **parallel structure** where every model is built **independent** of one another. This reduces variance and hence, minimizes overfitting.

Example: **Random Forest**, a bagging ensemble model for decision trees only.

5.b) Boosting (Concept)



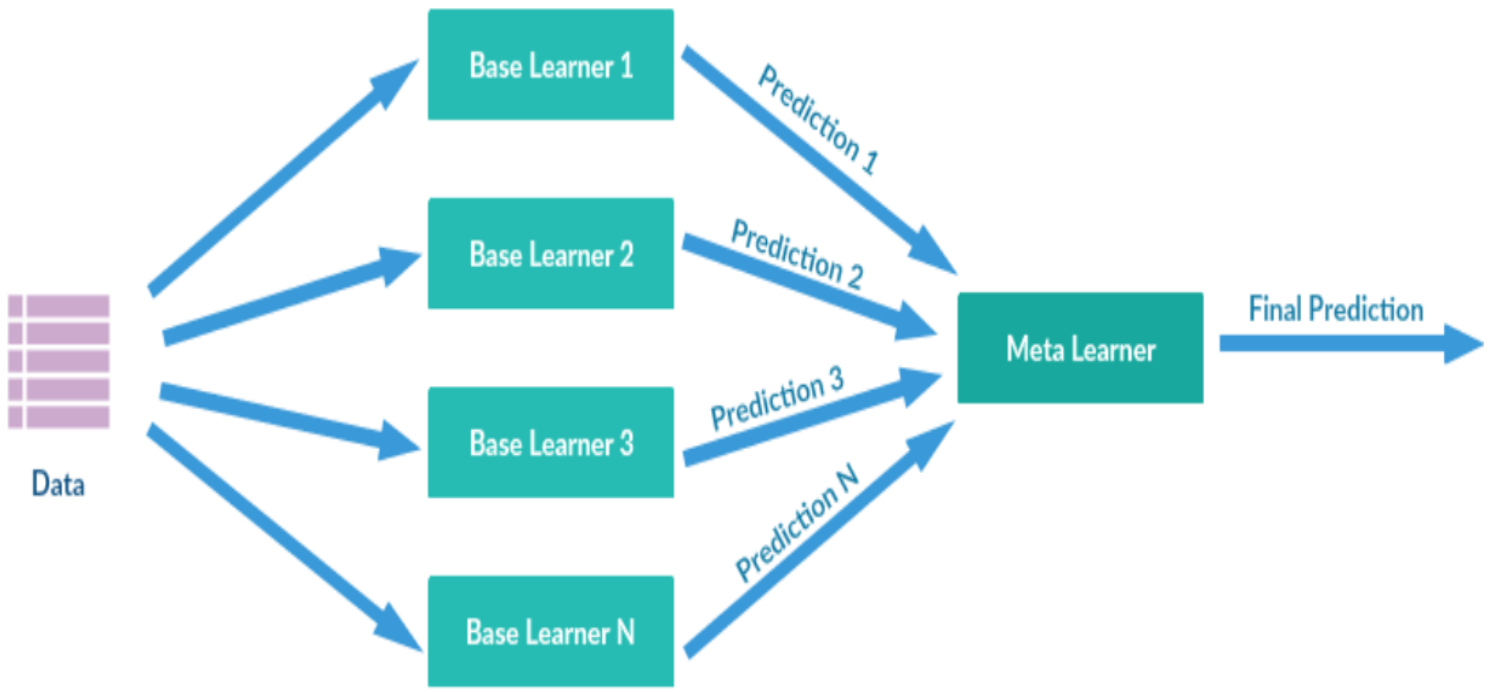
A technique combining multiple models in a **series/sequential structure** where every model is built **dependent** on the one before it.

The goal is to **re-train misclassified prediction** from the previous model evaluation. Hence, it reduces bias and minimizes underfitting.

No. of iteration is used defined.

SMART SOLUTIONS

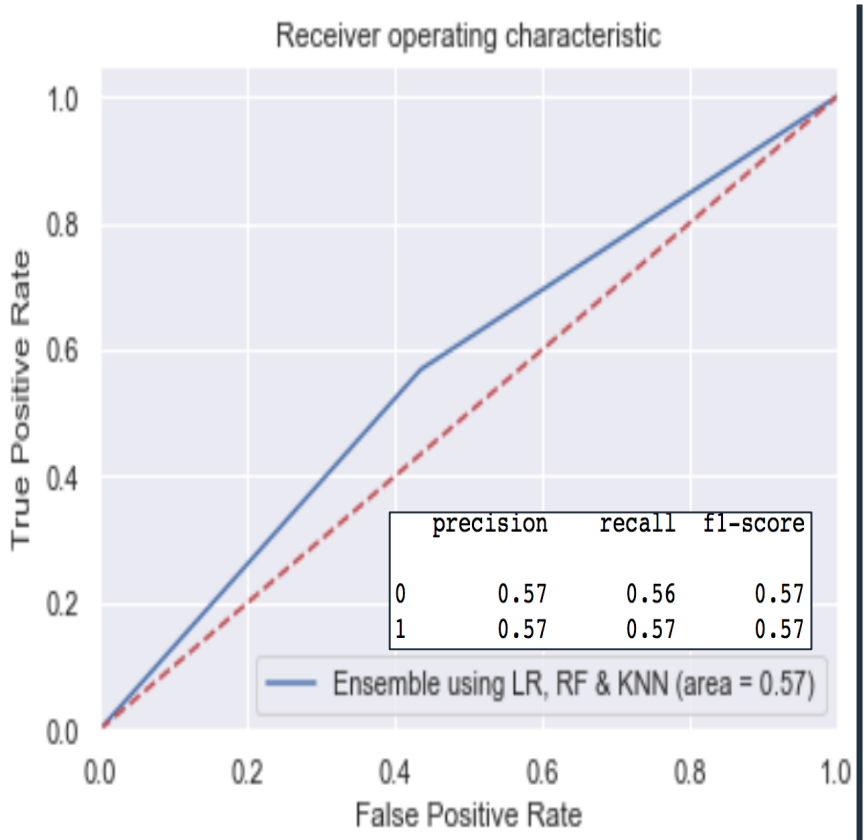
5.c) Stacking (Concept)



A technique combining multiple models in a **two-layered process** where 1st layer is bagging of predictions from different models and 2nd layer is using these **predictions as features** to train another model, “Meta Learner”.

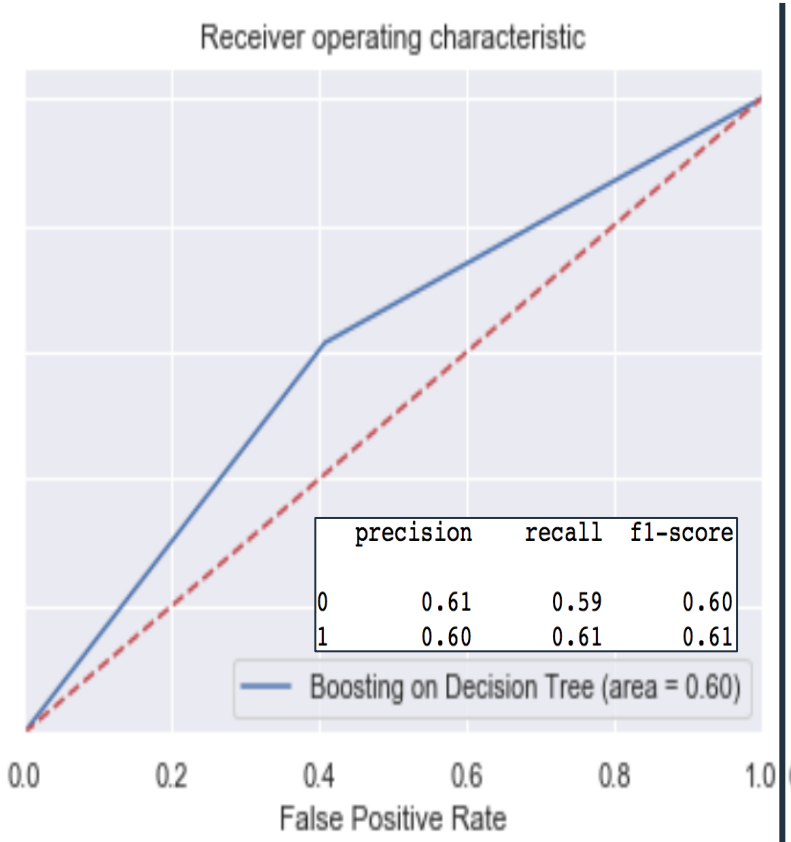
Prediction from “Meta Learner” is the final prediction. This architecture increases the prediction power in general.

5. Ensemble Learning Models (Results)



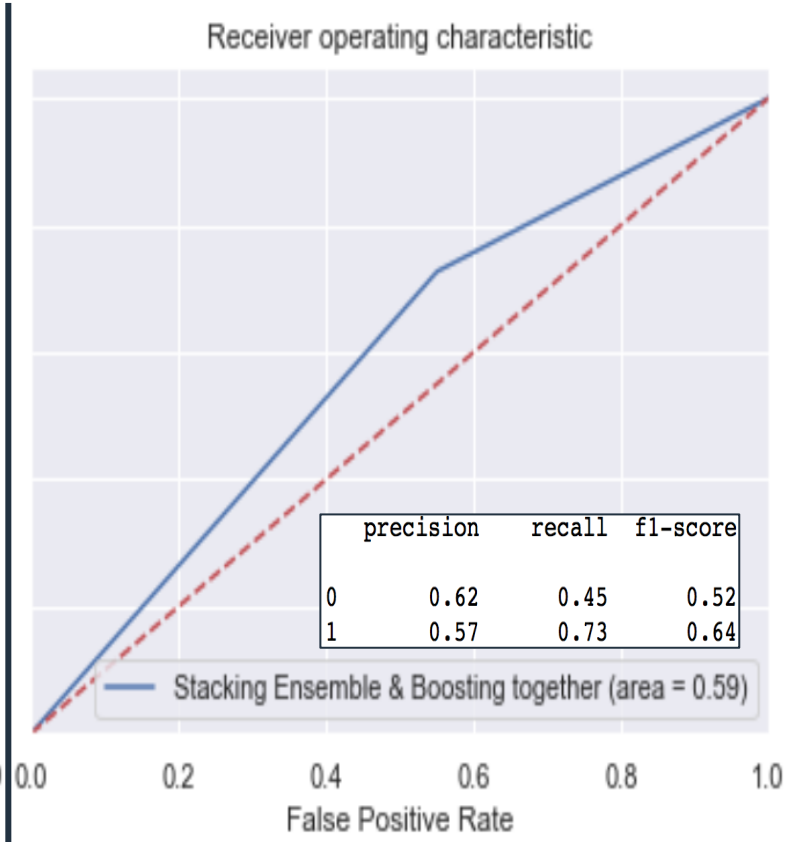
Bagging Model:

- Logistic Regression
- Random Forest (200 Trees)
- 9-Nearest Neighbors



Boosting Model:

- Base Model: Decision Tree (Max Depth = 3)
- Boost Model: Adaptive Boosting (iteration = 60)
- **Best AUC**

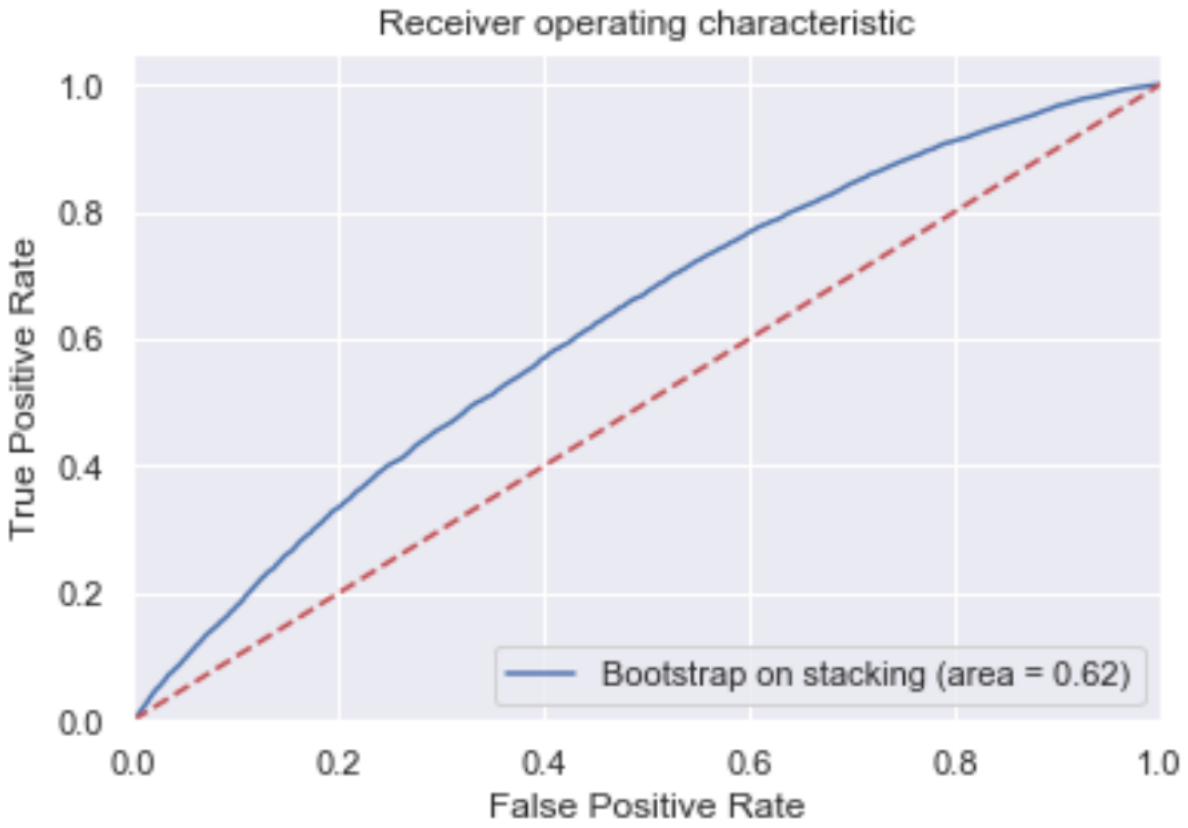


Stacking Model:

- Base Models: Bagging & Boosting Models (Left)
- Meta Learner: Logistic Regression
- **Best F1**

SMART SOLUTIONS

5. Ensemble Learning Models (Best Result)



	precision	recall	f1-score
0	0.67	0.27	0.39
1	0.54	0.87	0.67

Best result is generated when the **Stacking Model is trained from samples bootstrapped from the entire training dataset** (78.3% data)

200 stacking models were trained on unique bootstrapped sample set of 500 data (Labels stratified) each.

200 predictions were combined through averaging.

SMART SOLUTIONS

6. Conclusion

- Logistic Regression and Random Forest performed better than other basic supervised learning techniques.
- Boosting gave the best AUC while stacking gave the best F1 for ensemble model evaluation.
- Training Ensemble model like Stacking on bootstrapped of entire data showed significant increase in model performance and has potential to improve it even further with more tuning (i.e. increase sample size, increase number of models generated).
- Due to the complex nature of the dataset, application of advanced predictive modelling such as deep neural network can be useful (Note: Time complexity can increase also).

SMART SOLUTIONS

SMART SOLUTIONS