PwC Employee Reviews Analysis

Source: glassdoor.com

Manel RAHALI
Soumaya ATOUI
Hadiqa Javaid
21/11/2025

AGENDA

• Business use case	3
• Presenting DataFrame	4
• Data quality issues	5
• New DataFrame	7
• Univariate analysis : Rating distribution	14
• Business POV	15
• Correlation matrix	16
• Preprocessing	18
• Models	21
• Recommandations and results	

Business Use Case

Help PwC HR teams identify employees who may be at risk of low satisfaction

Understand which countries, roles or functions consistently show higher risk levels

Build a model that predicts the at_risk label derived from ratings ≤ 3

Presenting the DataFrame

1 . Content

2 . Size

3 . How & When?

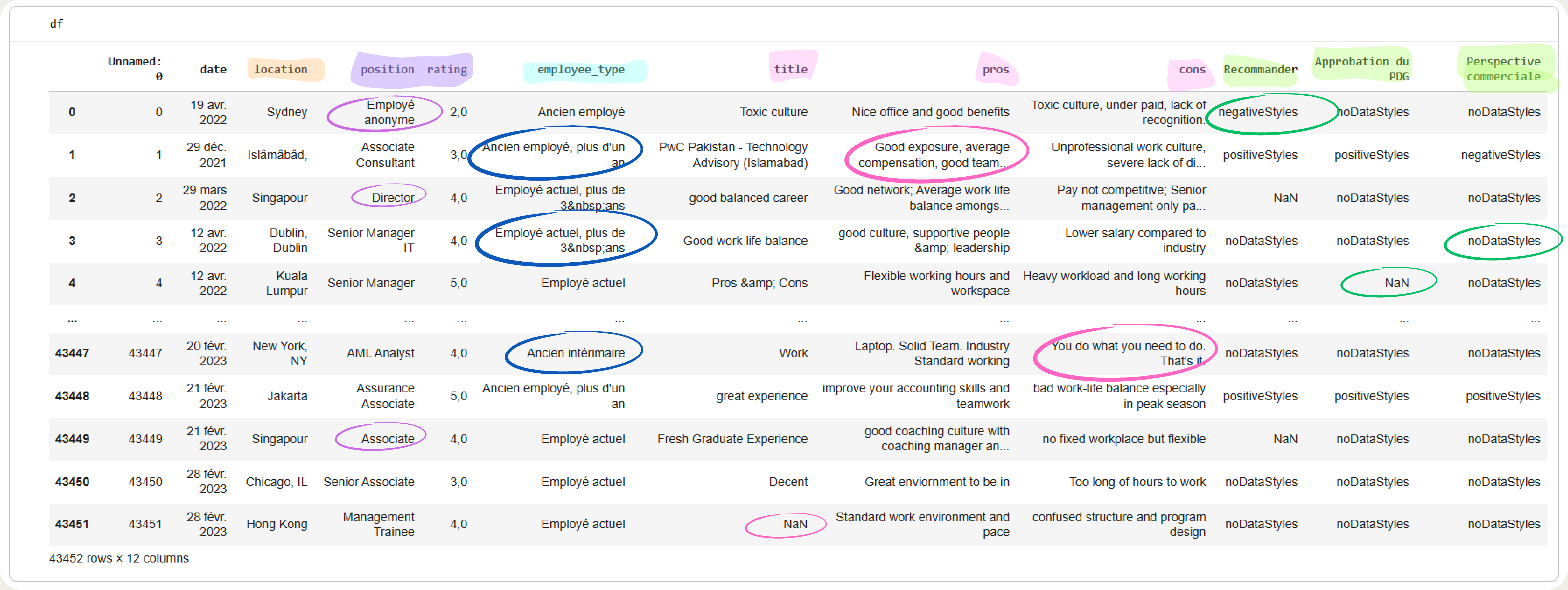
Reviews from :
Currrent & Former
Employees of PwC

12 Columns
43K+ Rows
(before cleaning)

GLASSDOOR

2008 - 2024

Data quality issues

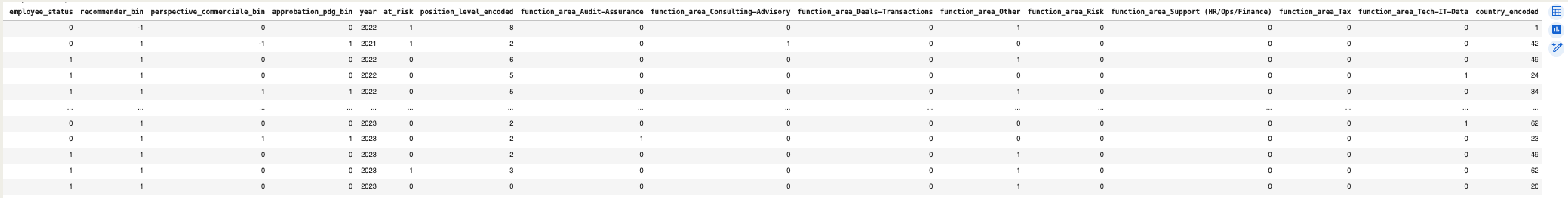


```
df
```

	Unnamed: 0	date	location	position	rating	employee_type	title	pros	cons	Recommander	Approbation du PDG	Perspective commerciale
0	0	19 avr. 2022	Sydney	Employé anonyme	2,0	Ancien employé	Toxic culture	Nice office and good benefits	Toxic culture, under paid, lack of recognition.	negativeStyles	noDataStyles	noDataStyles
1	1	29 déc. 2021	Islâmâbâd,	Associate Consultant	3,0	Ancien employé, plus d'un an	PwC Pakistan - Technology Advisory (Islamabad)	Good exposure, average compensation, good team...	Unprofessional work culture, severe lack of di...	positiveStyles	positiveStyles	negativeStyles
2	2	29 mars 2022	Singapour	Director	4,0	Employé actuel, plus de 3&nbsp;ans	good balanced career	Good network; Average work life balance amongs...	Pay not competitive; Senior management only pa...	NaN	noDataStyles	noDataStyles
3	3	12 avr. 2022	Dublin, Dublin	Senior Manager IT	4,0	Employé actuel, plus de 3&nbsp;ans	Good work life balance	good culture, supportive people &amp; leadership	Lower salary compared to industry	noDataStyles	noDataStyles	noDataStyles
4	4	12 avr. 2022	Kuala Lumpur	Senior Manager	5,0	Employé actuel	Pros &amp; Cons	Flexible working hours and workspace	Heavy workload and long working hours	noDataStyles	NaN	noDataStyles
...	...	...	...	...	...	...	...	...	...	...	...	...
43447	43447	20 févr. 2023	New York, NY	AML Analyst	4,0	Ancien intérimaire	Work	Laptop. Solid Team. Industry Standard working	You do what you need to do. That's it.	noDataStyles	noDataStyles	noDataStyles
43448	43448	21 févr. 2023	Jakarta	Assurance Associate	5,0	Ancien employé, plus d'un an	great experience	improve your accounting skills and teamwork	bad work-life balance especially in peak season	positiveStyles	positiveStyles	positiveStyles
43449	43449	21 févr. 2023	Singapour	Associate	4,0	Employé actuel	Fresh Graduate Experience	good coaching culture with coaching manager an...	no fixed workplace but flexible	NaN	noDataStyles	noDataStyles
43450	43450	28 févr. 2023	Chicago, IL	Senior Associate	3,0	Employé actuel	Decent	Great enviornment to be in	Too long of hours to work	noDataStyles	noDataStyles	noDataStyles
43451	43451	28 févr. 2023	Hong Kong	Management Trainee	4,0	Employé actuel	NaN	Standard work environment and pace	confused structure and program design	noDataStyles	noDataStyles	noDataStyles

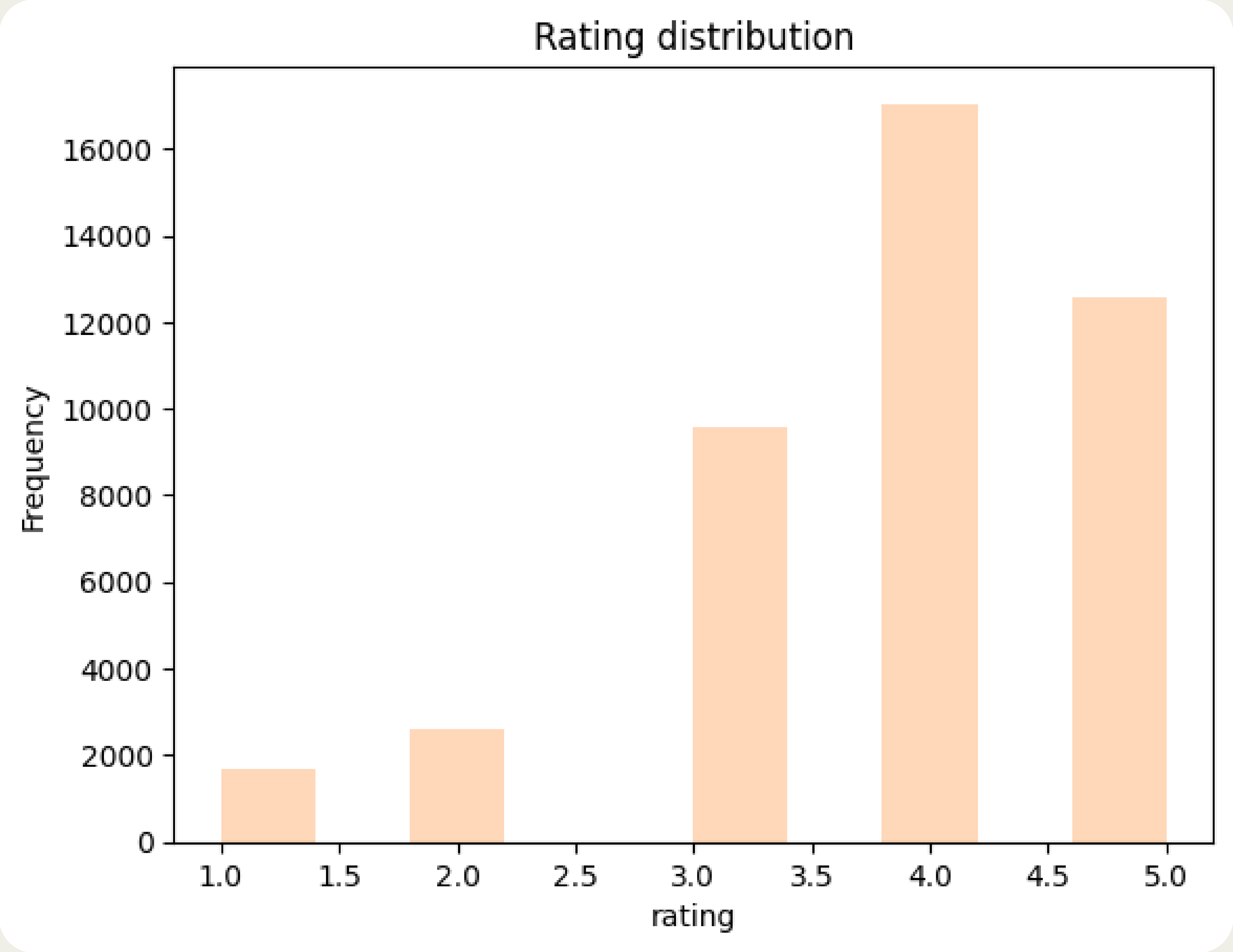
43452 rows × 12 columns

Final DataFrame:



employee_status	recommender_bin	perspective_commerciale_bin	approbation_pdg_bin	year	at_risk	position_level_encoded	function_area_Audit-Assurance	function_area_Consulting-Advisory	function_area_Deals-Transactions	function_area_Other	function_area_Risk	function_area_Support (HR/Ops/Finance)	function_area_Tax	function_area_Tech-IT-Data	country_encoded
0	-1	0	0	2022	1	8	0	0	0	1	0	0	0	0	1
0	1	-1	1	2021	1	2	0	1	0	0	0	0	0	0	42
1	1	0	0	2022	0	6	0	0	0	1	0	0	0	0	49
1	1	0	0	2022	0	5	0	0	0	0	0	0	0	1	24
1	1	1	1	2022	0	5	0	0	0	1	0	0	0	0	34
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	1	0	0	2023	0	2	0	0	0	0	0	0	0	1	62
0	1	1	1	2023	0	2	1	0	0	0	0	0	0	0	23
1	1	0	0	2023	0	2	0	0	0	1	0	0	0	0	49
1	1	0	0	2023	1	3	0	0	0	1	0	0	0	0	62
1	1	0	0	2023	0	0	0	0	0	1	0	0	0	0	20

Univariate analysis : Rating distribution

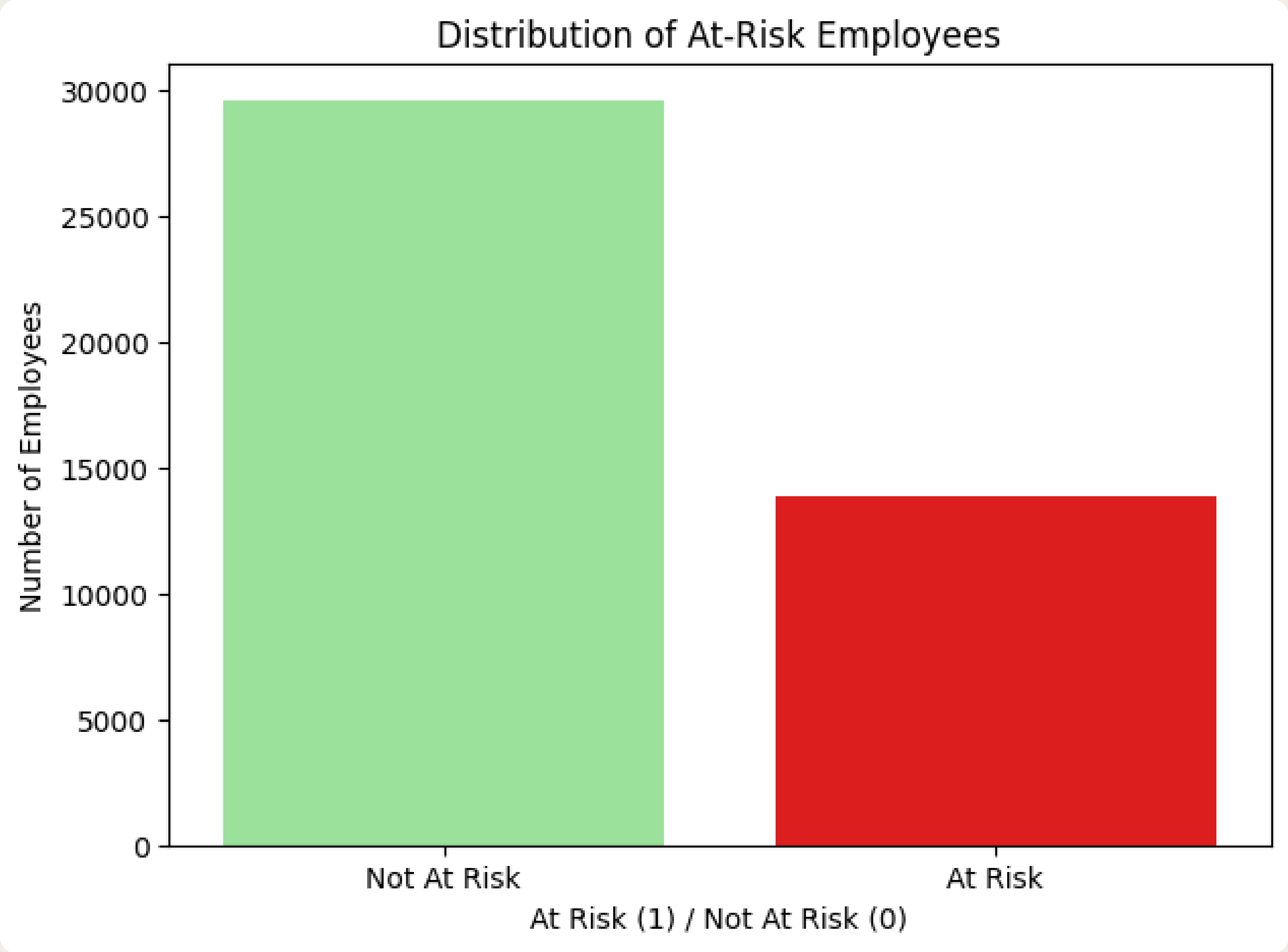


★★★⯪☆

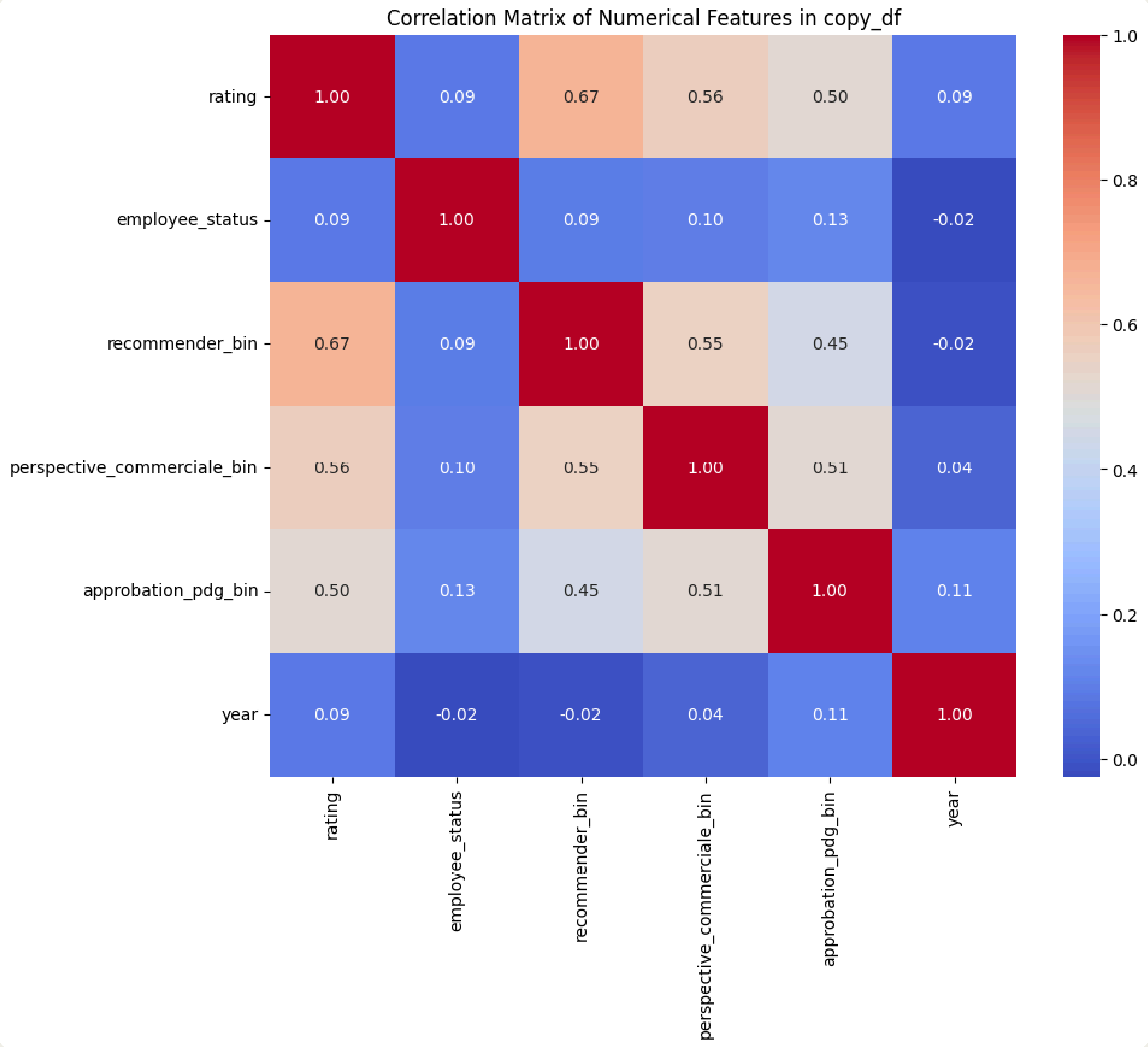
3.83

Average Rating

Distribution of risk vs. not at risk



Correlation Matrix



Correlation Matrix of Numerical Features in copy_df

	rating	employee_status	recommender_bin	perspective_commerciale_bin	approbation_pdg_bin	year
rating	1.00	0.09	0.67	0.56	0.50	0.09
employee_status	0.09	1.00	0.09	0.10	0.13	-0.02
recommender_bin	0.67	0.09	1.00	0.55	0.45	-0.02
perspective_commerciale_bin	0.56	0.10	0.55	1.00	0.51	0.04
approbation_pdg_bin	0.50	0.13	0.45	0.51	1.00	0.11
year	0.09	-0.02	-0.02	0.04	0.11	1.00

ML models used

Random Forest

test 1: 100% accuracy

test 2: 74 % accuracy

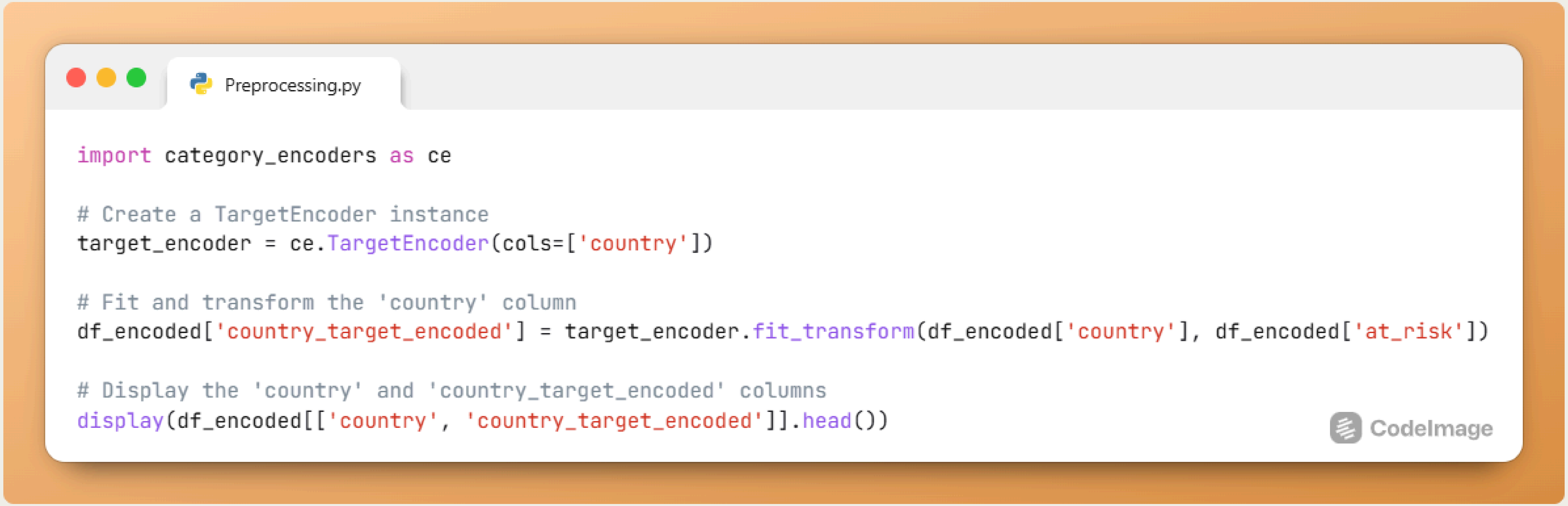
LightGBM

70% Accuracy

CatBoost

93% Accuracy

Preprocessing (the country example)



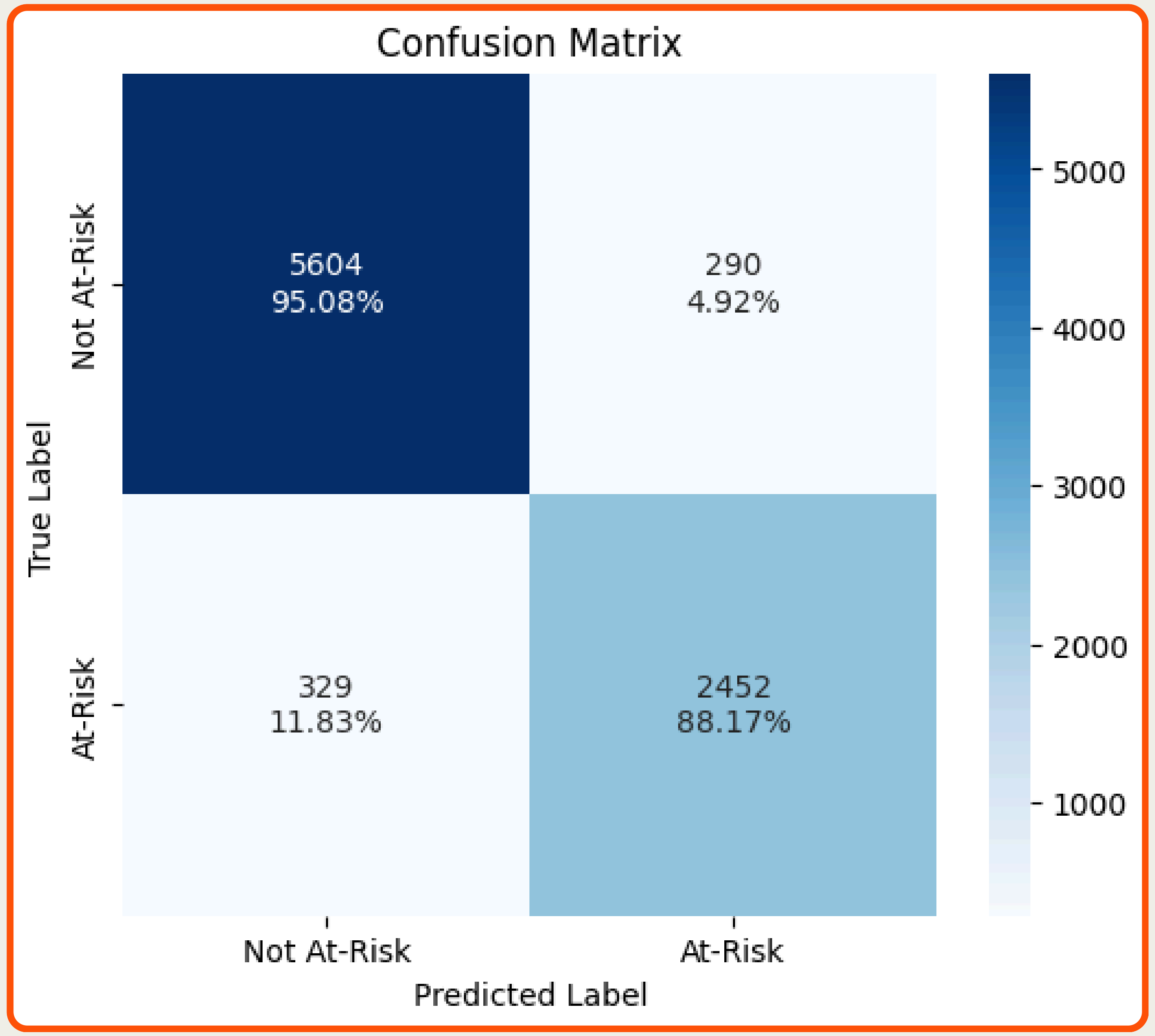
```python
import category_encoders as ce

# Create a TargetEncoder instance
target_encoder = ce.TargetEncoder(cols=['country'])

# Fit and transform the 'country' column
df_encoded['country_target_encoded'] = target_encoder.fit_transform(df_encoded['country'], df_encoded['at_risk'])

# Display the 'country' and 'country_target_encoded' columns
display(df_encoded[['country', 'country_target_encoded']].head())
```

Results: CatBoost Model



Accuracy: *0.9286*

Precision: *0.8942*

Recall: *0.8817*

F1-Score: *0.8879*

ROC/AUC: *0.9322*

Confusion Matrix

CatBoost Model Performance

Recommandations

recommanding regular feedback collection

Associates
Audit assurance

Thank you!

QUESTIONS?