Loan Default Prediction with Explainable AI

How-to make effective Decisions using Machine Intelligence

Arian Mohammadi

Mina Mohammadi

Input Data

PREPROCESSING

EDA

Feature Selection

Evaluation

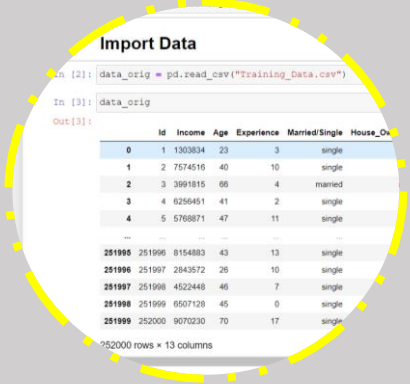
Modeling

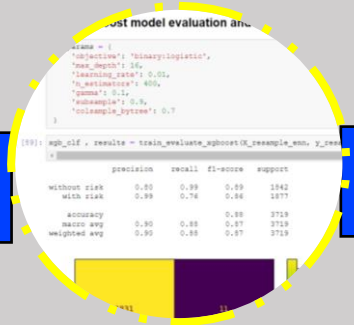
Finding the best model

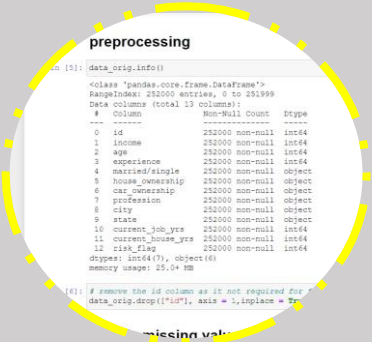
Feature importance

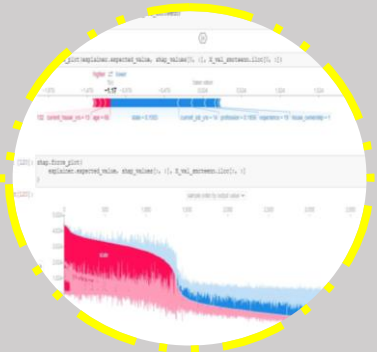
XAI

SHAP









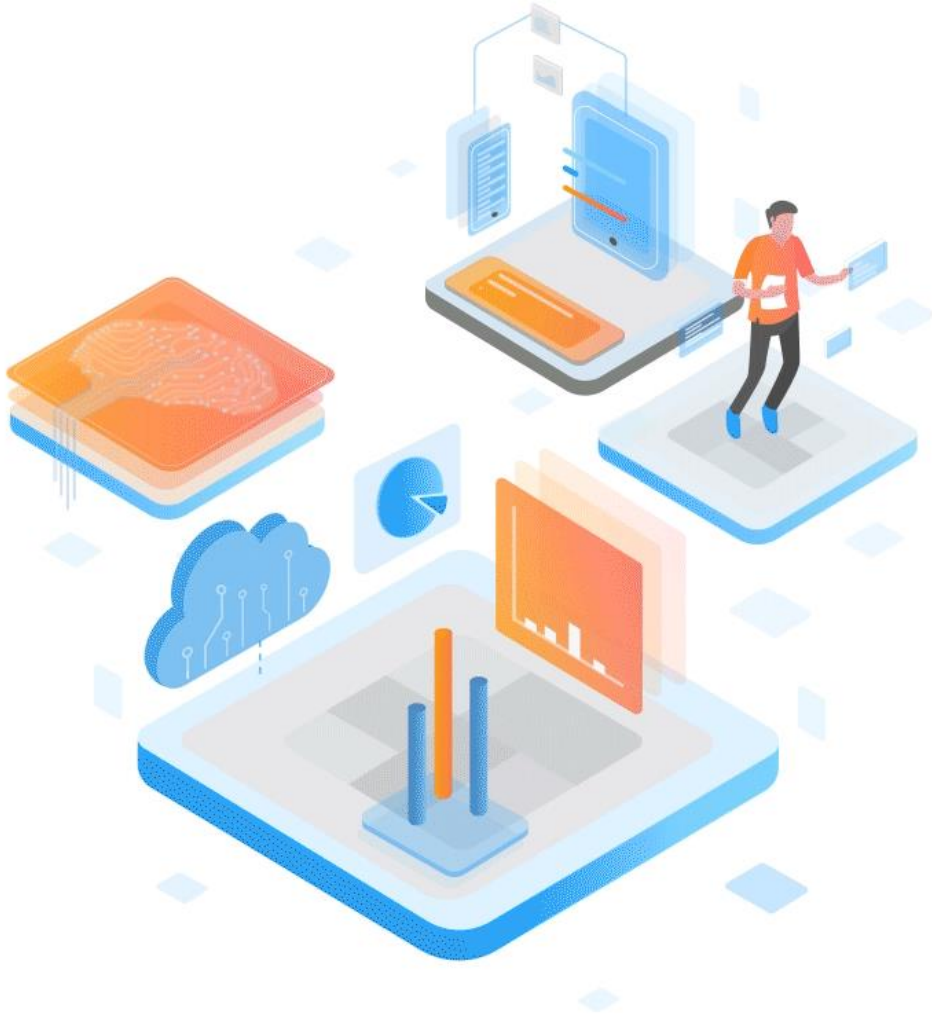
Data Understanding

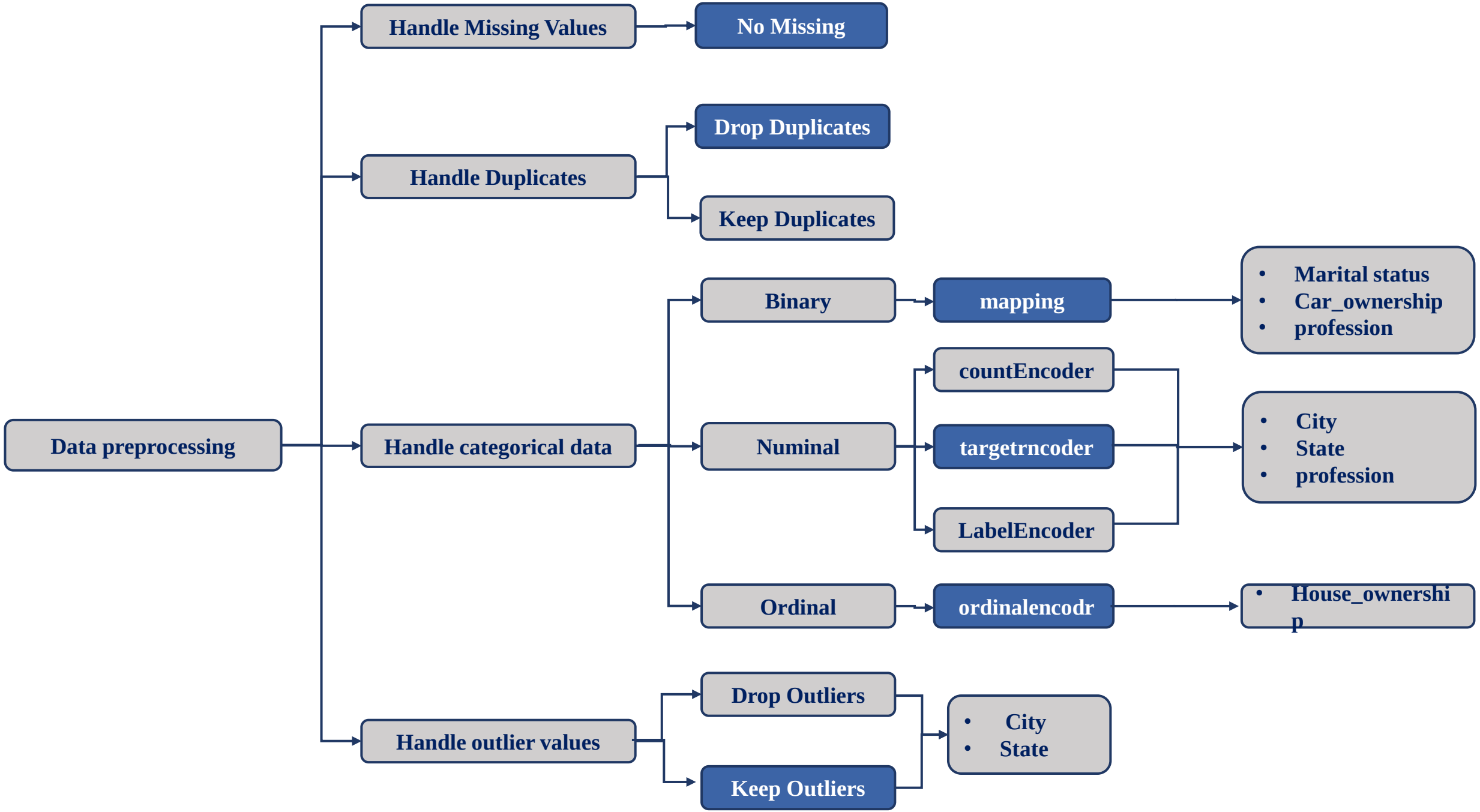
About DataSet

Summary	
shape	252000
NaN values	0
Duplicated values	208810
feature count	13
Categorical features	8
Problem type	Binary classification

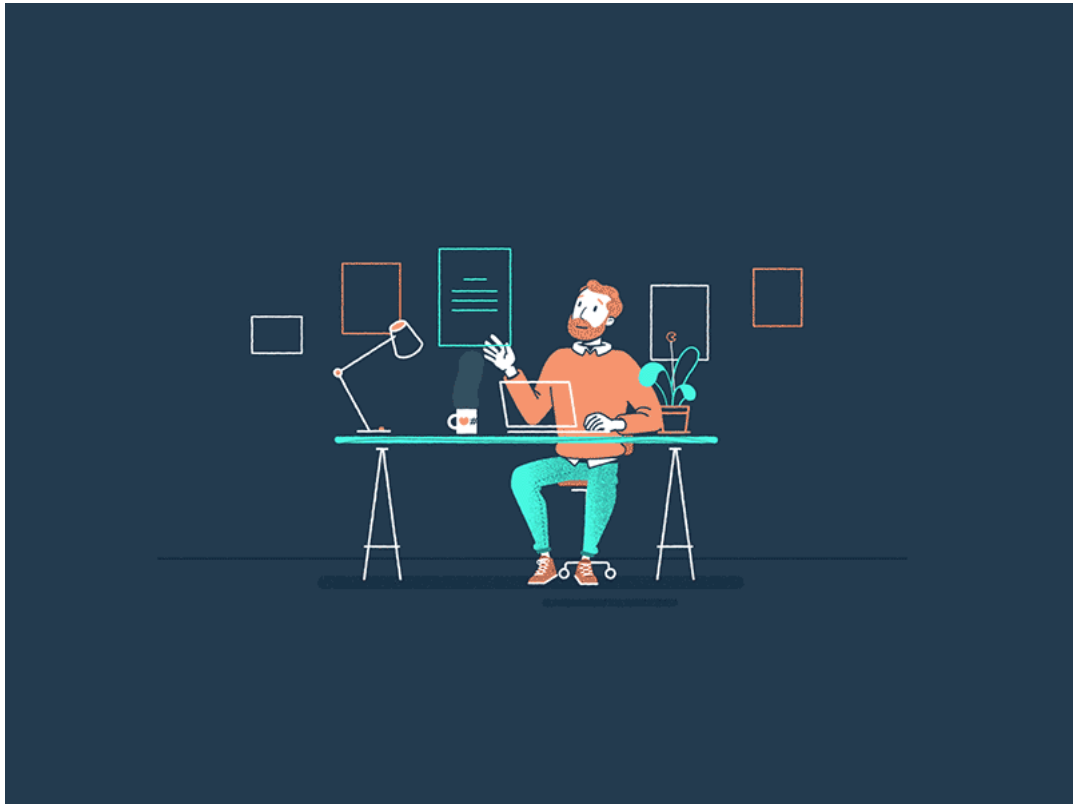
	Features	Description
1	income	Income of the user
2	age	Age of the user
3	experience	Professional experience of the user in years
4	profession	Profession
5	married	Whether married or single
6	house_ownership	Owned or rented or neither
7	car_ownership	Does the person own a car
8	current_job_years	Years of experience in the current job
9	current_house_years	Number of years in the current residence
10	city	City of residence
11	state	State of residence
12	risk_flag	Defaulted on a loan

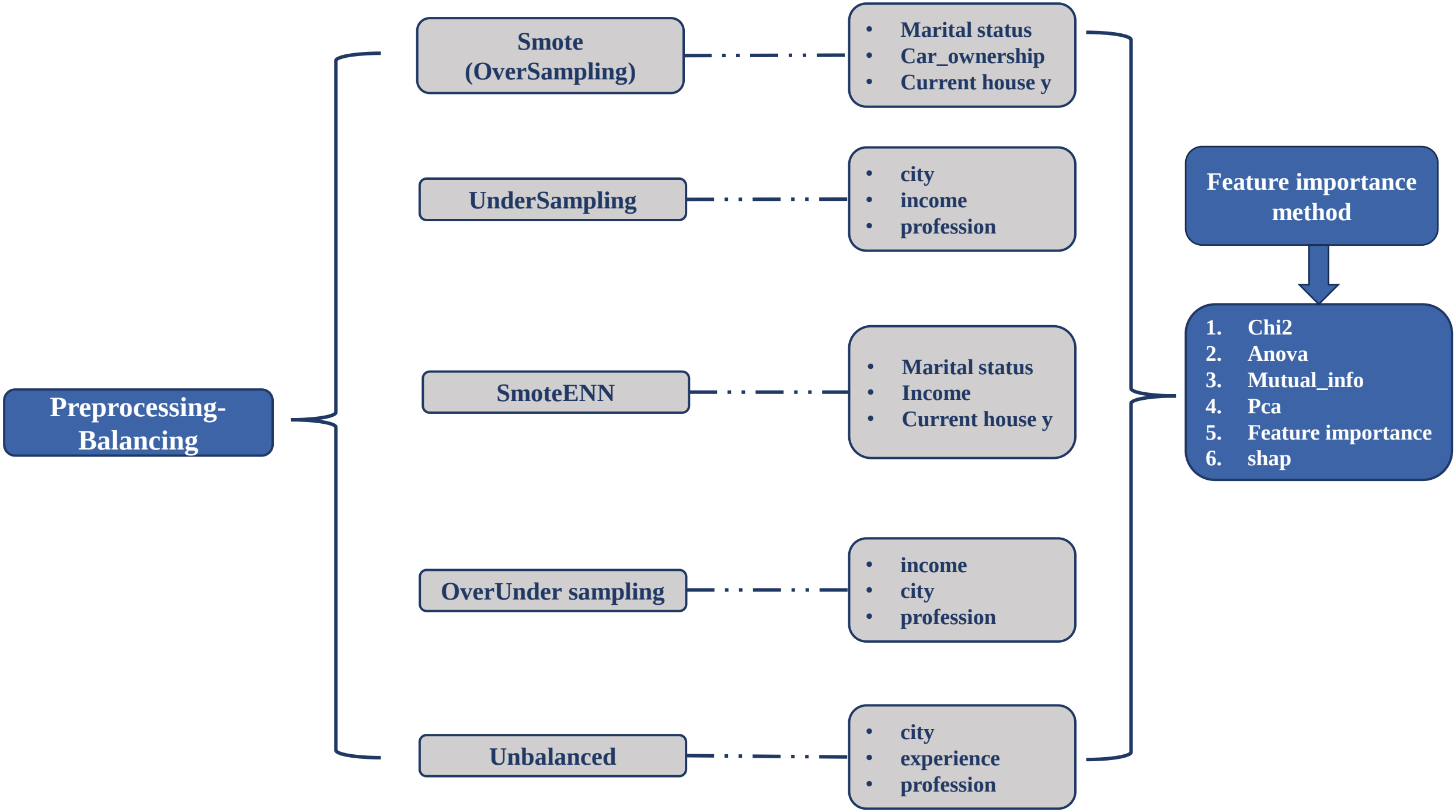
Data Preprocessing




Data preprocessing
Handle Missing Values
No Missing
Handle Duplicates
Drop Duplicates
Keep Duplicates
Handle categorical data
Binary
mapping
Marital status
Car_ownership
profession
Numinal
countEncoder
targetrncoder
LabelEncoder
City
State
profession
Ordinal
ordinalencodr
House_ownership
Handle outlier values
Drop Outliers
Keep Outliers
City
State

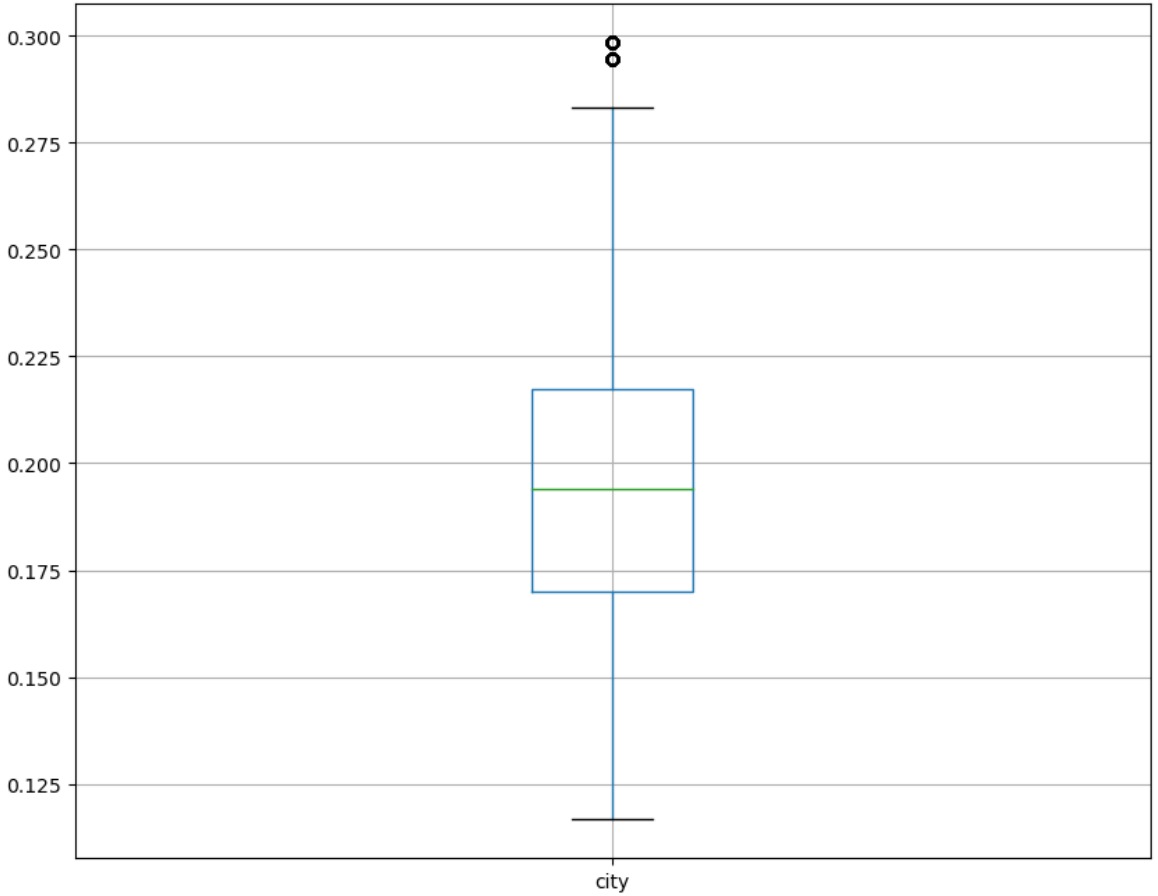
Sampling
+
Feature Engineering




Preprocessing-Balancing
Smote (OverSampling)
Marital status
Car_ownership
Current house y
UnderSampling
city
income
profession
SmoteENN
Marital status
Income
Current house y
OverUnder sampling
income
city
profession
Unbalanced
city
experience
profession
Feature importance method
1. Chi2
2. Anova
3. Mutual_info
4. Pca
5. Feature importance
6. shap

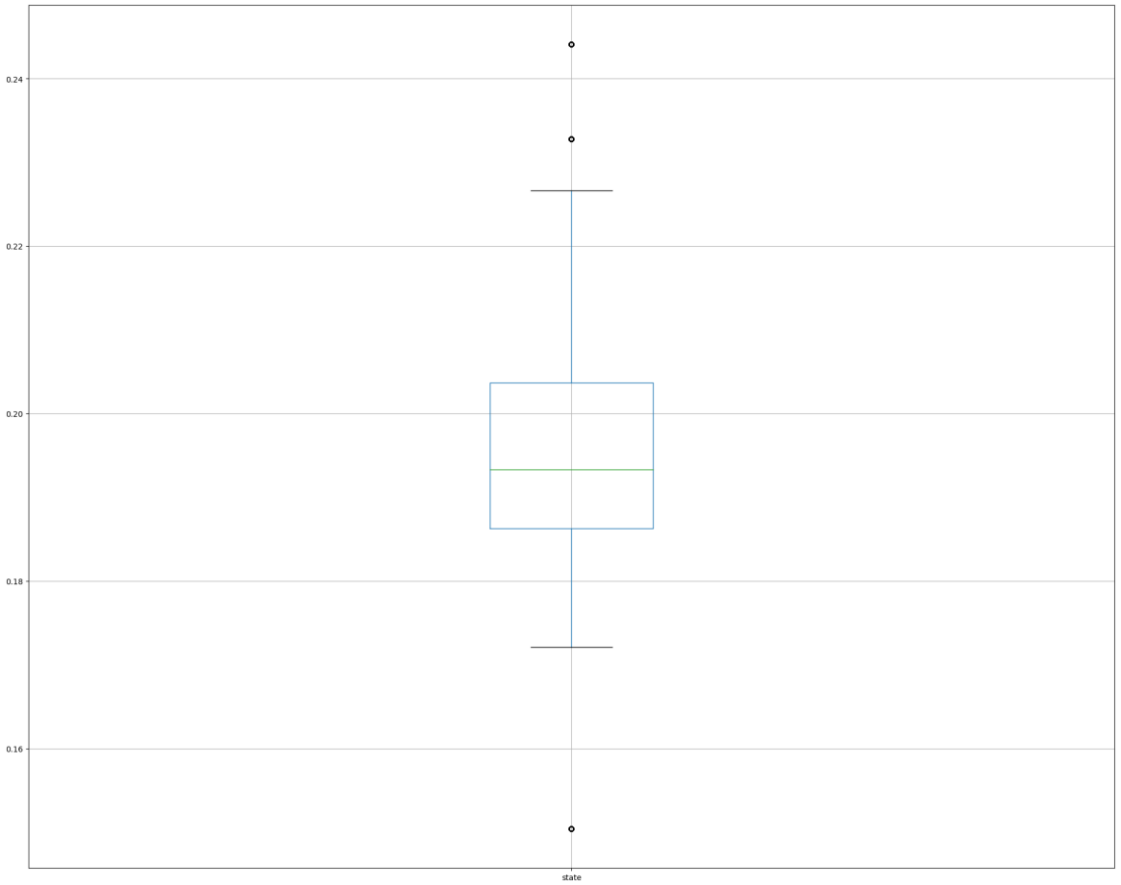
Exploratory Data Analysis

BoxPlot for OutLiers



['Bhubaneswar', 'Yamunanagar']

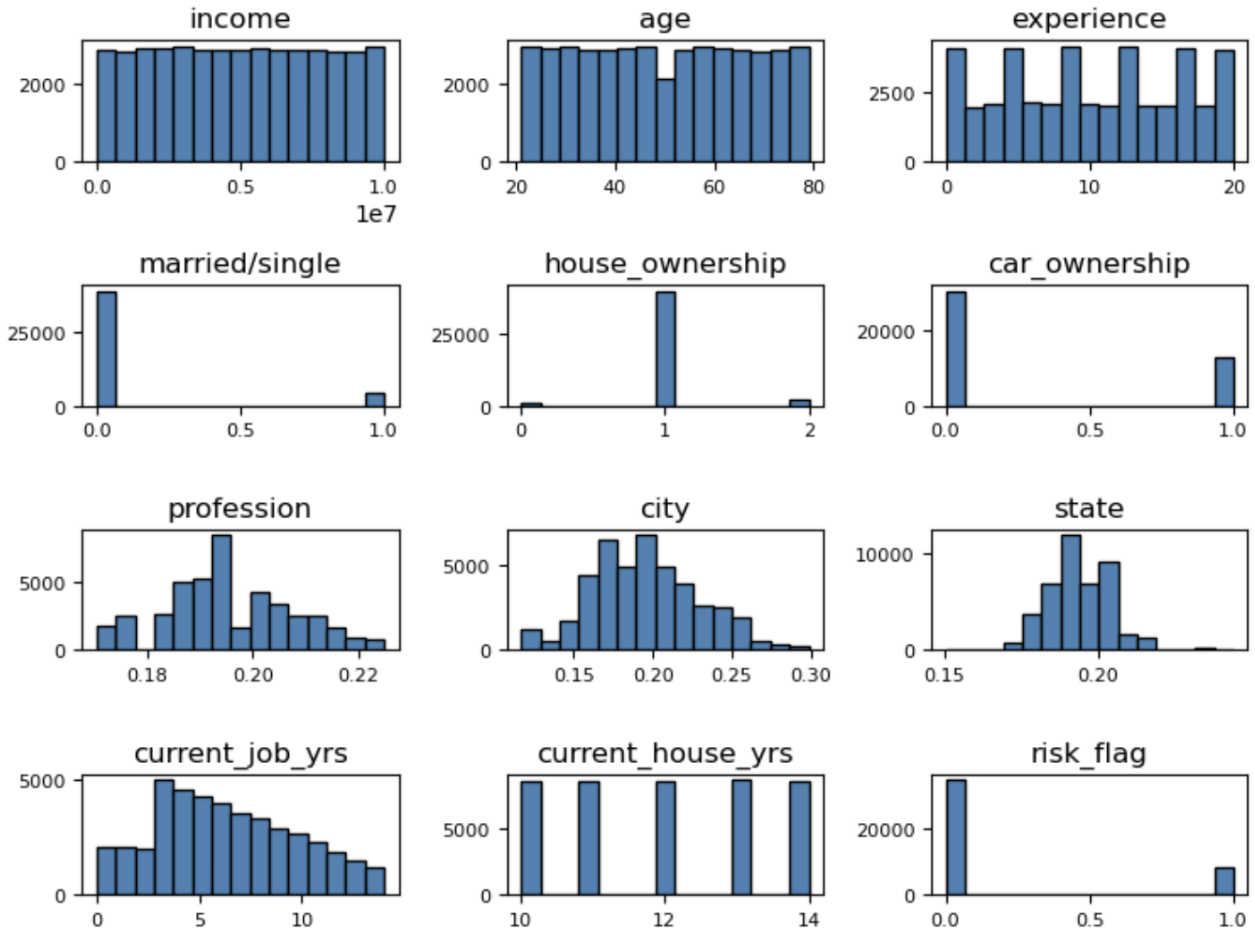
these 2 cities have more records with a target label of class 1



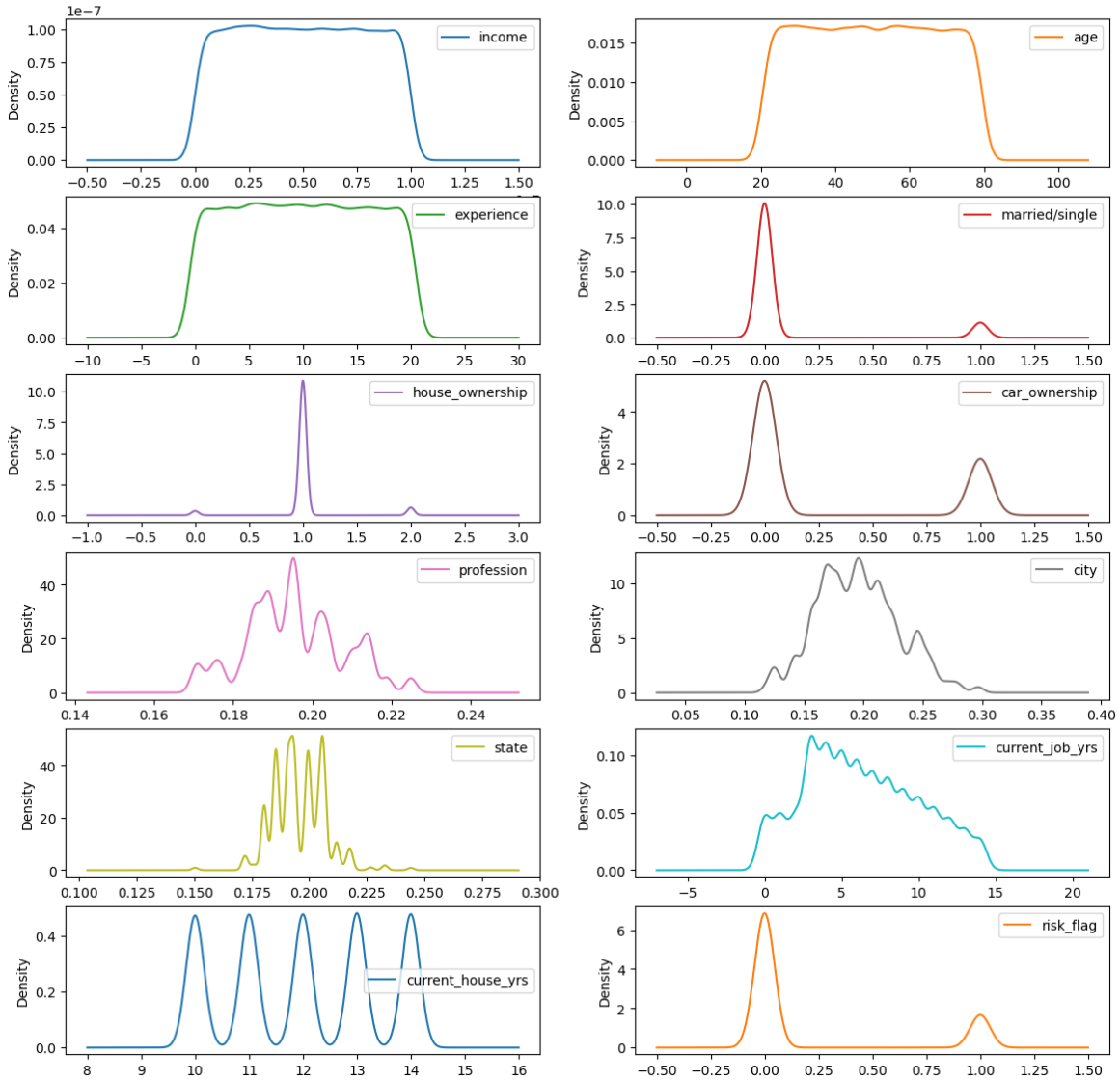
Up:['Puducherry', 'Manipur']
Down : ['Mizoram']

Up : these 2 states have more records with a target label of class 1

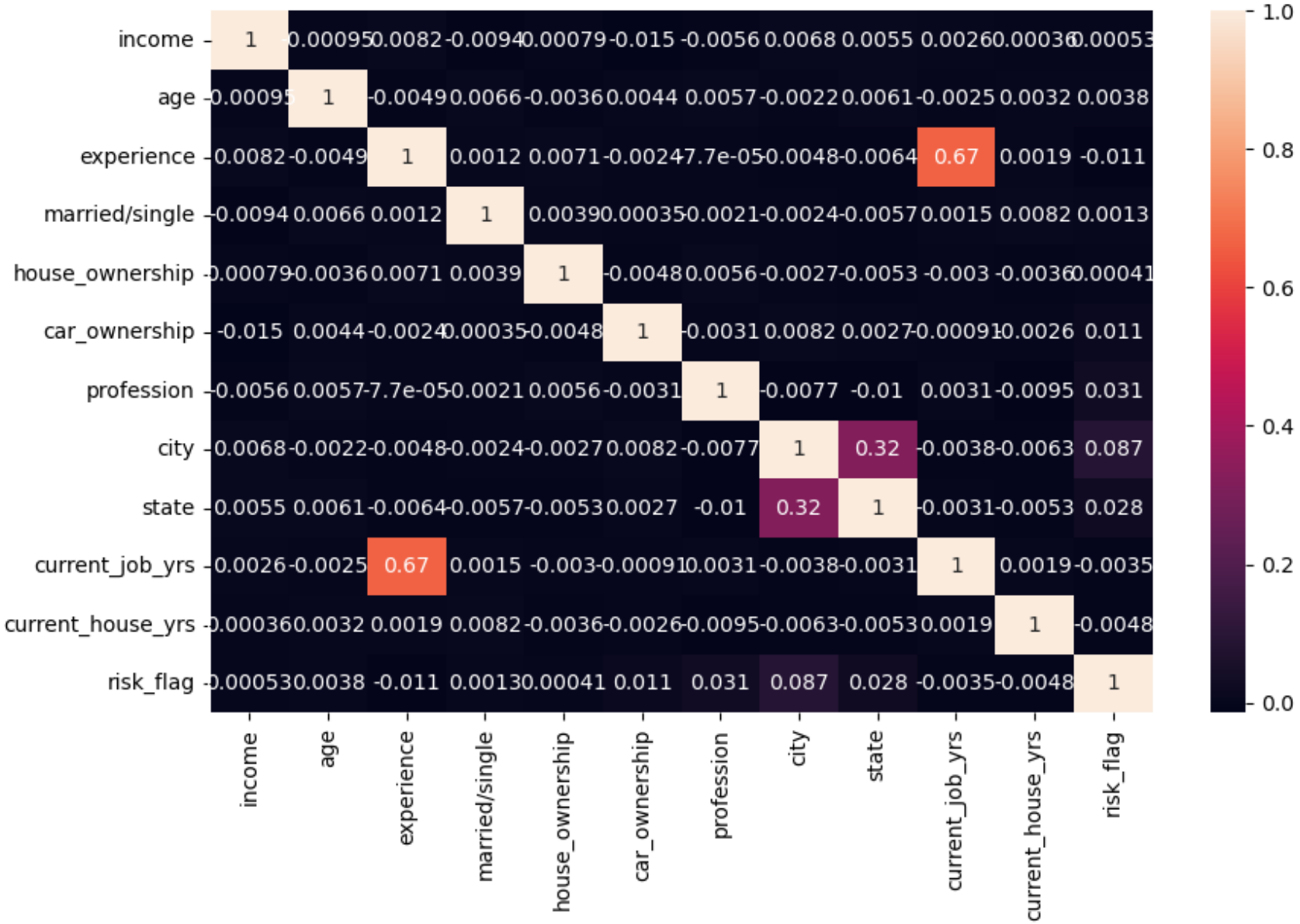
Comprehensive scheme of data distribution



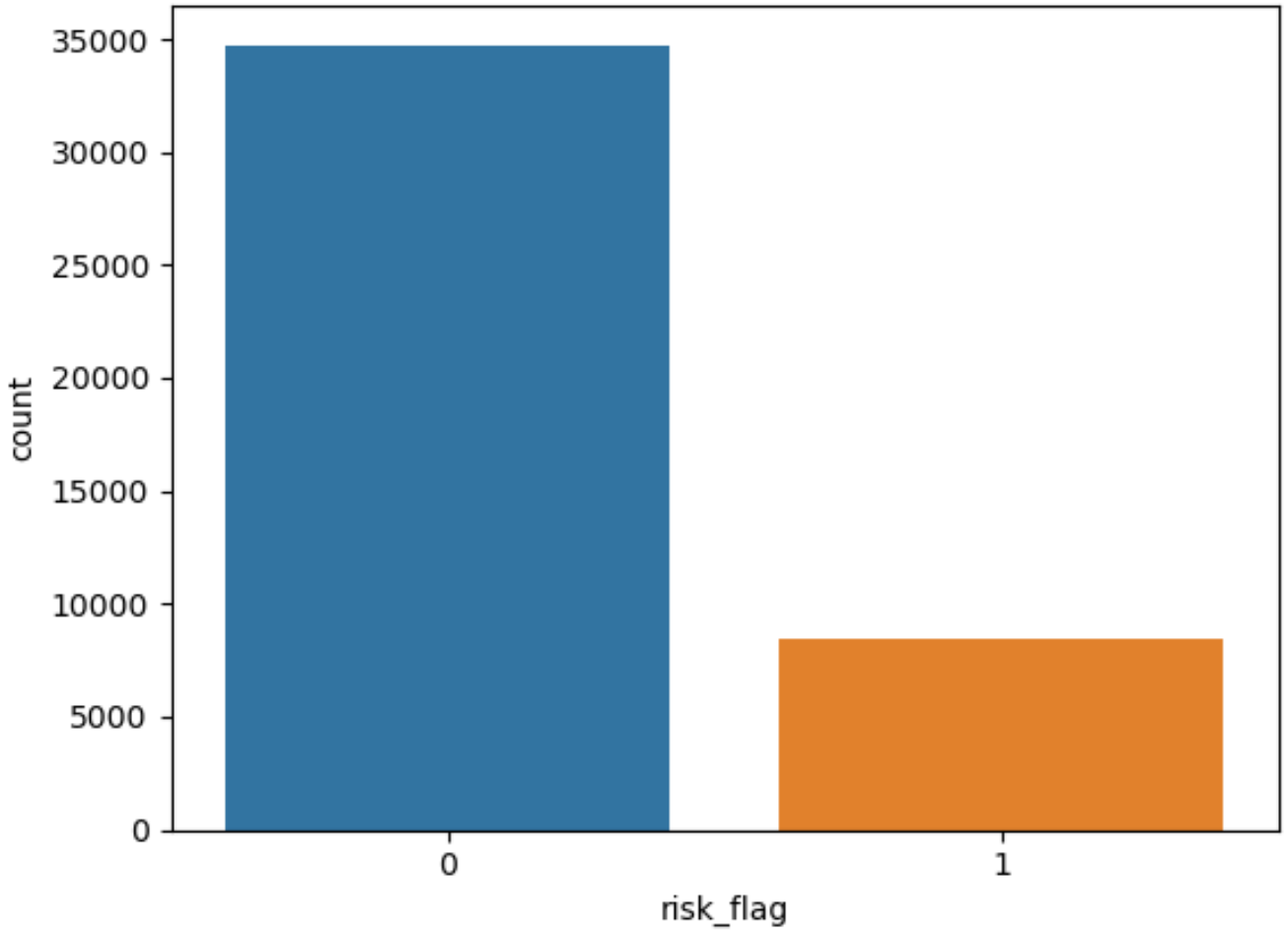
Density plot of Numerical features



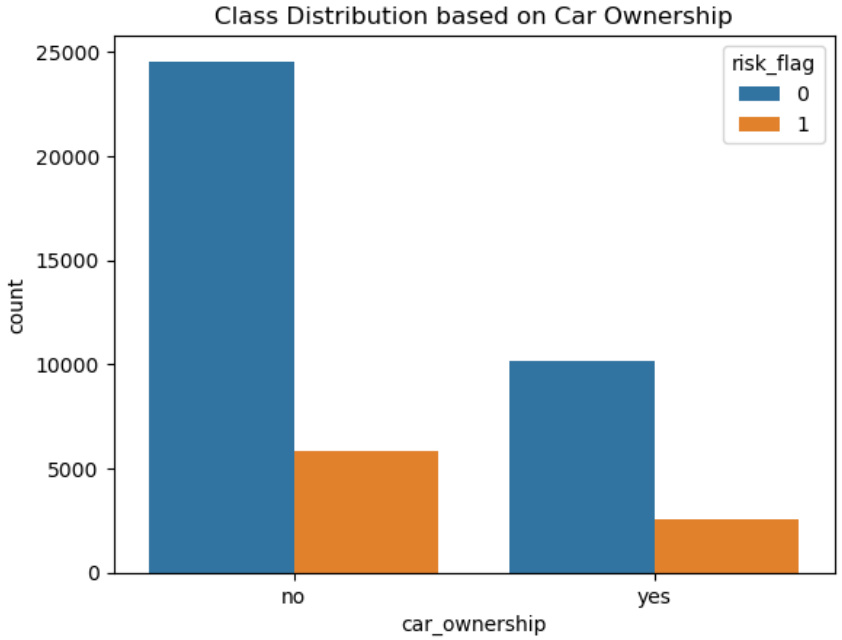
correlation

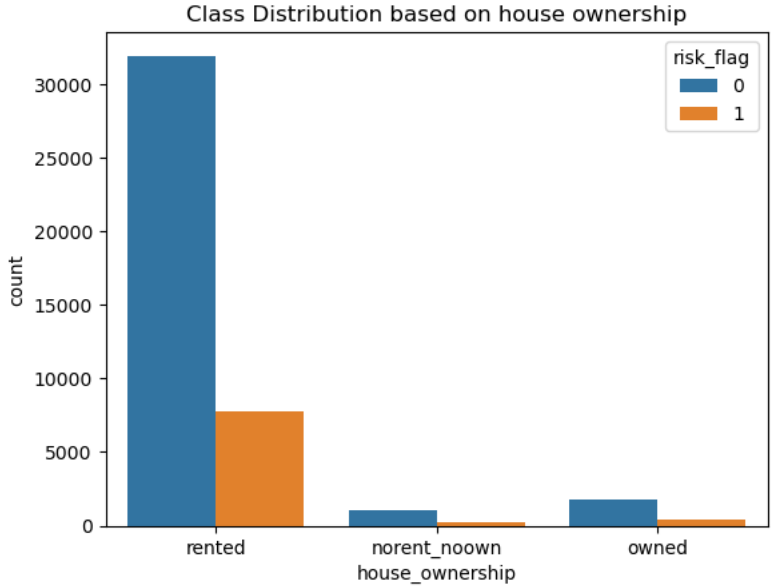


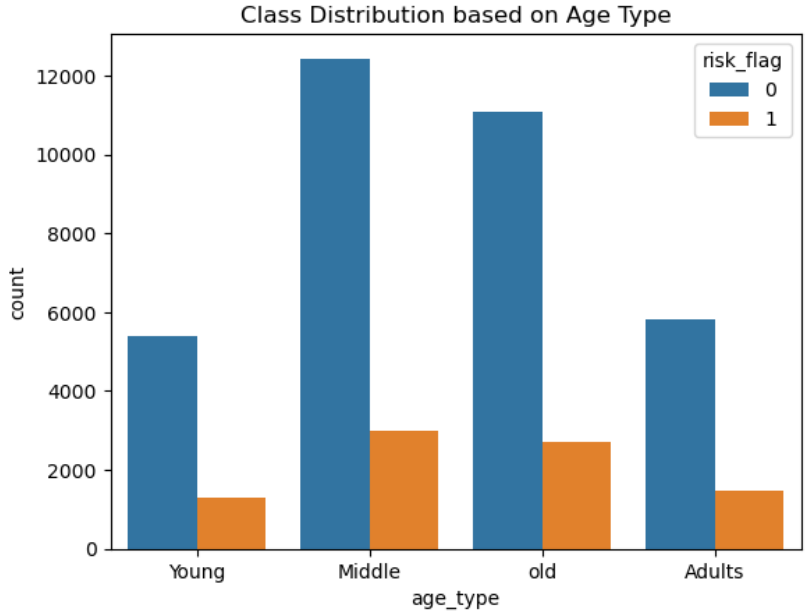
Risk Flag (Target Features)

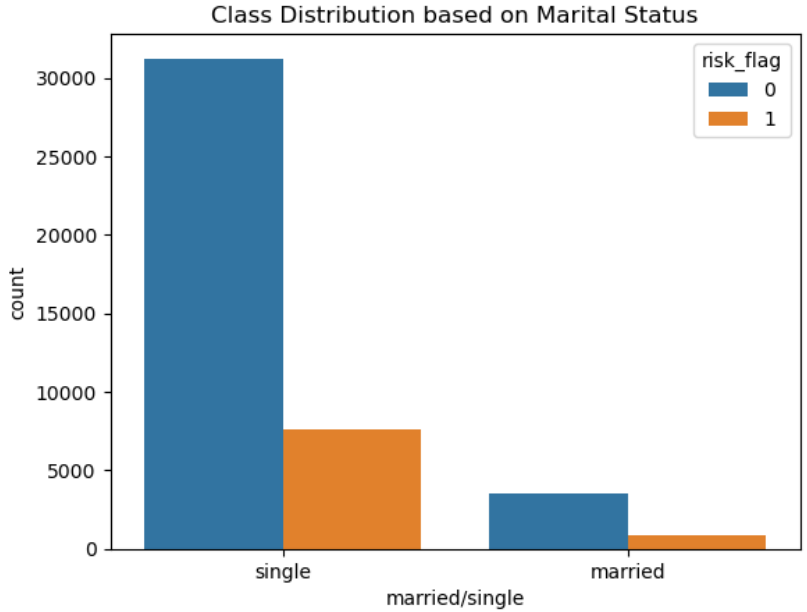


Class distribution besed on features

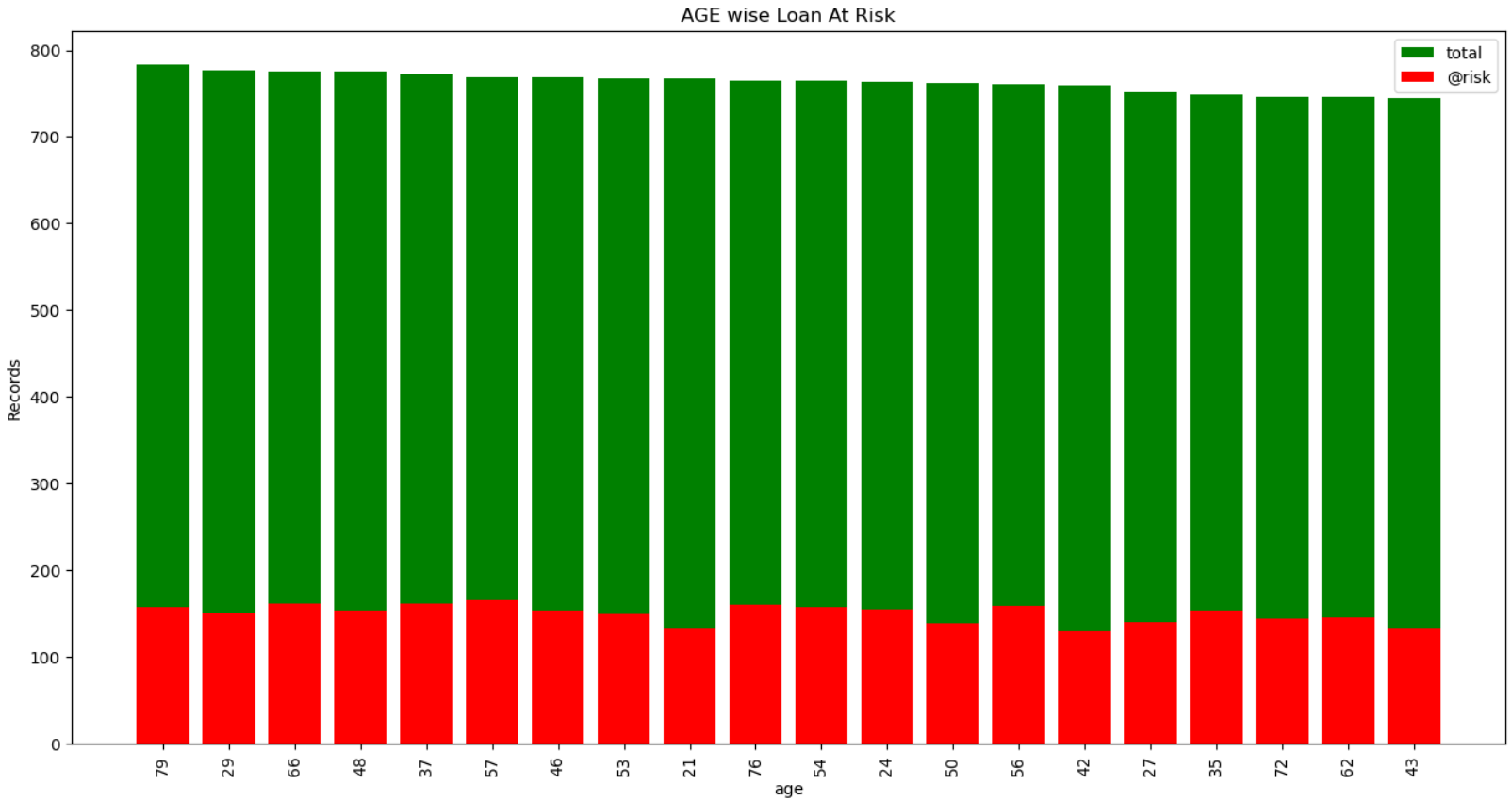








Age



Top 5 with LeasT Default Rate: AGE

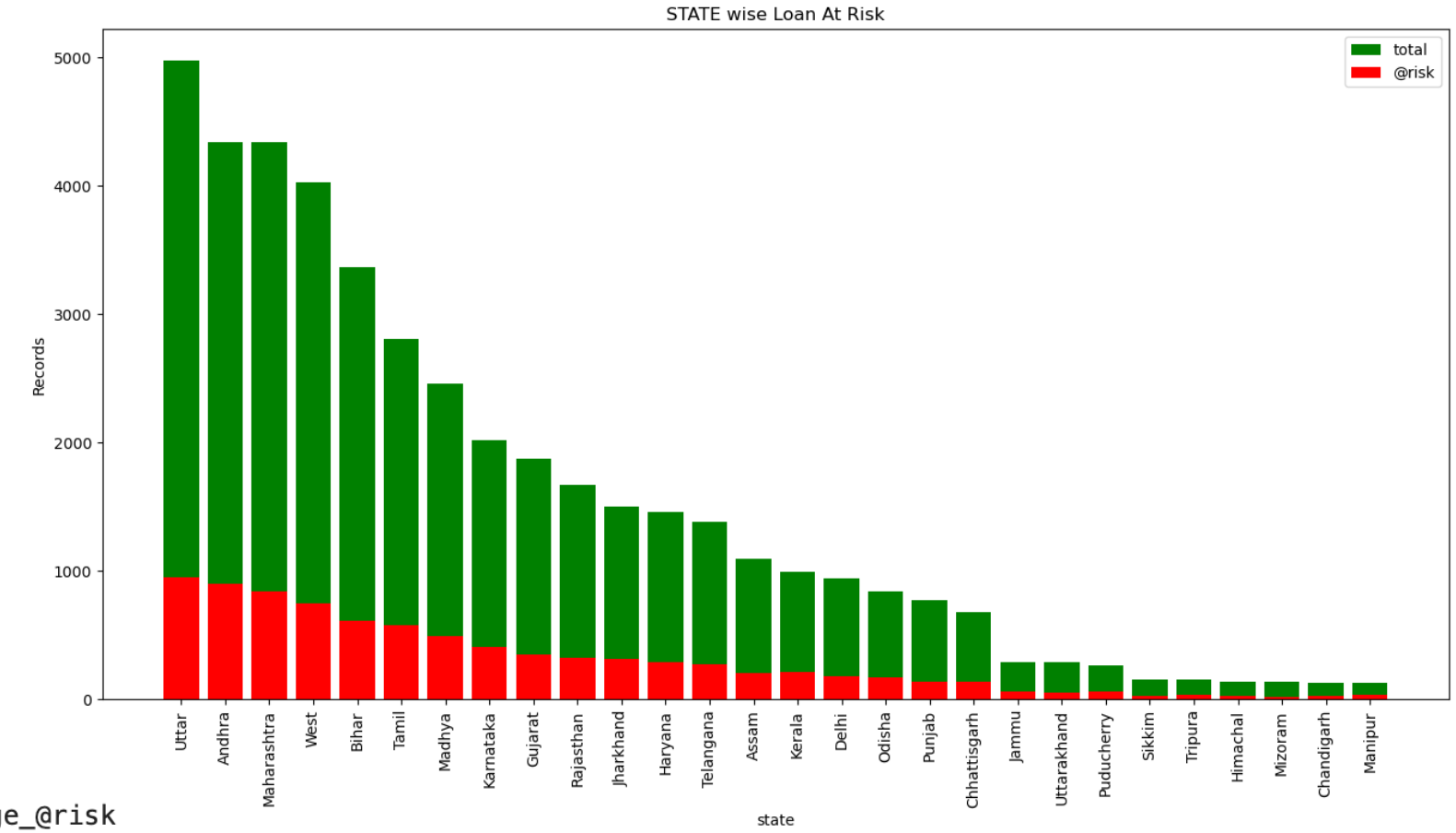
	age	risk_flag	total_records	Average_@risk
21	42	130	759	0.171278
0	21	134	767	0.174707
22	43	134	745	0.179866
29	50	139	762	0.182415
6	27	140	752	0.186170

==========

Top 5 with Most Default Rate: AGE

	age	risk_flag	total_records	Average_@risk
16	37	161	773	0.208279
45	66	162	776	0.208763
35	56	159	761	0.208936
55	76	160	765	0.209150
36	57	165	769	0.214564

State



Top 5 with LeasT Default Rate: STATE

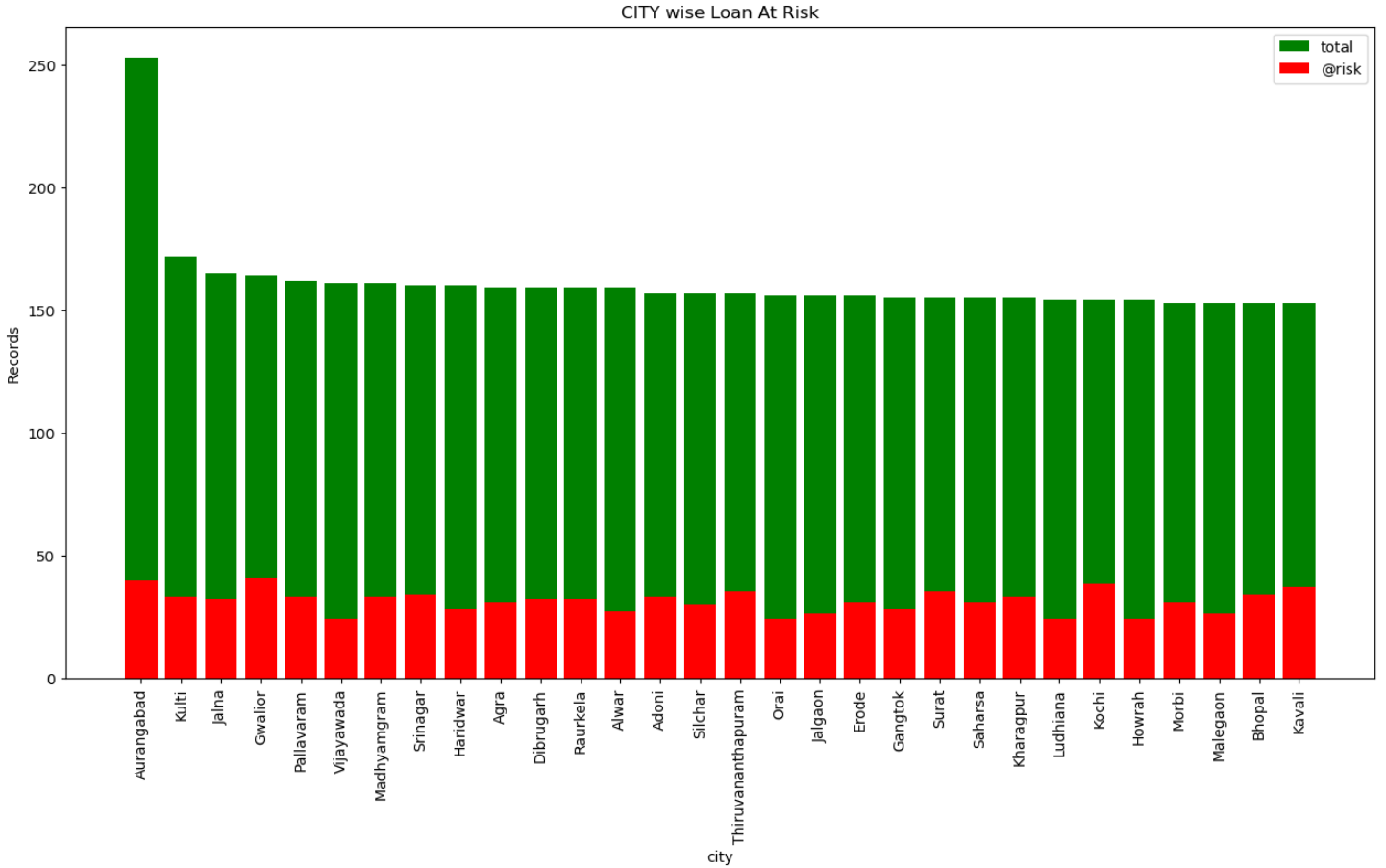
	state	risk_flag	total_records	Average_@risk
16	Mizoram	20	133	0.150376
19	Punjab	133	773	0.172057
26	Uttarakhand	51	290	0.175862
2	Bihar	607	3365	0.180386
21	Sikkim	28	155	0.180645

==========

Top 5 with Most Default Rate: STATE

	state	risk_flag	total_records	Average_@risk
9	Jammu	63	291	0.216495
12	Kerala	216	992	0.217742
24	Tripura	34	150	0.226667
18	Puducherry	61	262	0.232824
15	Manipur	31	127	0.244094

City



Top 5 with LeasT Default Rate: CITY

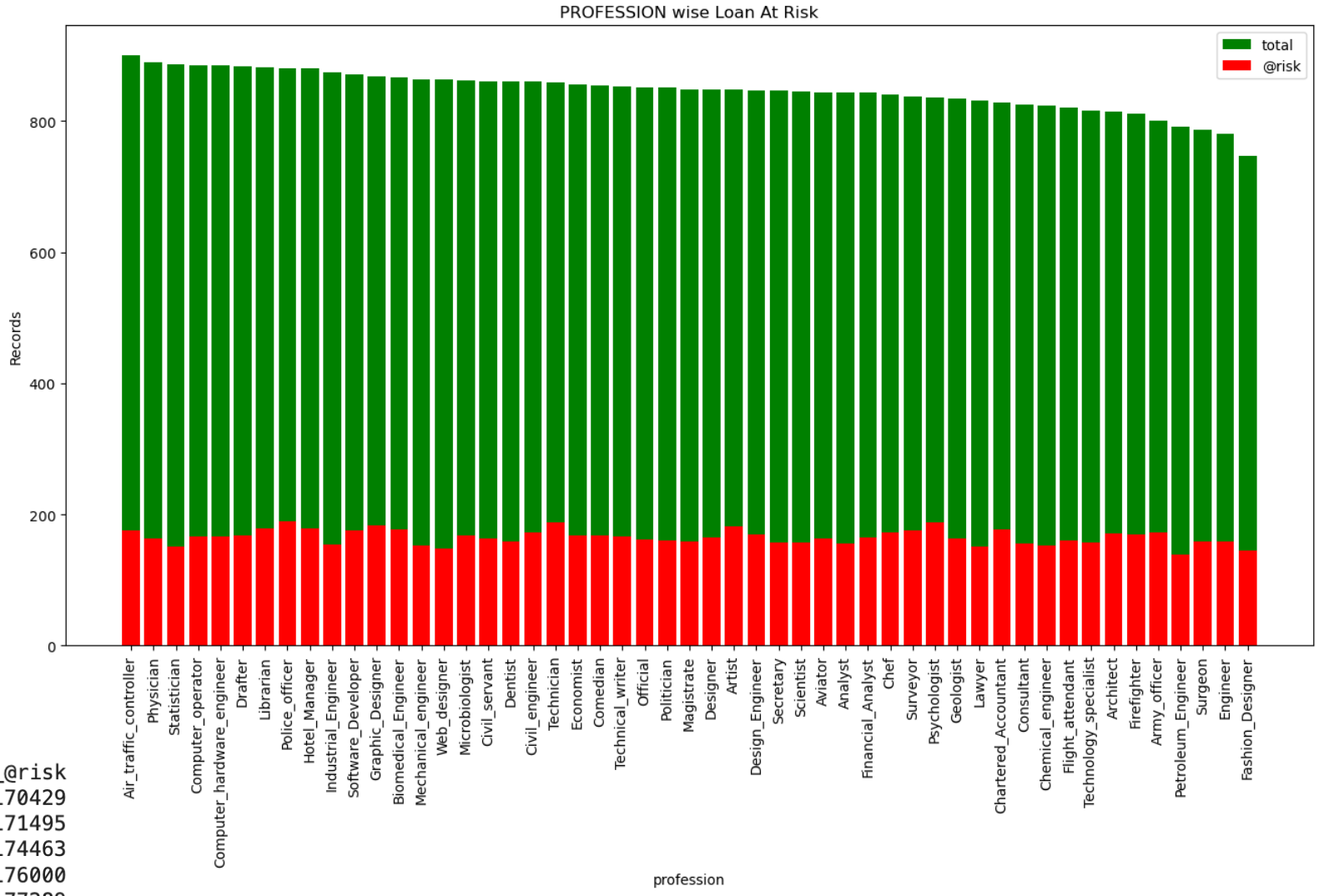
	city	risk_flag	total_records	Average_@risk
312	Vijayawada	24	161	0.149068
218	Orai	24	156	0.153846
120	Howrah	24	154	0.155844
173	Ludhiana	24	154	0.155844
24	Aurangabad	40	253	0.158103

==========

Top 5 with Most Default Rate: CITY

	city	risk_flag	total_records	Average_@risk
291	Thiruvananthapuram	35	157	0.222930
282	Surat	35	155	0.225806
152	Kavali	37	153	0.241830
159	Kochi	38	154	0.246753
111	Gwalior	41	164	0.250000

profession



Top 5 with LeasT Default Rate: PROFESSION

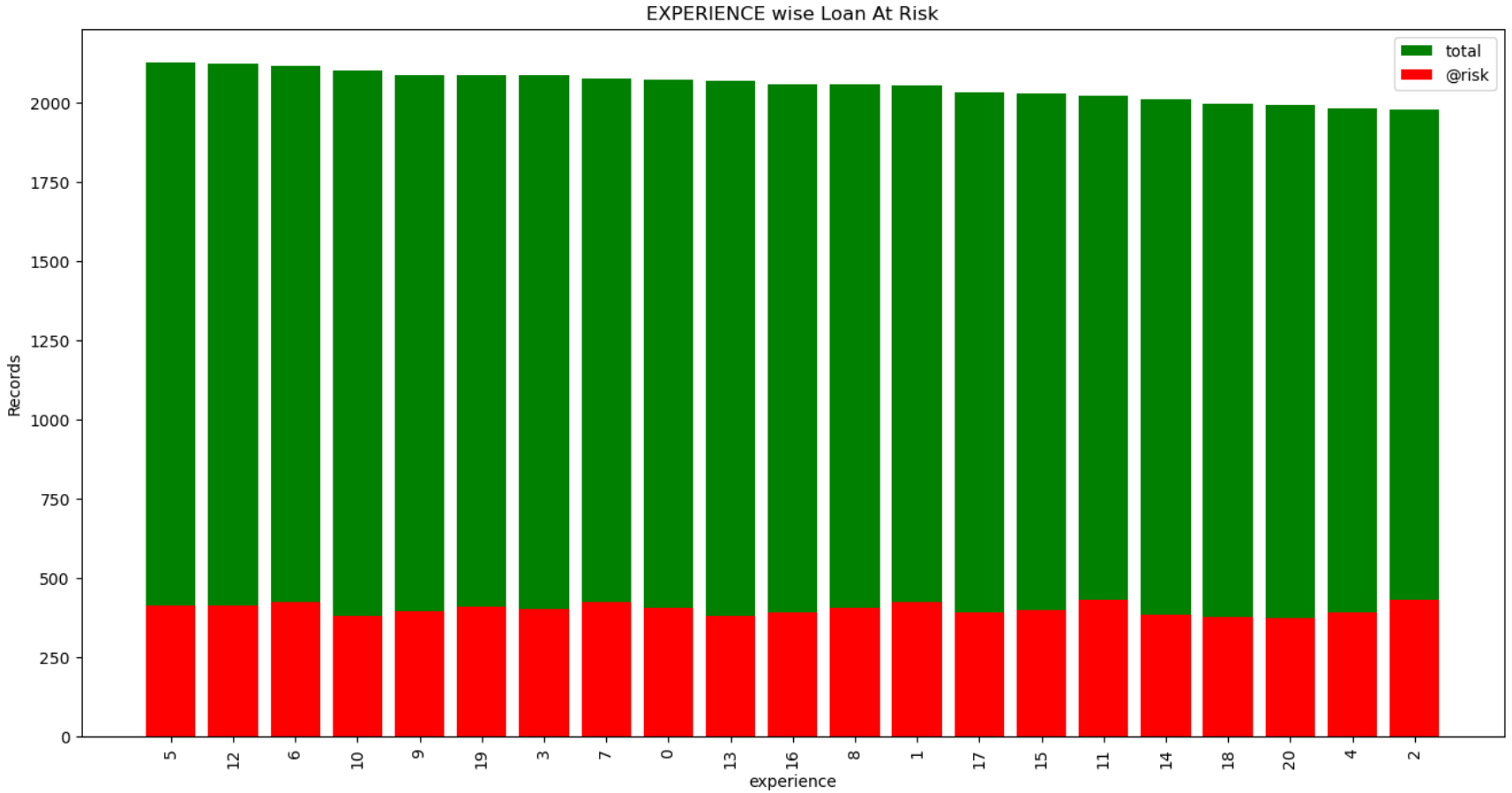
	profession	risk_flag	total_records	Average_@risk
44	Statistician	151	886	0.170429
50	Web_designer	148	863	0.171495
36	Petroleum_Engineer	138	791	0.174463
29	Industrial_Engineer	154	875	0.176000
33	Mechanical_engineer	153	863	0.177289

==========

Top 5 with Most Default Rate: PROFESSION

	profession	risk_flag	total_records	Average_@risk
7	Chartered_Accountant	177	828	0.213768
38	Police_officer	189	881	0.214529
3	Army_officer	172	801	0.214732
48	Technician	188	859	0.218859
40	Psychologist	188	836	0.224880

Experience



Top 5 with LeasT Default Rate: EXPERIENCE

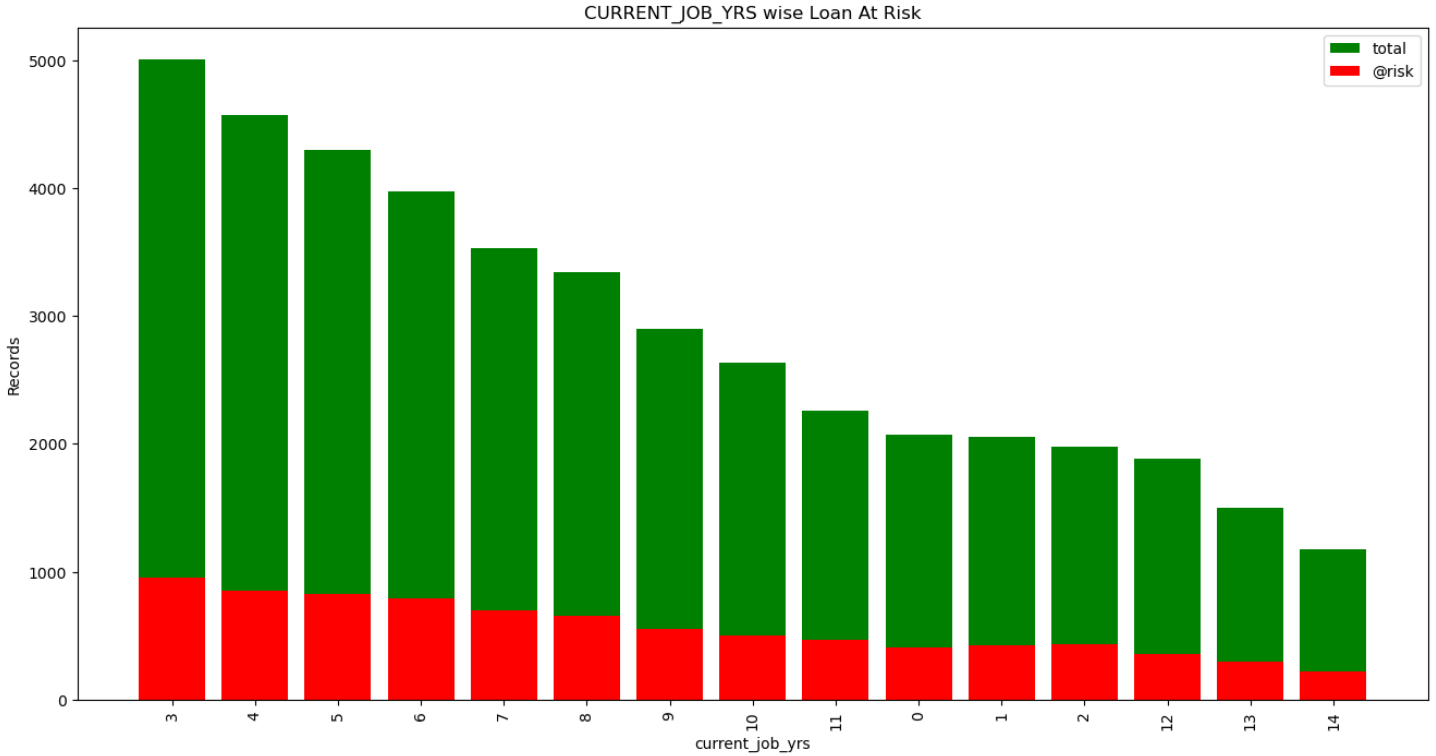
	experience	risk_flag	total_records	Average_@risk
10	10	380	2103	0.180694
13	13	380	2069	0.183664
20	20	372	1994	0.186560
18	18	376	1999	0.188094
16	16	390	2061	0.189229

==========

Top 5 with Most Default Rate: EXPERIENCE

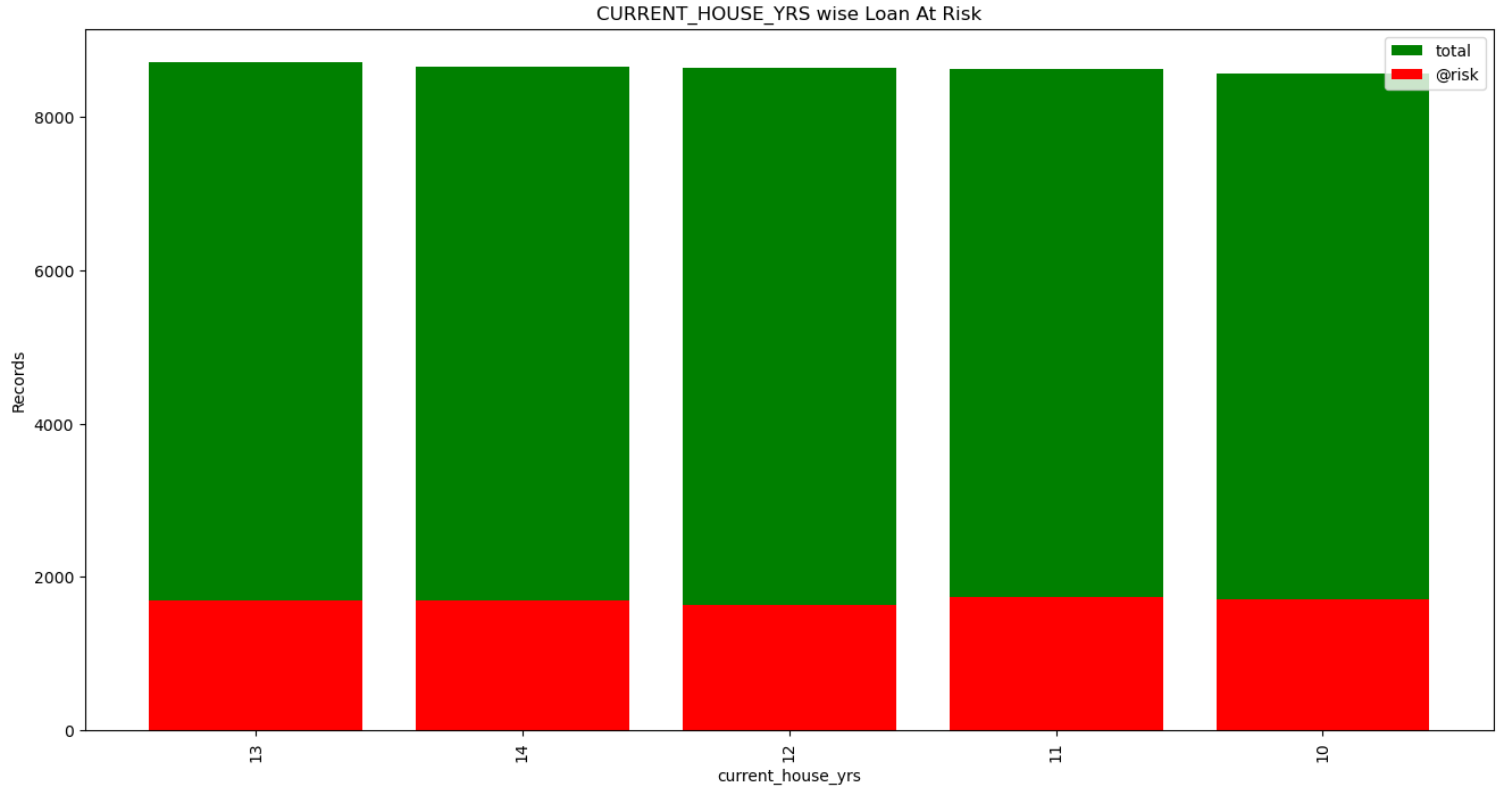
	experience	risk_flag	total_records	Average_@risk
6	6	423	2119	0.199622
7	7	424	2076	0.204239
1	1	424	2054	0.206426
11	11	430	2023	0.212556
2	2	431	1980	0.217677

Current job yrs




```
Top 5 with LeasT Default Rate: CURRENT_JOB_YRS
    current_job_yrs  risk_flag  total_records  Average_@risk
4                 4        854           4571       0.186830
14               14        222           1176       0.188776
10               10        499           2637       0.189230
12               12        357           1883       0.189591
9                 9        551           2904       0.189738
================================================
Top 5 with Most Default Rate: CURRENT_JOB_YRS
    current_job_yrs  risk_flag  total_records  Average_@risk
7                 7        702           3535       0.198586
6                 6        794           3973       0.199849
11               11        465           2262       0.205570
1                 1        424           2054       0.206426
2                 2        431           1980       0.217677
```

Current house yrs



Top 5 with LeasT Default Rate: CURRENT_HOUSE_YRS

	current_house_yrs	risk_flag	total_records	Average_@risk
2	12	1628	8640	0.188426
3	13	1687	8713	0.193619
4	14	1694	8652	0.195793
0	10	1704	8561	0.199042
1	11	1730	8624	0.200603

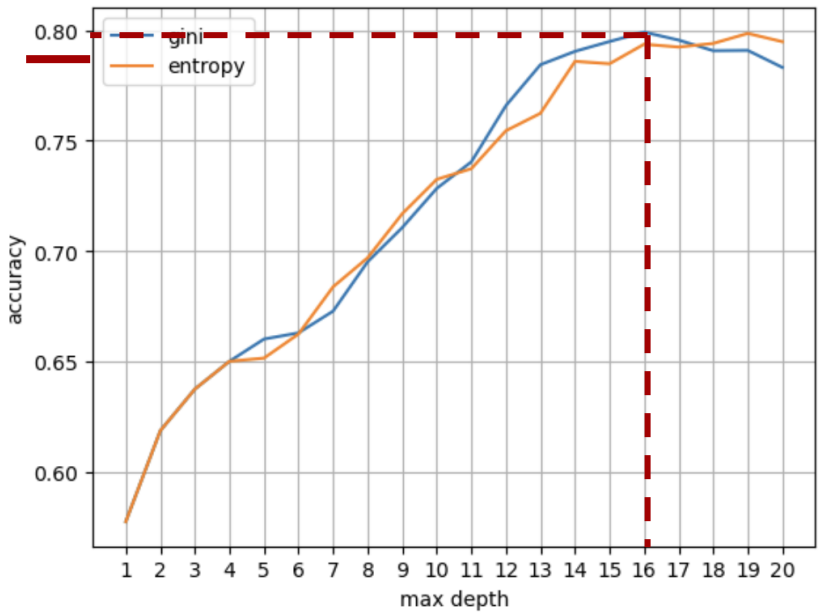
==========

Top 5 with Most Default Rate: CURRENT_HOUSE_YRS

	current_house_yrs	risk_flag	total_records	Average_@risk
2	12	1628	8640	0.188426
3	13	1687	8713	0.193619
4	14	1694	8652	0.195793
0	10	1704	8561	0.199042
1	11	1730	8624	0.200603

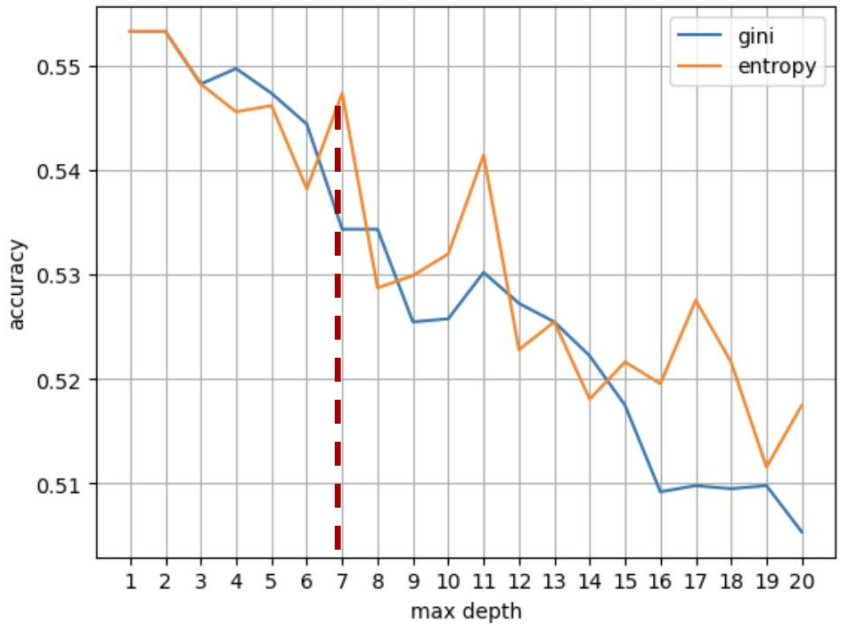
Model Evaluation

- Decision Tree balancing methods

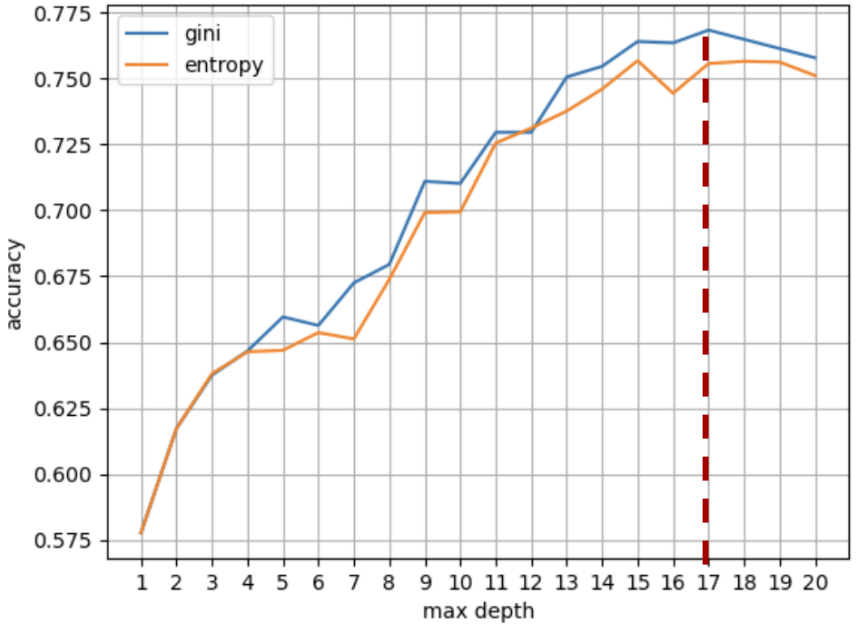


Smote method
Over sampling

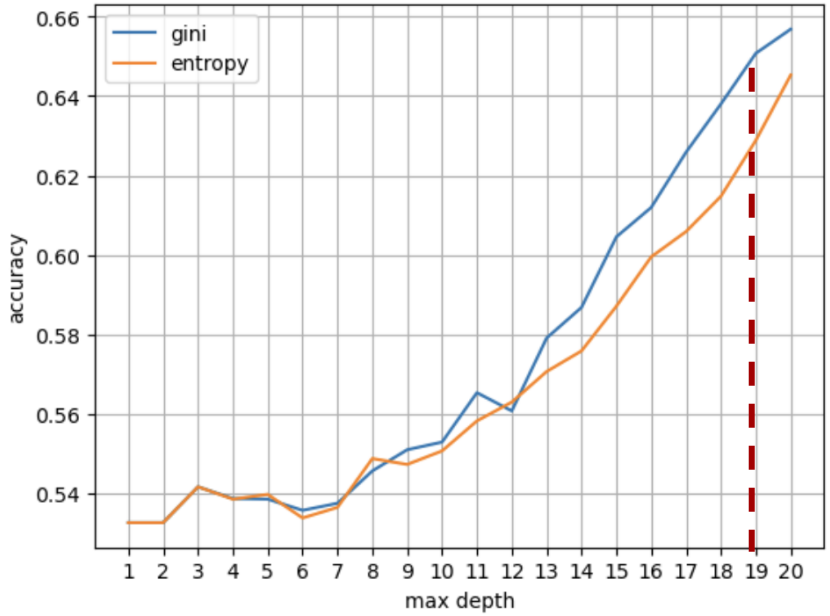
Best method



Undersampling



SmoteENN

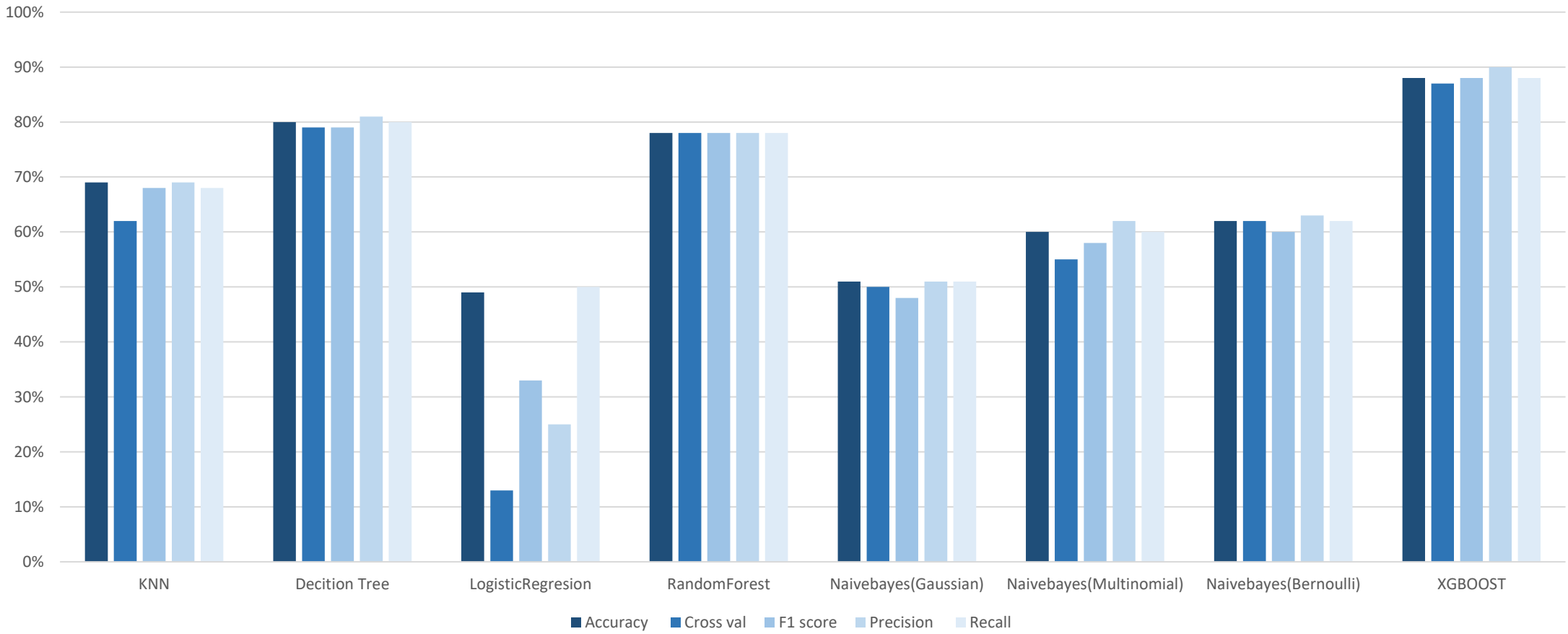


Under over sampling

Evaluation results of models

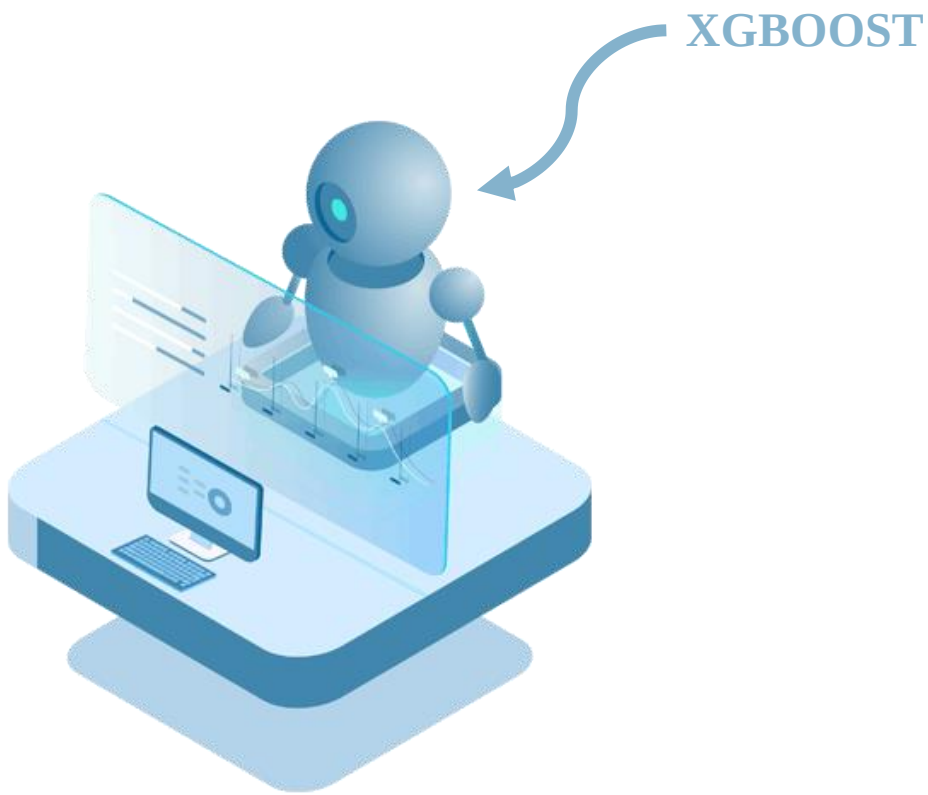
Algorithms	Best sampling method	Accuracy	Cross val	F1 score	Precision	Recall
KNN	Smote	69%	62%	68%	69%	68%
Decition Tree	Smote	80%	79%	79%	81%	80%
LogisticRegresion	In all sampling method accuracy is not acceptable	49%	13%	33%	25%	50%
RandomForest	Smote	78%	78%	78%	78%	78%
Naivebayes(Gaussian)	SmoteENN	51%	50%	48%	51%	51%
Naivebayes(Multinomial)	Smote	60%	55%	58%	62%	60%
Naivebayes(Bernoulli)	Smote	62%	62%	60%	63%	62%
XGBOOST	SmoteENN	88%	87%	88%	90%	88%

Evaluation results of models

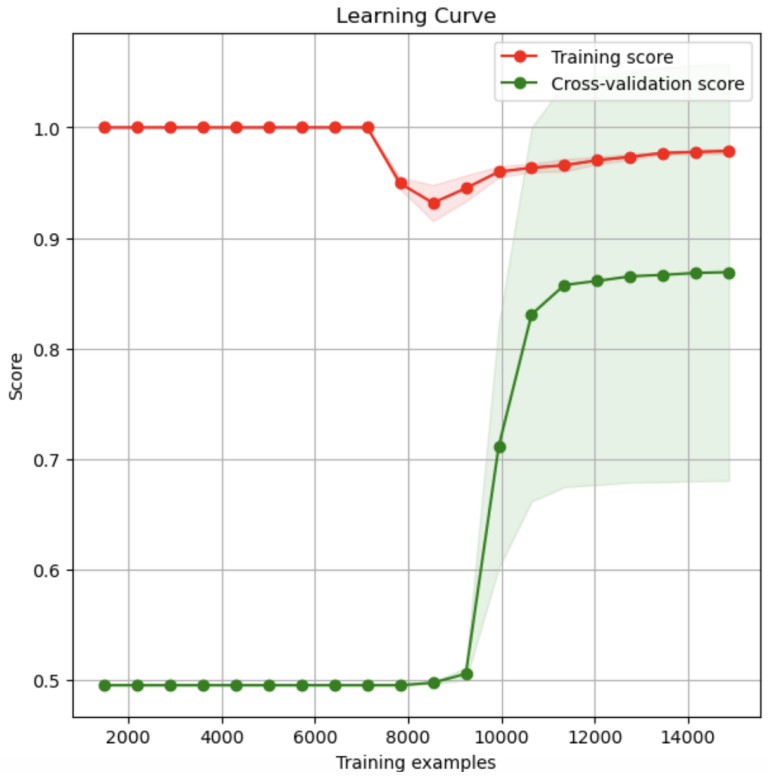


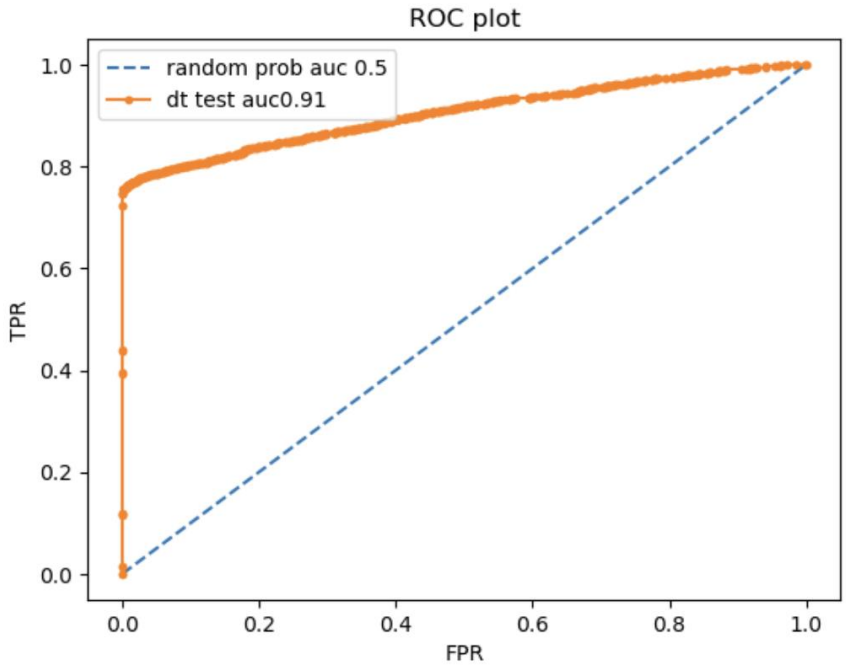
XGBOOST – Best Model

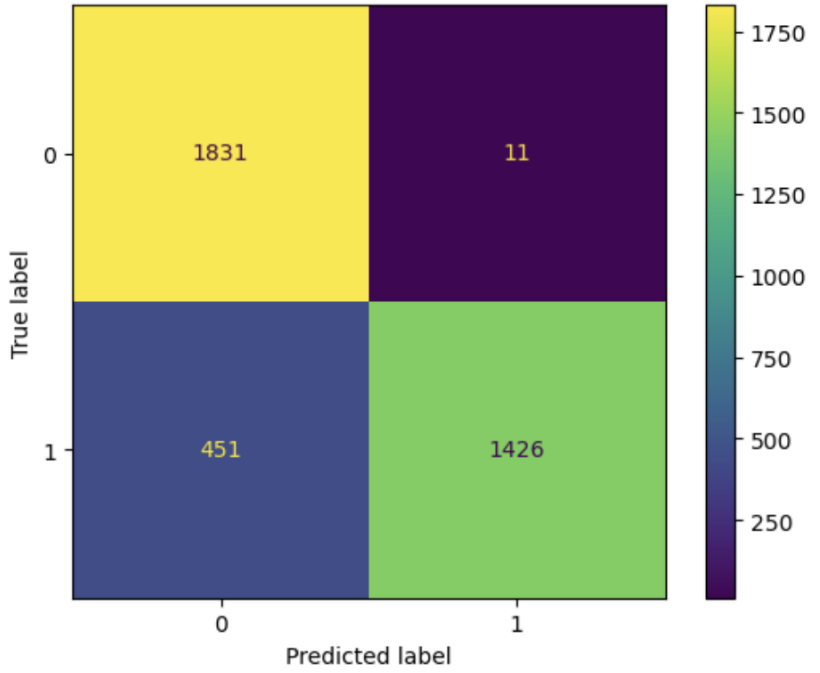
parameters	Best parameters
objective	Binary hinge
Max depth	16
Nestimators	400
gamma	0.1
subsample	0.9
Colsample bytree	0.7
Learning rate	0.01



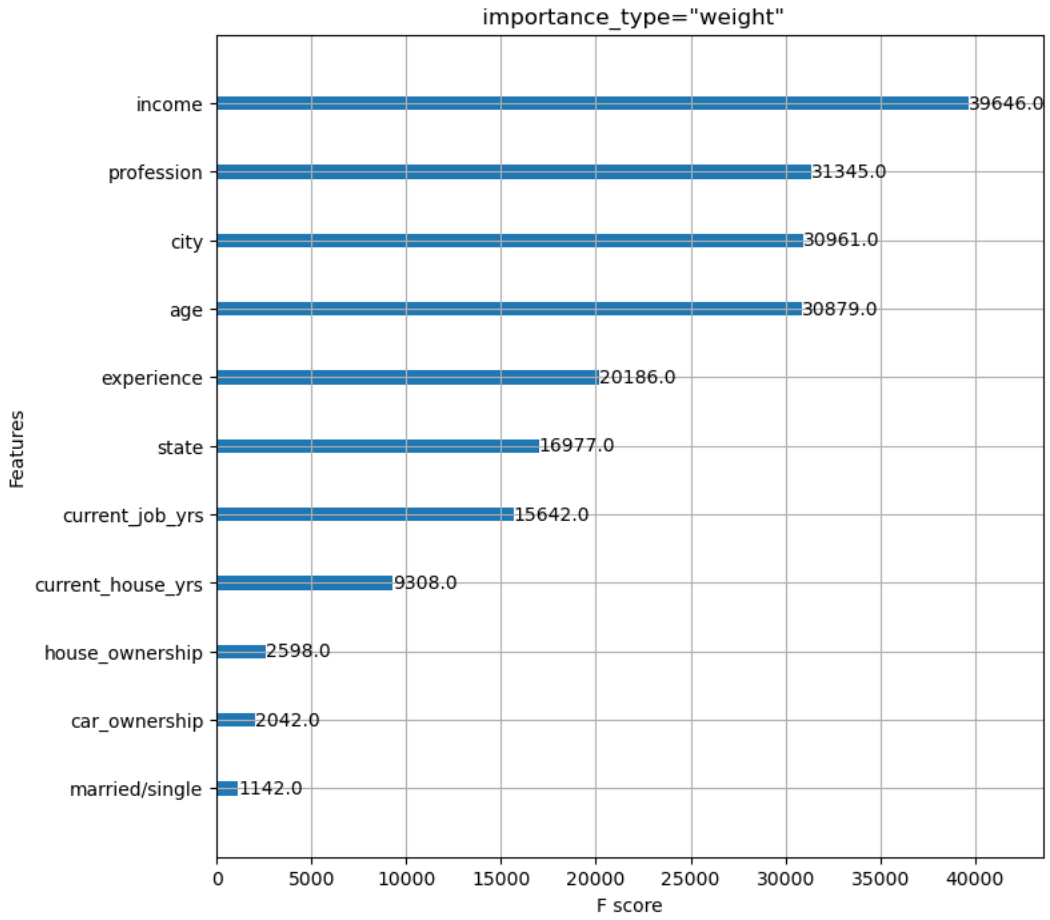
Learning Curve + ROC + confusion matrix of XGBOOST



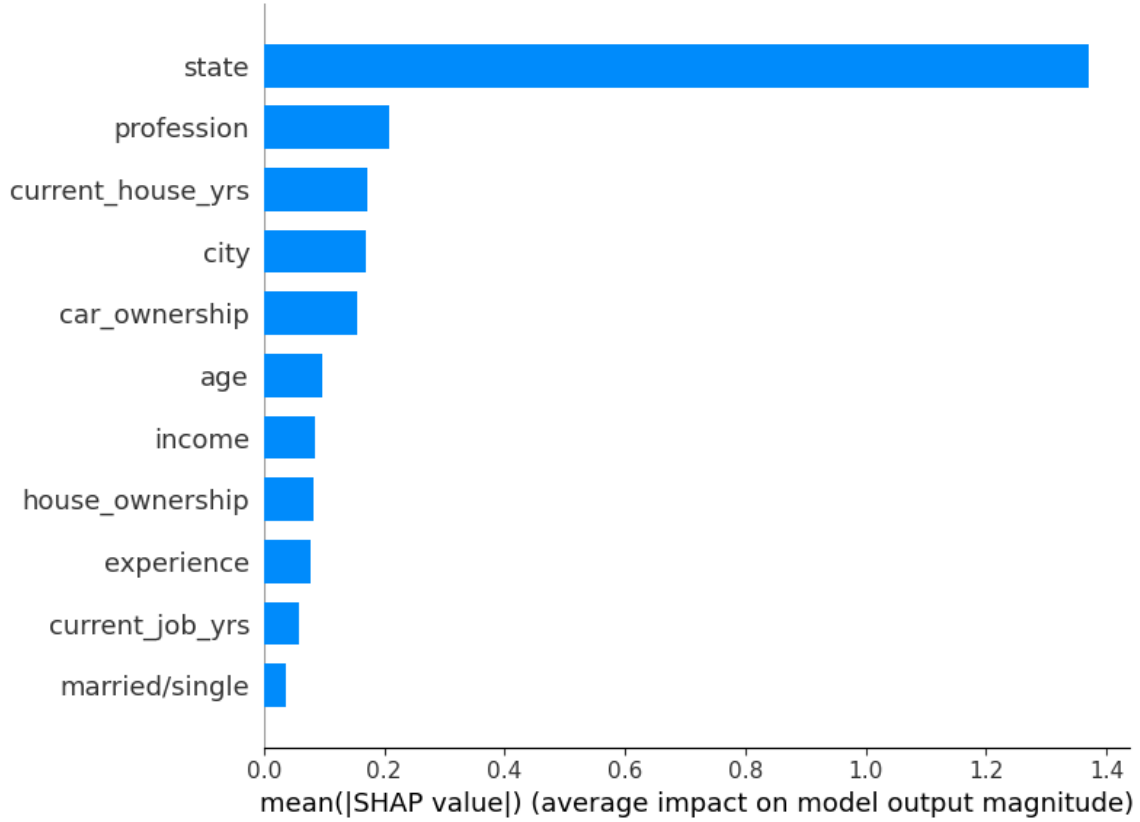




Feature importance xgboost

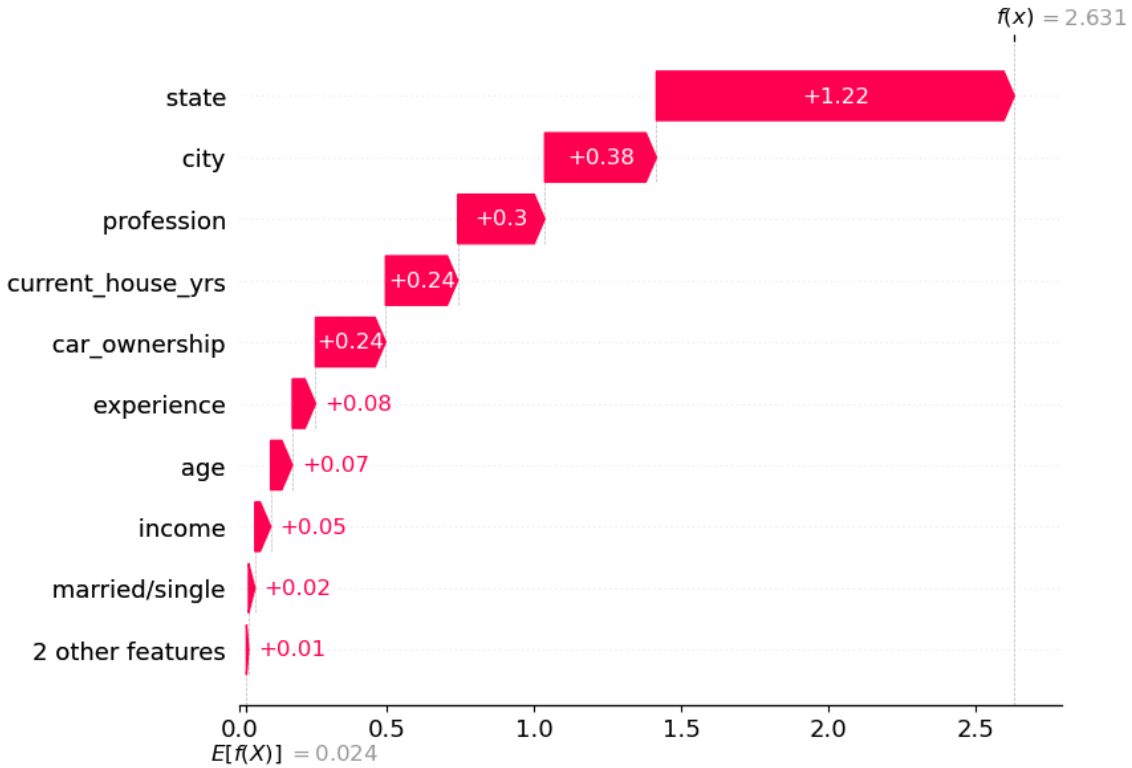


Feature importance shaply values

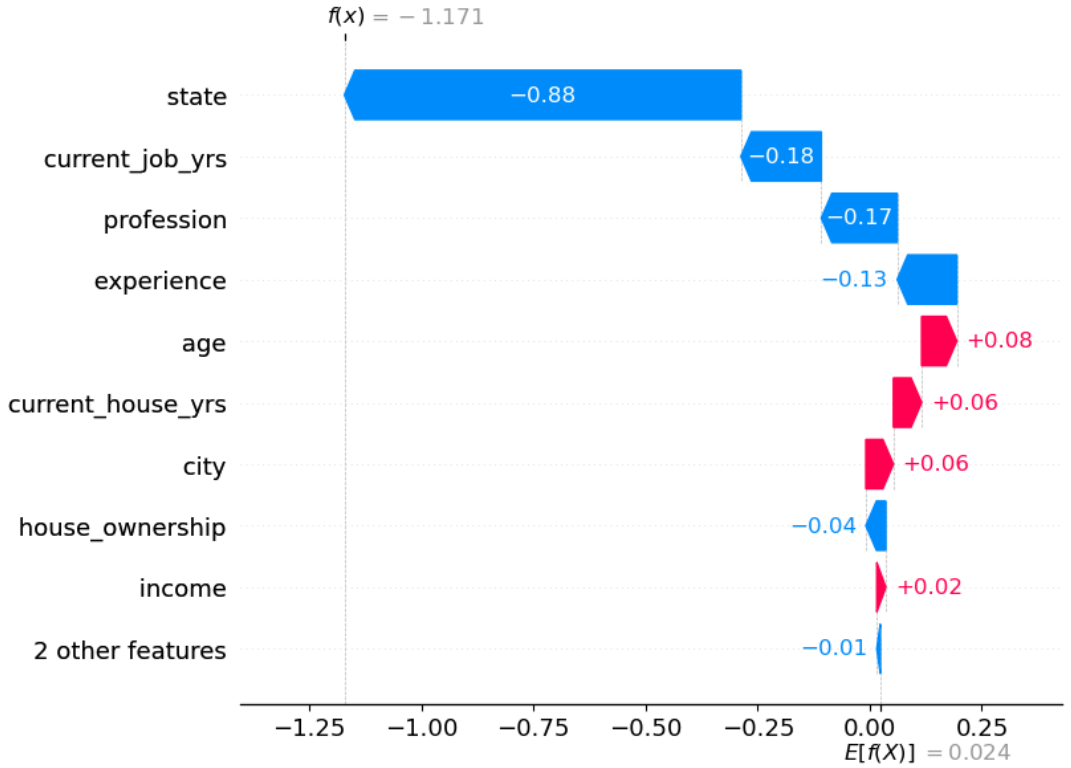


Drop “ Current job yrs ”and “ house ownership “ based on all of feature importances but accuracy and cross-validation did not improved

Shaply values class 1



Shaply values class 0



Recommendations

- According to shap values, you can see the features that are effective in predicting the first class that has loan risk along with the summary plot of whole dataset importances of features and characteristics of loan borrowers which contributes the most for the selected model.
- For better results, special attention should be paid to these highlighted features
- You can also add features like (in data collection) :

1. Financial Reserves
2. Loan Repayment History
3. Loan Term (Loan repayment period, Loan amount received)
4. the amount of expenses and the ratio of expenses to income should be measured
5. The purpose of the loan(what will it be used for)

The End