

Final Project DS 18

# Credit Default Prediction

Suci Rahma Nura

Juni 2023

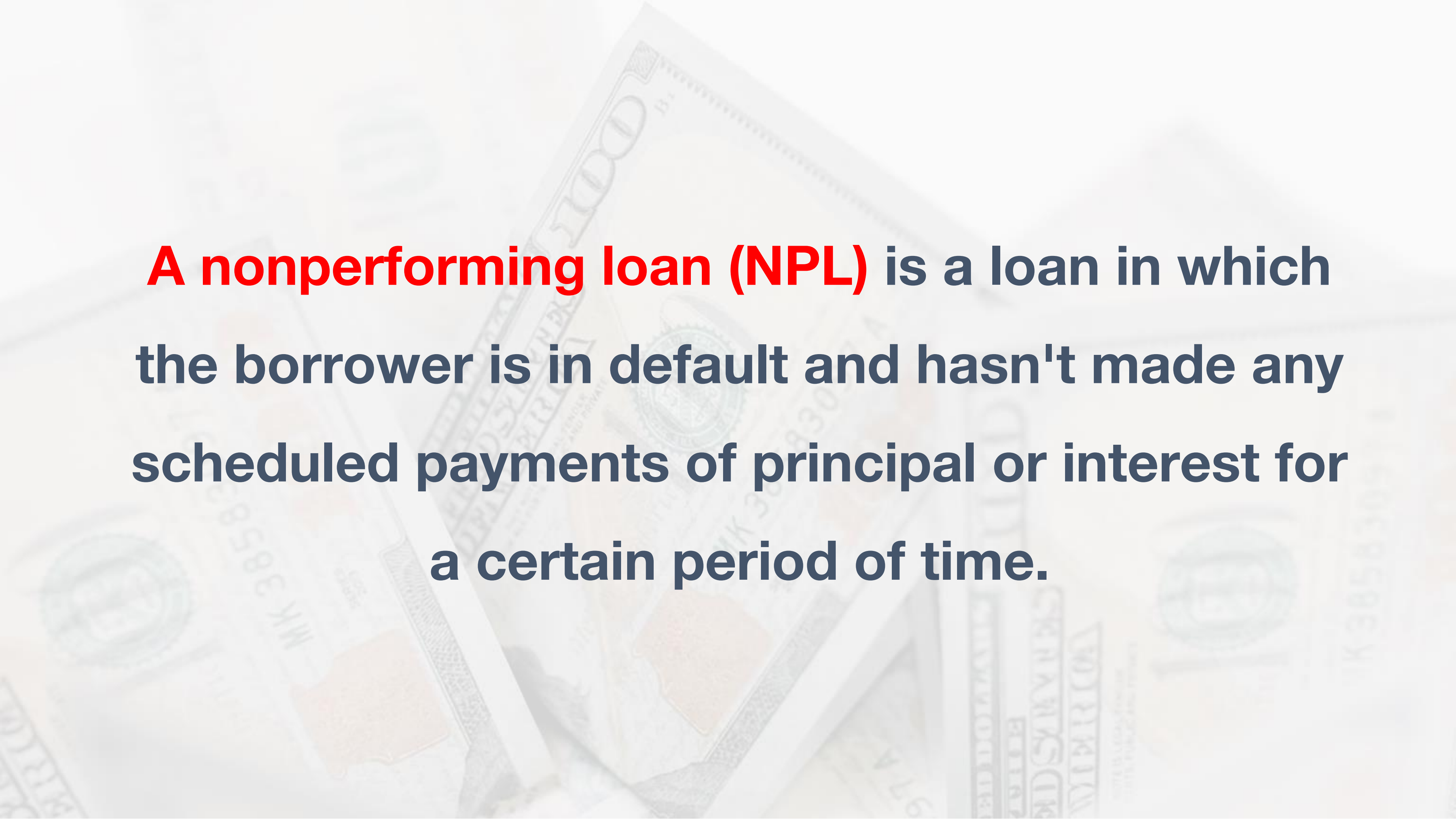


Indonesia's non-performing loan (NPL) ratio in April 2023 was reported to have **increased** by **2.5%** from March 2023, with an average ratio increase from January - March 2023 of **3.0%.**




The background of the image consists of several overlapping US dollar bills, including \$100 and \$10 bills, which are slightly blurred and faded. The bills are oriented in various directions, creating a sense of depth and texture. The text "What is NPL?" is centered over this background.

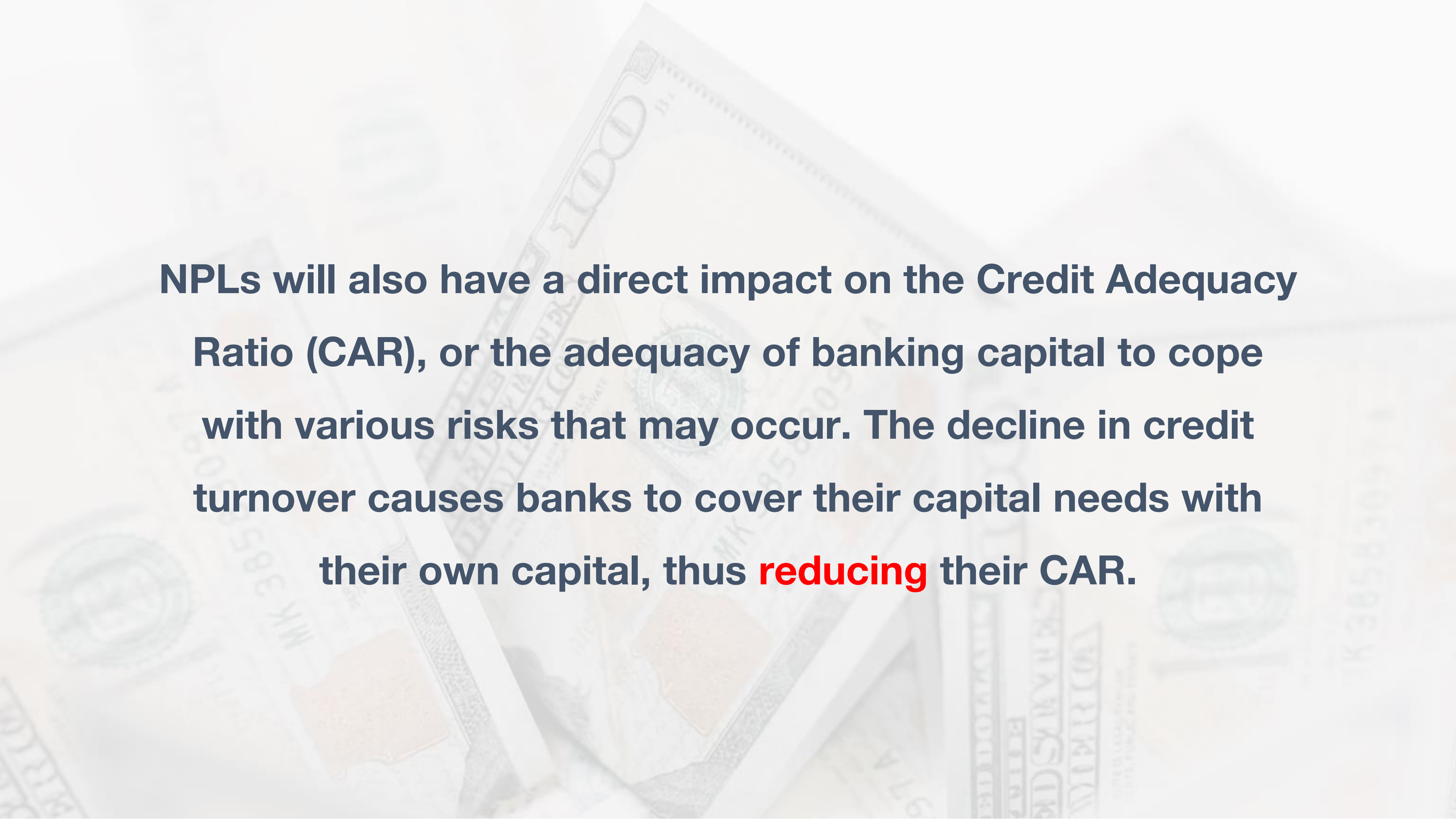
**What is NPL ?**

The background of the slide features a close-up, slightly blurred image of several US dollar bills. A \$100 bill is prominent in the center, with its green and white colors and intricate patterns visible. Other bills, including a \$20 bill, are partially visible in the foreground and background, creating a sense of depth and financial context.

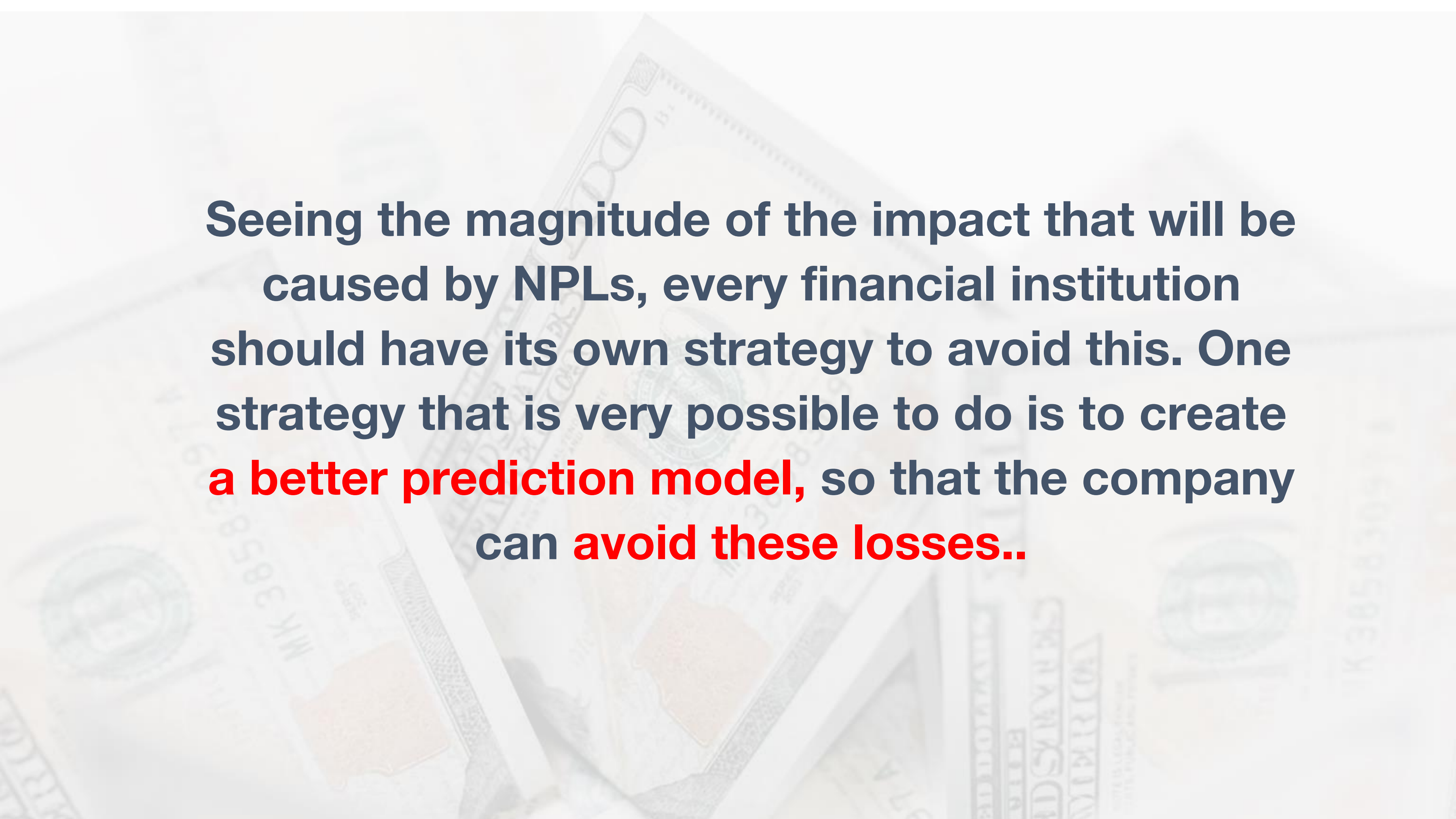
**A nonperforming loan (NPL)** is a loan in which the borrower is in default and hasn't made any scheduled payments of principal or interest for a certain period of time.

The background of the image consists of several overlapping US dollar bills, including a \$100 bill and a \$20 bill, which are slightly out of focus. The bills are scattered across the frame, with some showing the portrait of the president and others showing the serial numbers and denominations.

**NPLs for Financial Institutions are a sign of  
unsound business management. If these  
circumstances do not change, the impact may  
worsen the economic situation of the nation.**

The background of the slide features a blurred, overlapping image of several Euro banknotes. The notes are in various denominations, including 100, 50, and 20 Euros, and are oriented diagonally across the frame. The colors are muted, with a focus on the green and yellow tones of the currency.

**NPLs will also have a direct impact on the Credit Adequacy Ratio (CAR), or the adequacy of banking capital to cope with various risks that may occur. The decline in credit turnover causes banks to cover their capital needs with their own capital, thus **reducing** their CAR.**

The background of the slide features a blurred, overlapping image of several Euro banknotes. The notes are in various denominations, including 100, 50, and 20 Euros, and are oriented diagonally. The colors are muted, with a focus on the green and yellow tones of the currency.

**Seeing the magnitude of the impact that will be caused by NPLs, every financial institution should have its own strategy to avoid this. One strategy that is very possible to do is to create a better prediction model, so that the company can avoid these losses..**



The background of the image consists of several Indian 1000 Rupee banknotes. The notes are slightly out of focus, with the central one being more prominent. They are arranged in a way that they overlap each other, creating a sense of depth. The colors of the notes are primarily green and yellow, with some orange and red accents. The text 'So, let's check our' is in a dark blue font, and 'Credit Default Prediction' is in a bold red font. The exclamation mark is in a dark blue font.

**So, let's check our**  
**‘Credit Default Prediction ’!**



# **Table of Contents**

**Data Understanding and Exploratory Data Analysis**

**Data Preprocessing**

**Modelling with Balance Data**

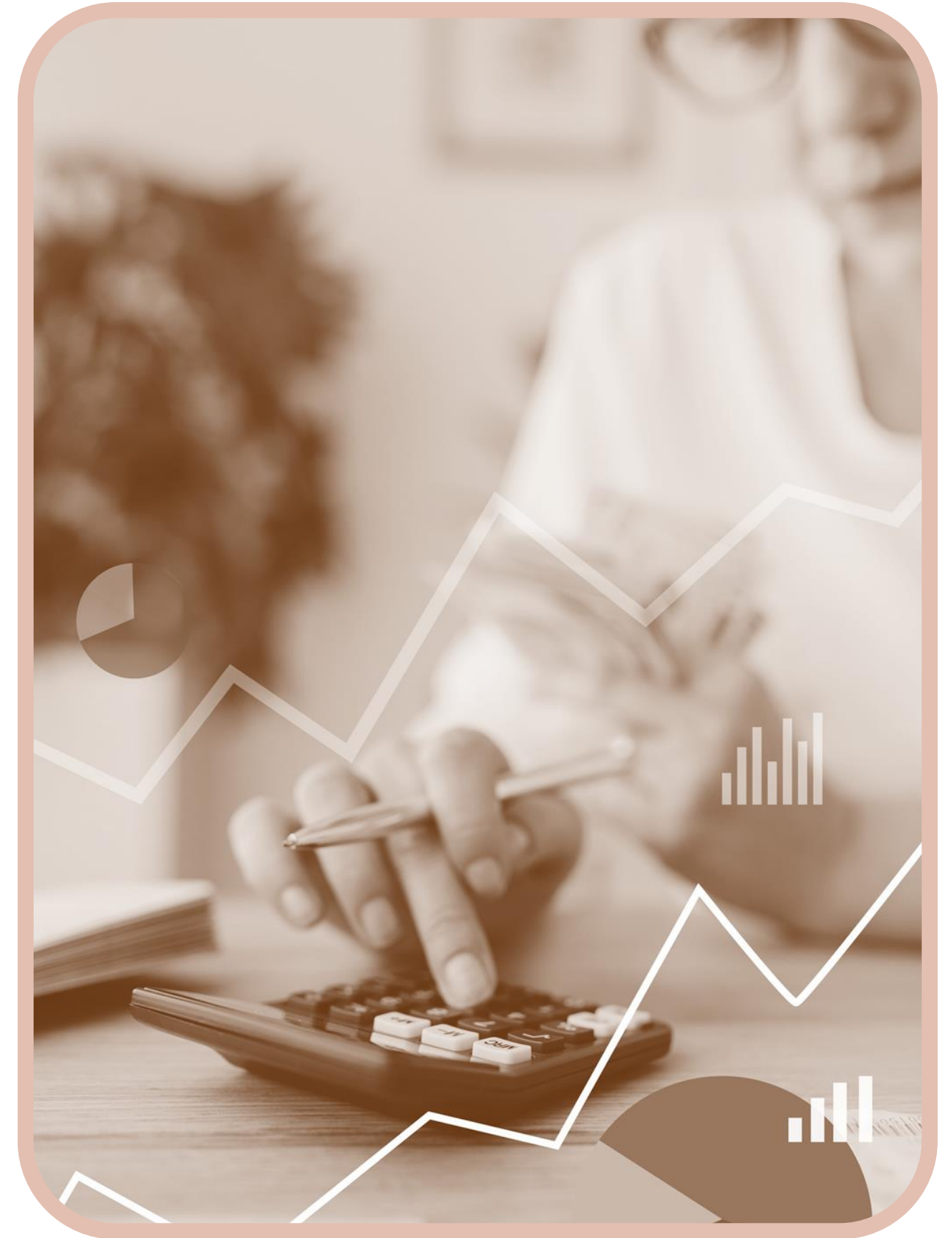
**Dalex**

**Business Case and Recommendation**

**Appendix**

# Objective

- What kind of debtors have a tendency to default?
- What variables affect the risk of default the most?
- What machine learning models are suitable for predicting debtors defaults?
- What impact will the model have on the business?



The background of the image is a collage of various US dollar bills, including \$100, \$50, and \$20 bills, which are slightly out of focus. A dark blue rectangular box is centered over the image, containing the text "Data Understanding" in white.

# Data Understanding



# Data Understanding

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 252000 entries, 0 to 251999
```

```
Data columns (total 13 columns):
```

#	Column	Non-Null Count	Dtype
0	Id	252000 non-null	int64
1	income	252000 non-null	int64
2	age	252000 non-null	int64
3	experience	252000 non-null	int64
4	married	252000 non-null	object
5	house_ownership	252000 non-null	object
6	car_ownership	252000 non-null	object
7	profession	252000 non-null	object
8	city	252000 non-null	object
9	state	252000 non-null	object
10	current_job_years	252000 non-null	int64
11	current house_years	252000 non-null	int64
12	risk_flag	252000 non-null	int64

```
dtypes: int64(7), object(6)
```

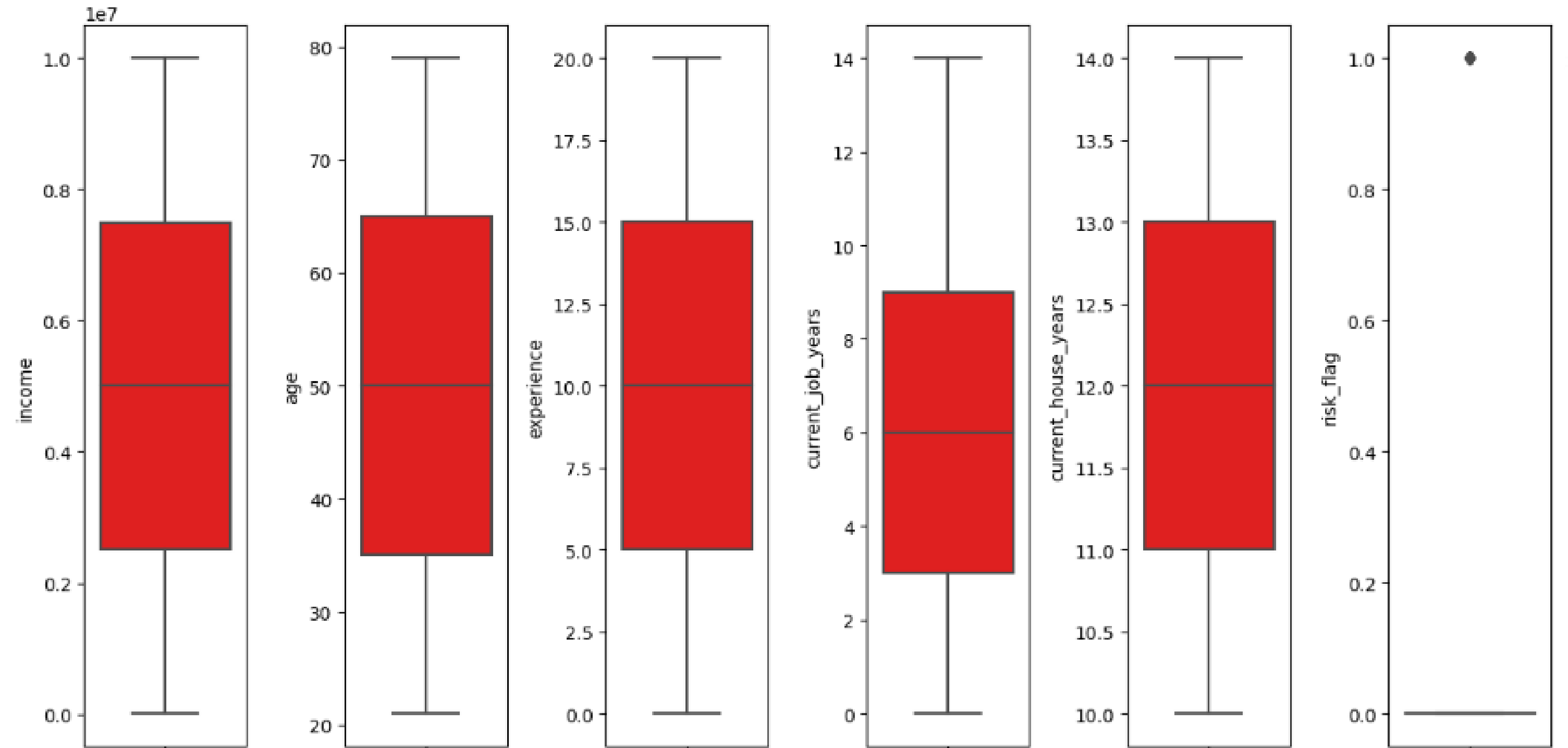
```
memory usage: 25.0+ MB
```

- This dataset is obtained from a Hackathon.  
[https://www.kaggle.com/datasets/gargvg/univai-dataset?select=univ.ai\\_Training+Data.csv](https://www.kaggle.com/datasets/gargvg/univai-dataset?select=univ.ai_Training+Data.csv)
- This dataset has 12 columns and 252000 rows
- Target feature : 'risk\_flag'
- No missing and duplicate values



# Exploratory Data Analysis

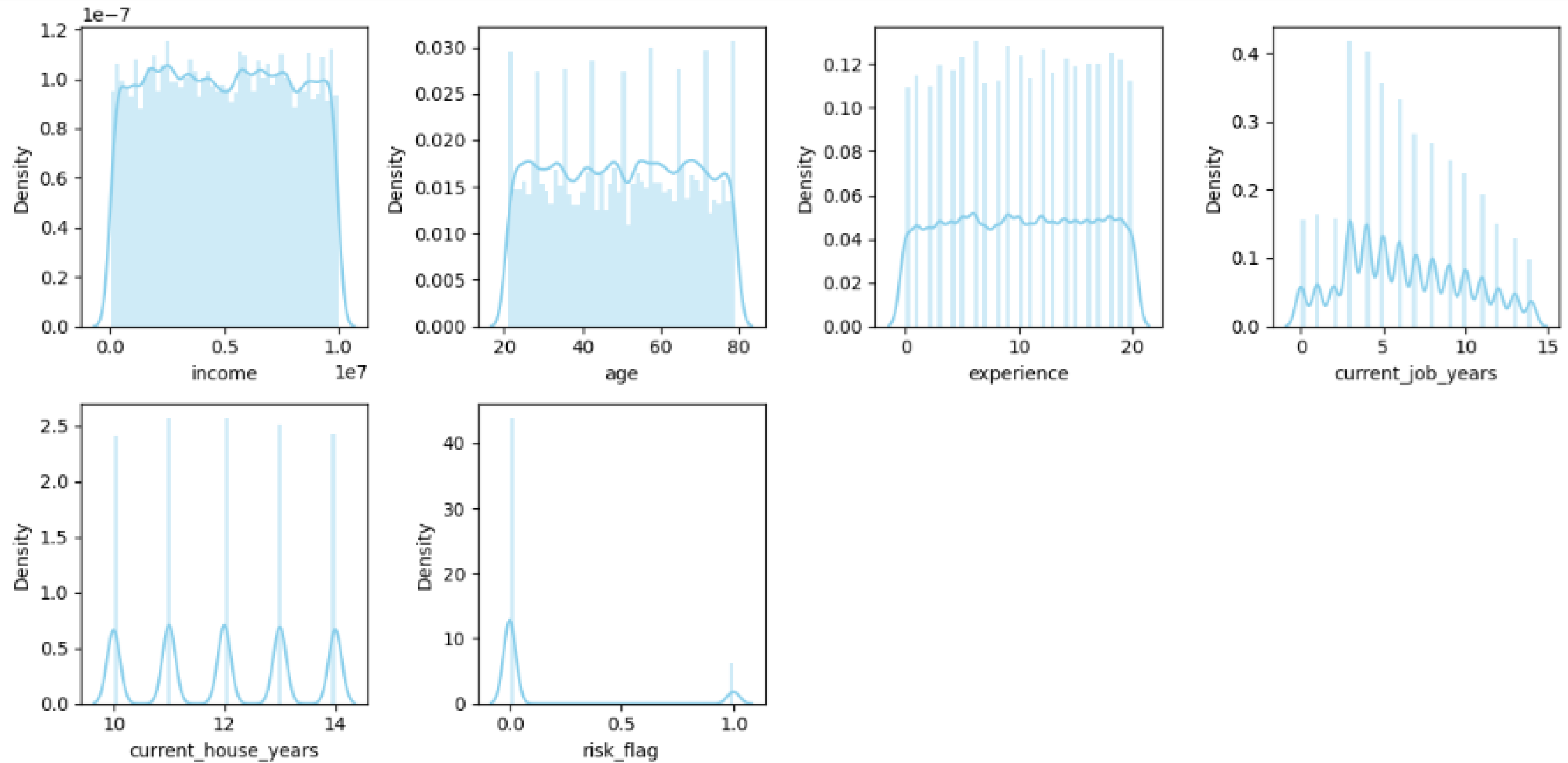
# EDA



We can see, that risk-flag is the only column that has outliers and it reasonable, since it only has 2 unique value ( 0 and 1). So we can say that there are no outliers in this dataset.

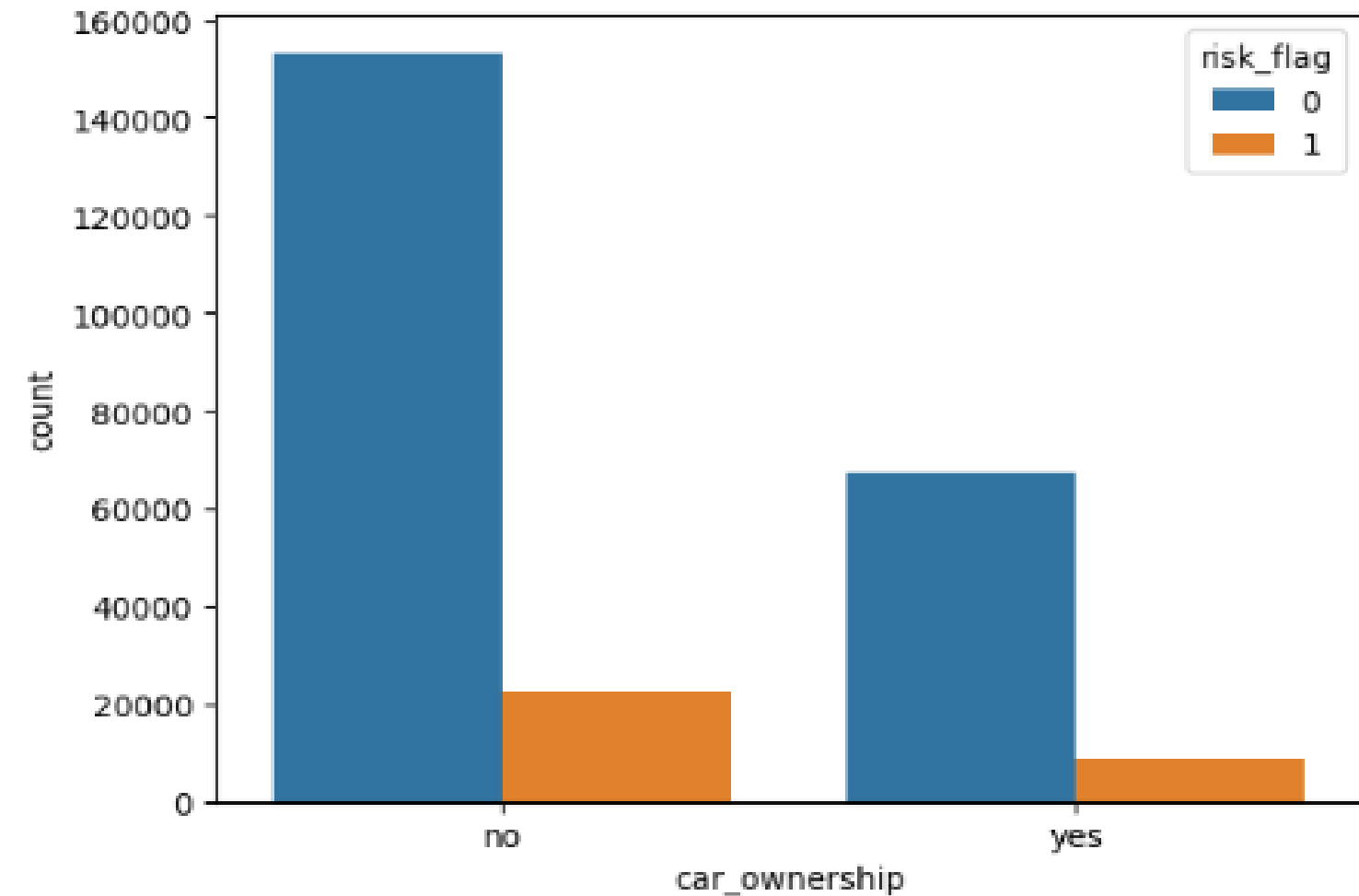
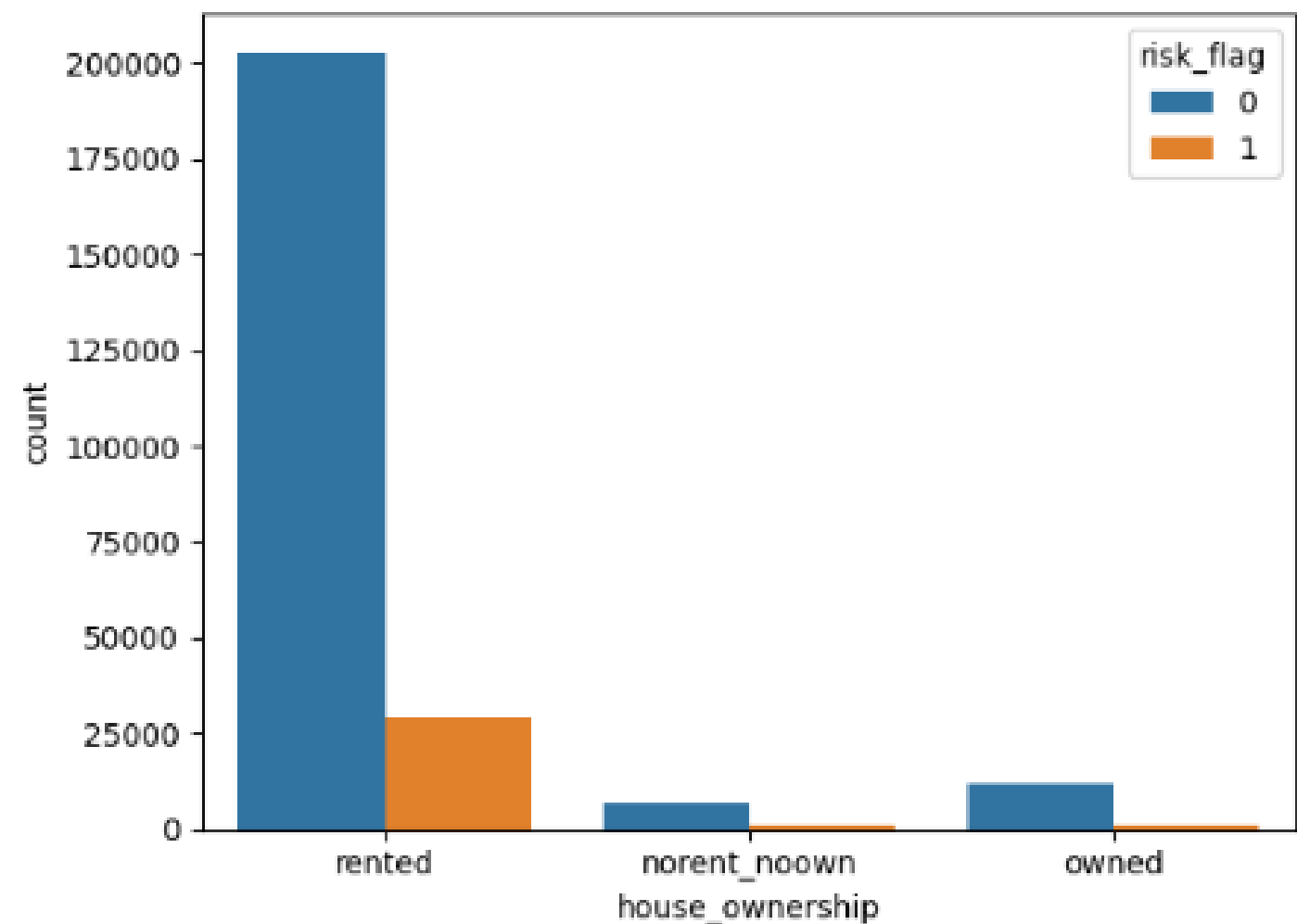


# EDA

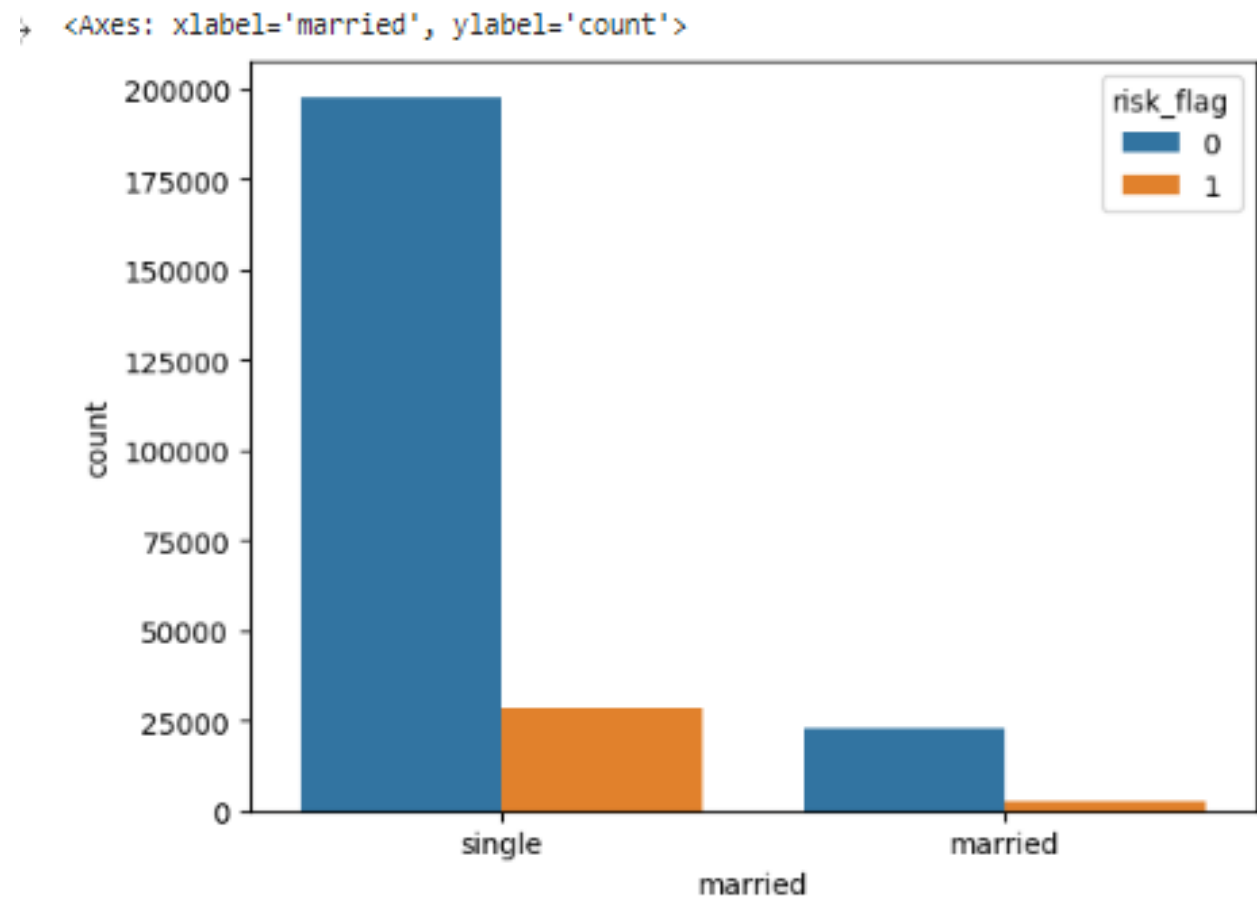


**Somehow, the columns are not simetrical, but not skew either**

# EDA

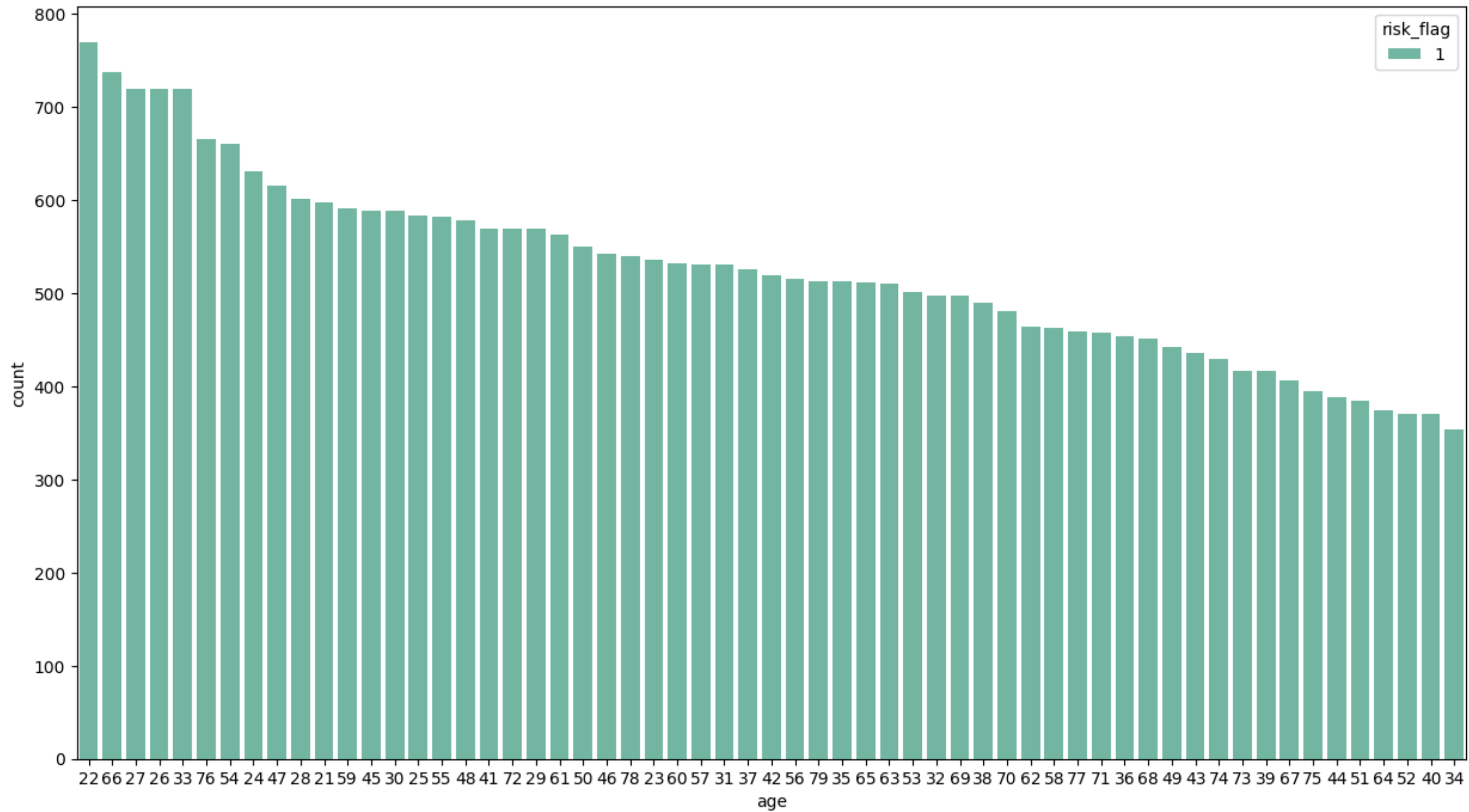


notes : blue bars indicate **non-risky** debtors and orange bars indicate **high-risk** debtors.



Based on this graph, debtors who are single, live in a rented house, and are not car owners have a higher risk of default than those who are married, live in their own house and own a car.

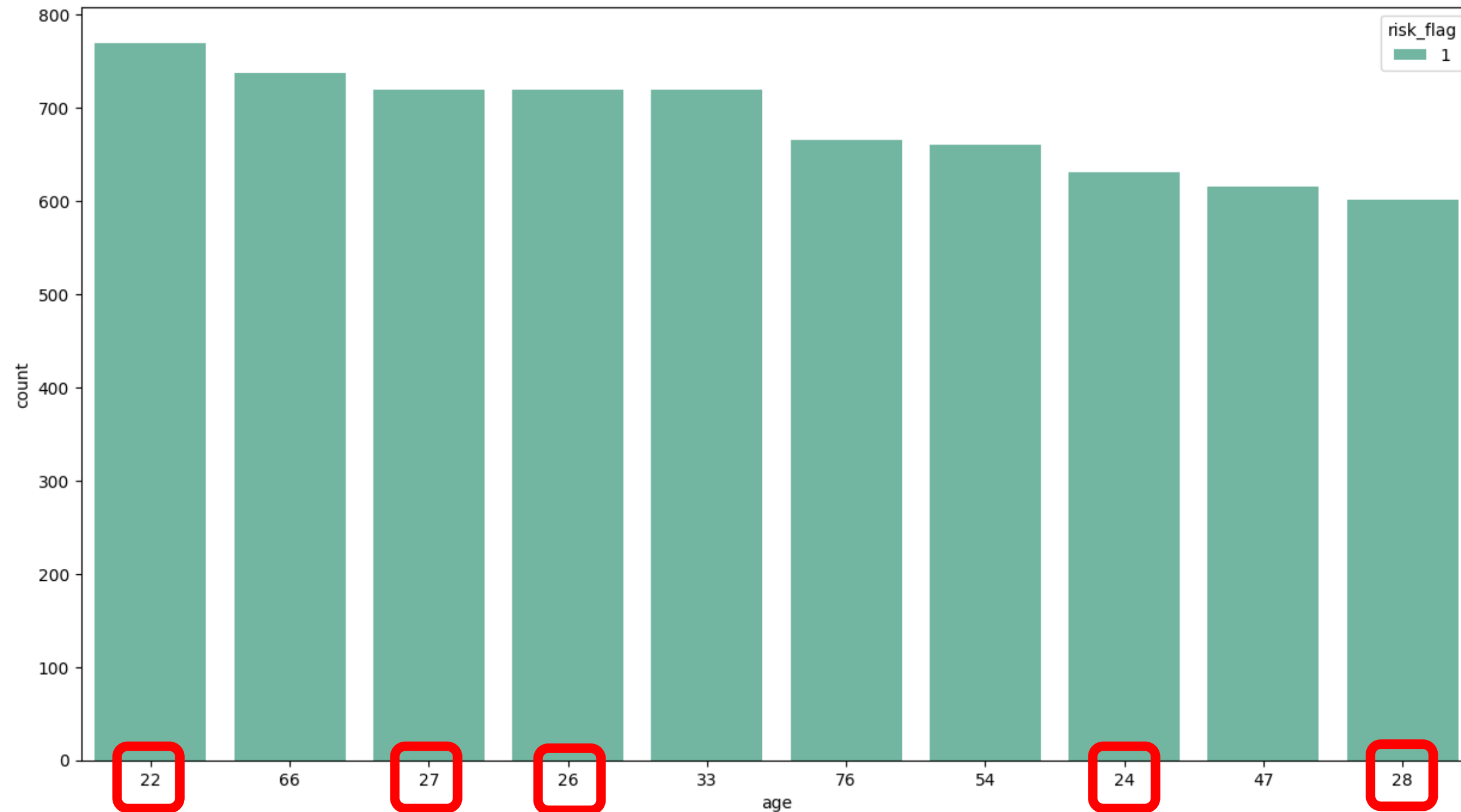
# EDA



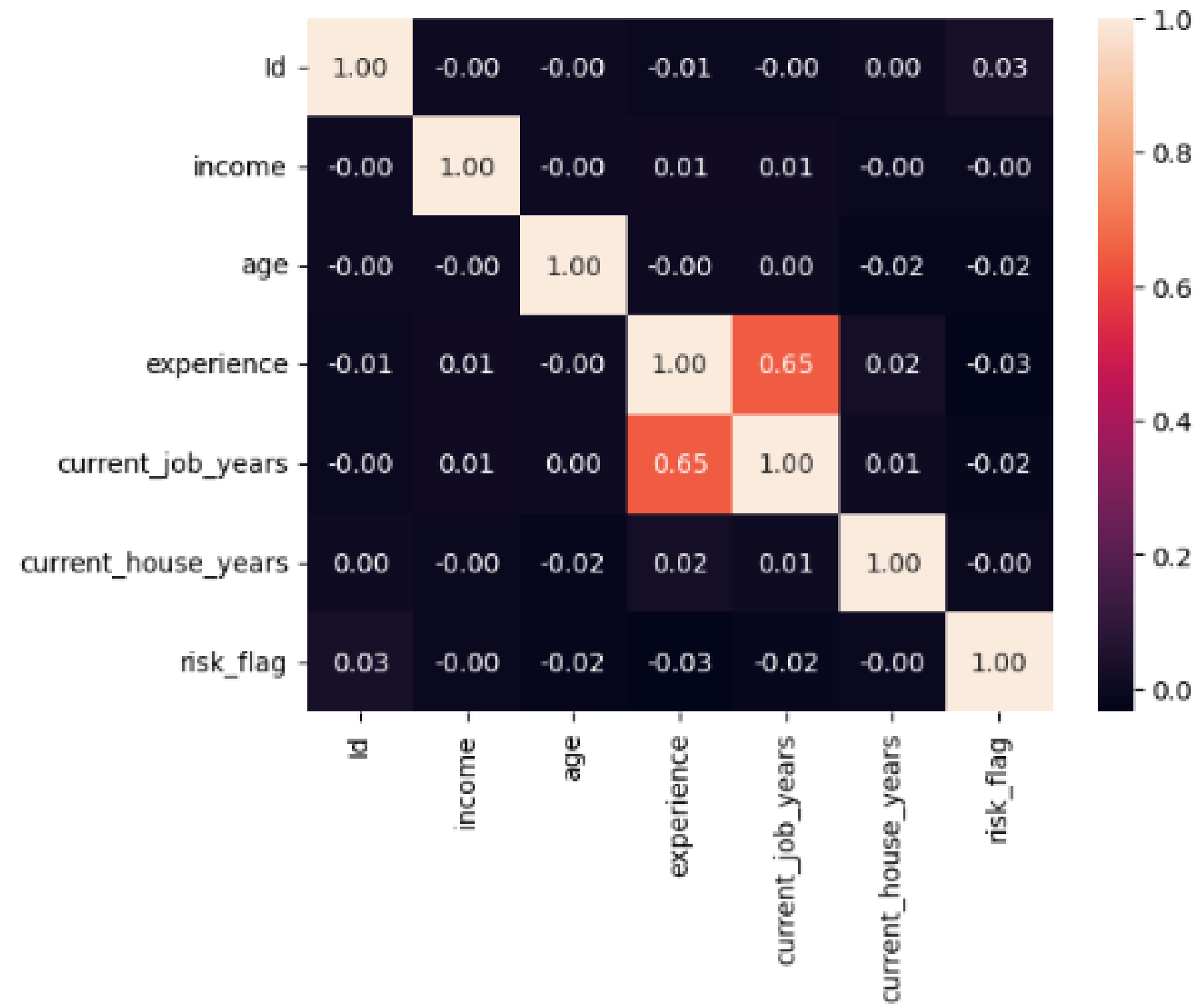
**There is no visible risk in certain age groups, which means that any age can be at risk of default.**



# EDA



**But, If we take the 10 ages with the highest risk, we can see that the 20s are more at risk than other ages.**



Based on the heatmap above, experience and current\_job\_years have a stronger correlation than the others, although the correlation is only 0.65, still < 0.7.

# EDA - Chi Square Test

```
Chi-square statistic: 111.89204667099783  
p-value: 3.773053705715196e-26  
marital status has a significant correlation with risk flag
```

```
Chi-square statistic: 182.98924138871385  
p-value: 1.8381930028370595e-40  
house_ownership has a significant correlation with risk flag
```

```
Chi-square statistic: 145.42374419378916  
p-value: 1.7350853850183746e-33  
car_ownership has a significant correlation with risk flag
```

The chi square test was conducted on categorical data such as married, house\_ownership, and car\_ownership. The result is that the variables married, house\_ownership, and car\_ownership **have a fairly strong correlation with the variable risk\_flag.**



The background of the image is a collage of various banknotes from different countries, including the United States, Canada, and the United Kingdom. The notes are overlapping and slightly blurred, creating a sense of depth and texture. The colors range from light greens and yellows to darker blues and browns.

# Data Pre Processing

# Data Pre Processing

In this case the methods that will be used are :

- **One hot coding** for the variables married and car\_ownership.
- **Frequency coding** for the profession, city, and state variables.
- **Scaling** on the age variable, experience variable, current\_job\_years, and current\_house\_years.

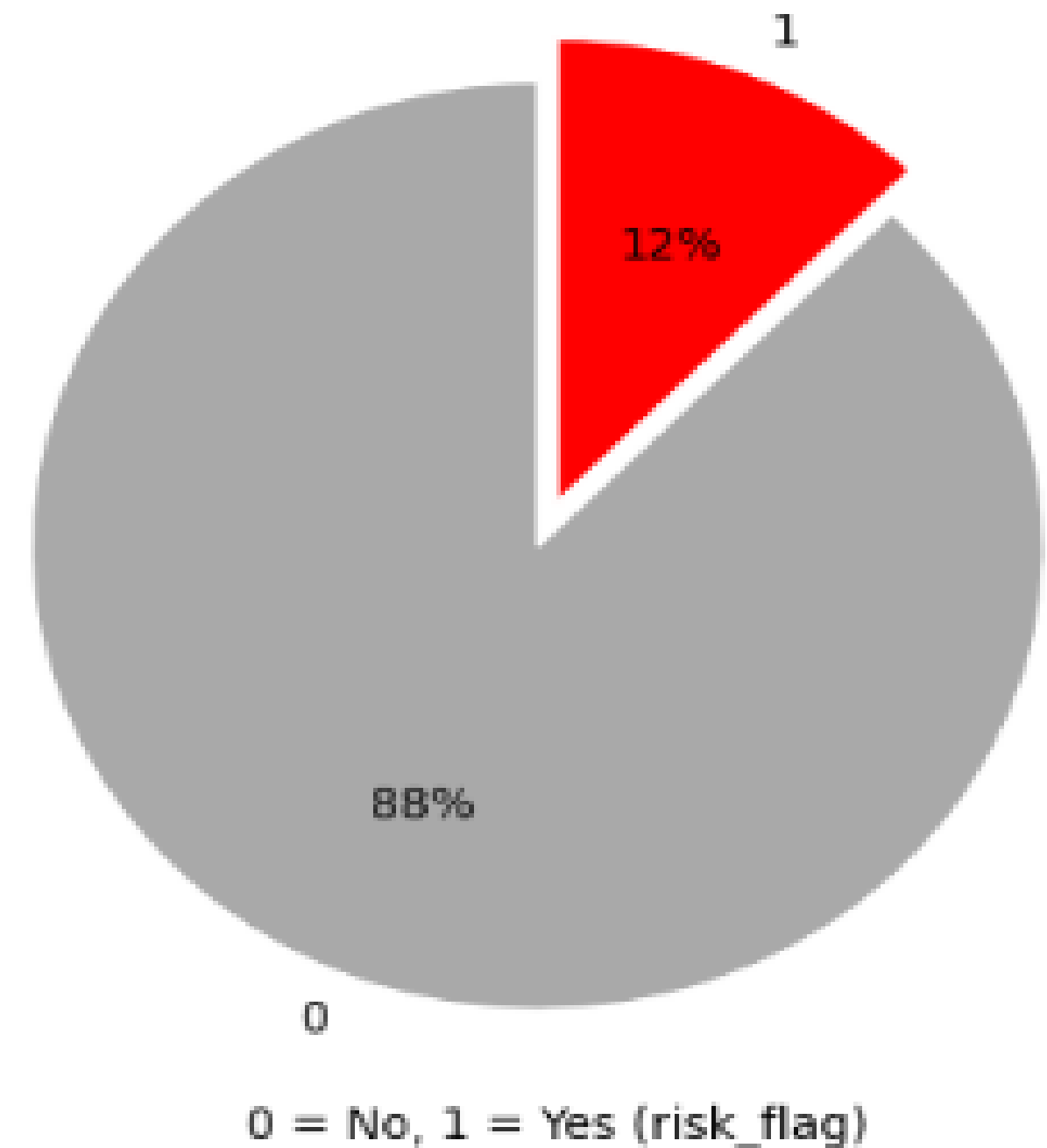
# Handling Imbalance

```
[83] #check distribution of target variable  
new_df1['risk_flag'].value_counts()
```

```
0    221004  
1     30996  
Name: risk_flag, dtype: int64
```

This dataset has a fairly imbalanced amount between risk\_flag that is worth 1 and risk\_flag that is worth 0, for that we need to do imbalance handling using smote.

The percentage of target variable





The background of the image is a collage of various banknotes from different countries, including the United States, Canada, and the United Kingdom. The notes are overlapping and slightly blurred, creating a sense of depth and financial context. A dark blue rectangular box is centered over the image, containing the title text in white.

# Modelling with Balance Data

# **Model with Balance Data**

**Machine learning model to be tested are :**

- 1. Logistic Regression**
- 2. Decision Tree**
- 3. KNN**
- 4. Random Forest**
- 5. Gaussian Naive Bayes**
- 6. Gradient Boosted Tree**



# Model with Balance Data

Because this data set is a **credit risk classification case** and the data is balanced, we will focus on the **recall** and **accuracy** evaluation matrix. The following is a recap of the model values:

Jenis Model	Accuracy	F1 Score	Recall	Precision
Logistic Regression	0.5071	0.2245	0.5751	0.1395
Decision Tree	0.8713	0.6157	0.8306	0.4891
KNN	0.8610	0.5002	0.5605	0.4516
Random Forest	<b>0.8869</b>	0.6339	0.7892	0.5296
Gaussian Naive Bayes	0.2662	0.2249	<b>0.8581</b>	0.1294
Gradient Boosted Tree	0.6444	0.2634	0.5126	0.1773

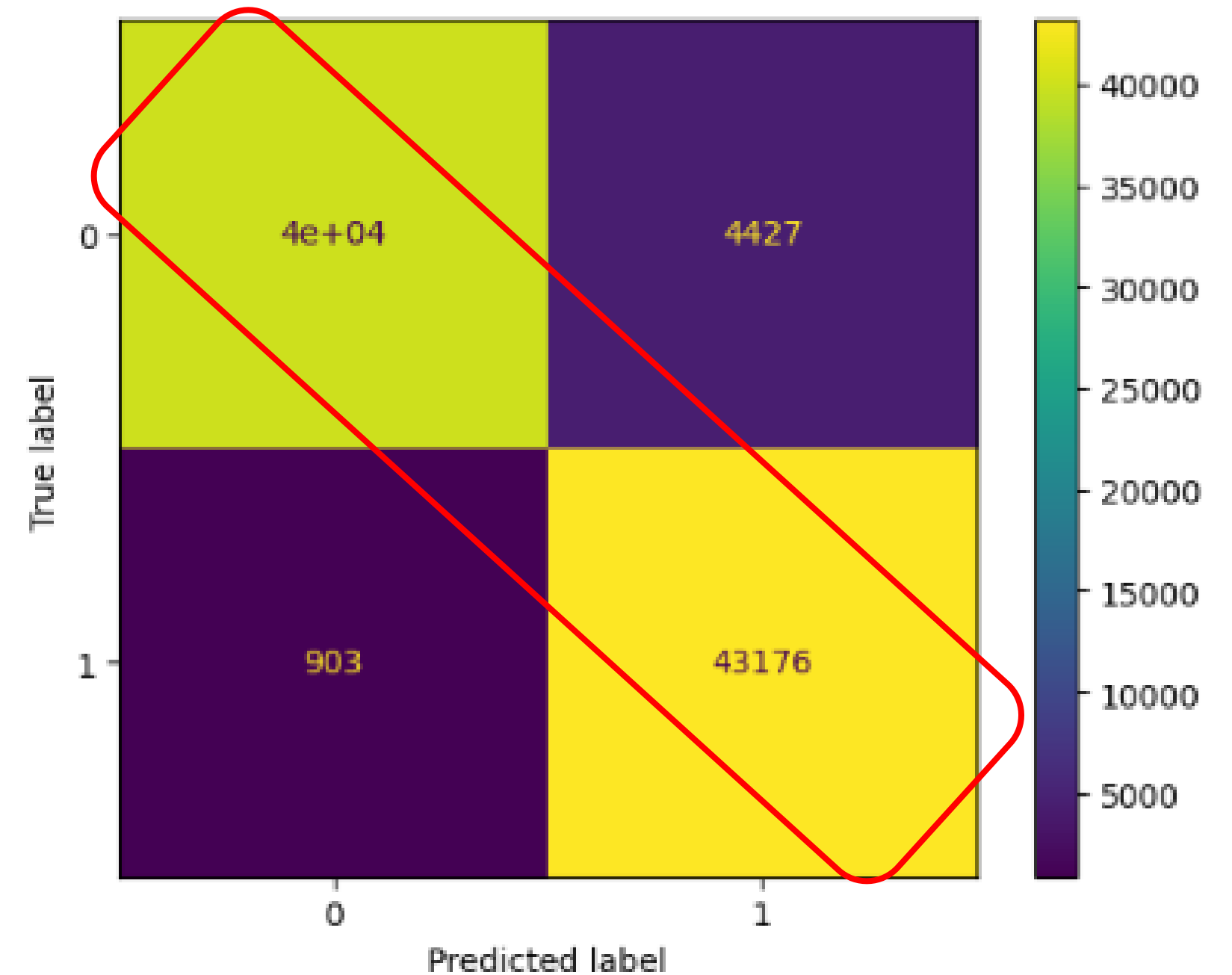
The accuracy value of Random Forest is higher than the recall of Gaussian Naive Bayes, so we will use **Random Forest** as the model because it has the highest evaluation matrix value.

# 88.69%

Based on the test results of several methods, the best model is Random Forest using matrix evaluation **Accuracy** with an evaluation rate is 88.69%.

# Confusion Matrix

- 4427 means 4427 not high risk that predicted as high risk.
- 903 means 903 have high risk that predicted as not high risk.
- 4e+04 means 4e+04 that correctly predicted as not high risk
- 43176 means 43176 that correctly predicted as high risk.



The background of the image is a collage of various banknotes from different countries, including the United States, Canada, and the United Kingdom. The notes are overlapping and slightly blurred, creating a sense of depth and texture. A semi-transparent dark blue rectangle is centered over the image, serving as a backdrop for the text.

# Dalex



# Dalex

Next, we will use dalex to see which variables are most highly correlated with the target variable after data processing.

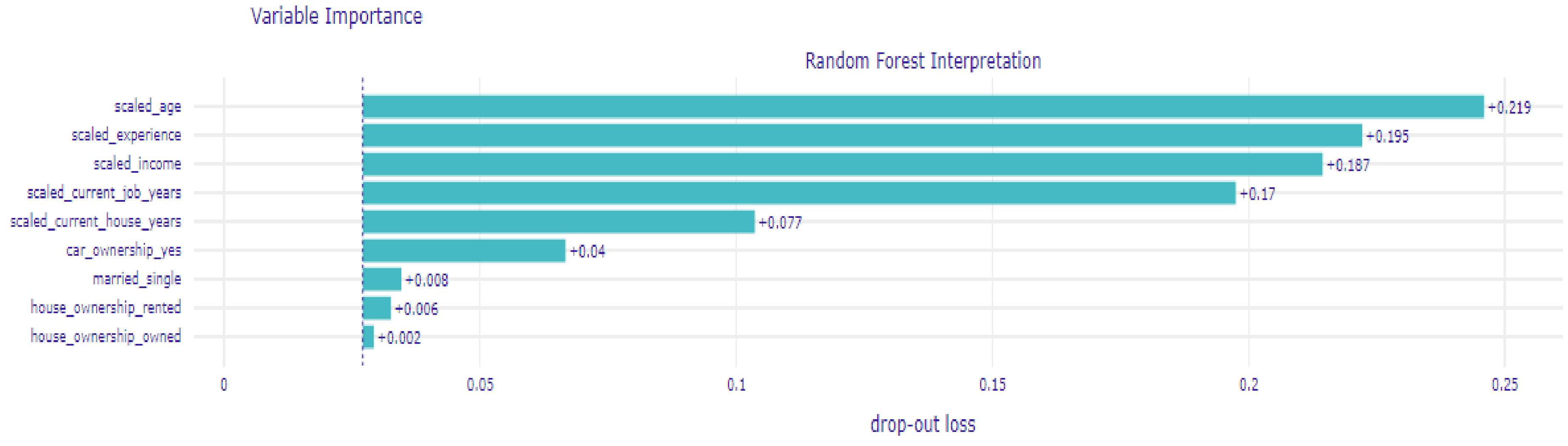
```
[98] # make predictions on a test dataset using a trained random forest model  
y_pred_rf = rf.predict(X_test)  
y_pred_rf
```

```
[ ] y_pred_rf = pd.Series(y_pred_rf)  
y_pred_rf.value_counts()
```

```
!pip install dalex
```

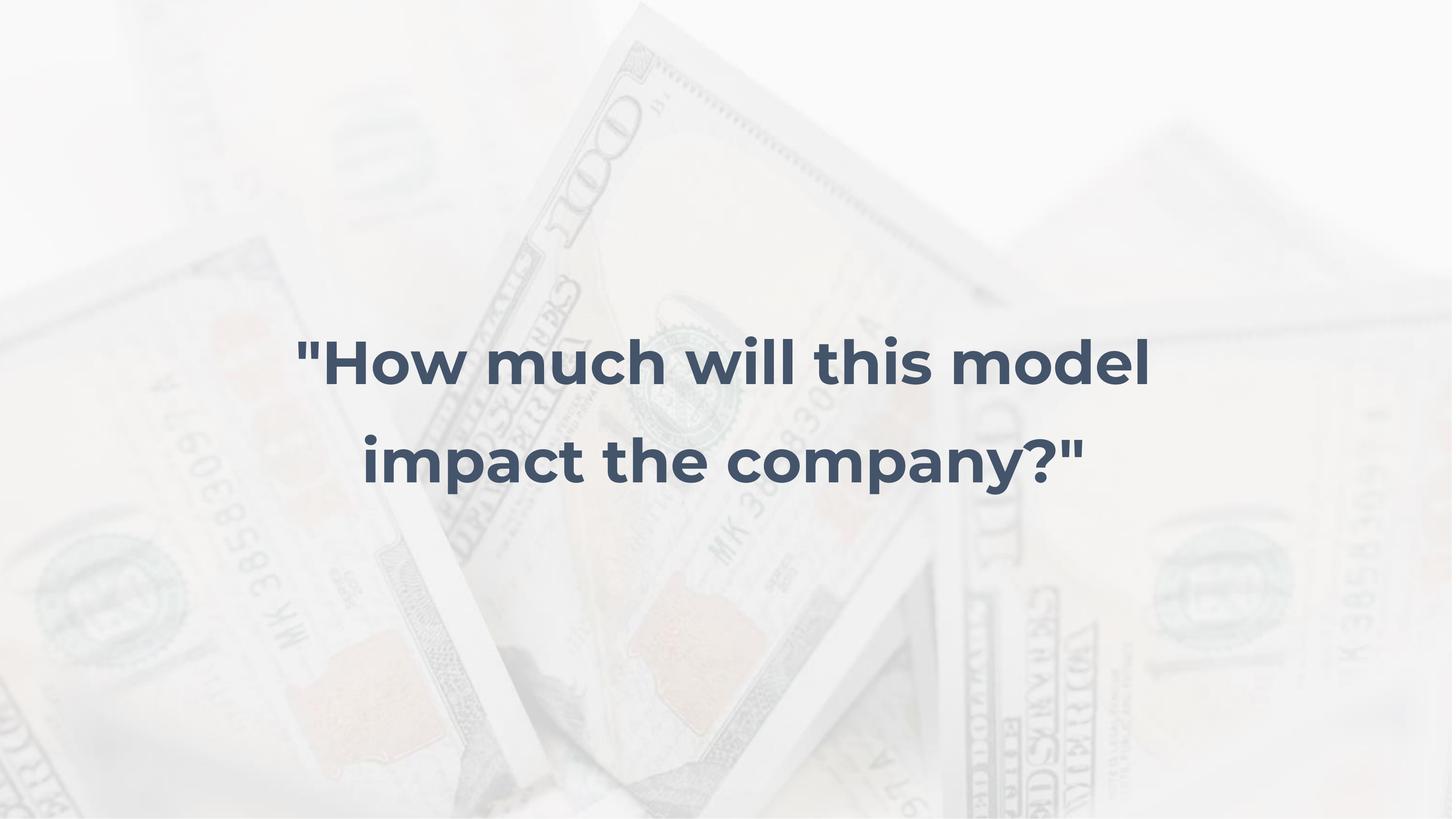
```
[101] # import dalex to explain complex model  
import dalex as dx
```

```
[102] ## initiate explainer for the best model  
var_exp = dx.Explainer(rf, X_train, y_train, label = "Random Forest Interpretation")
```



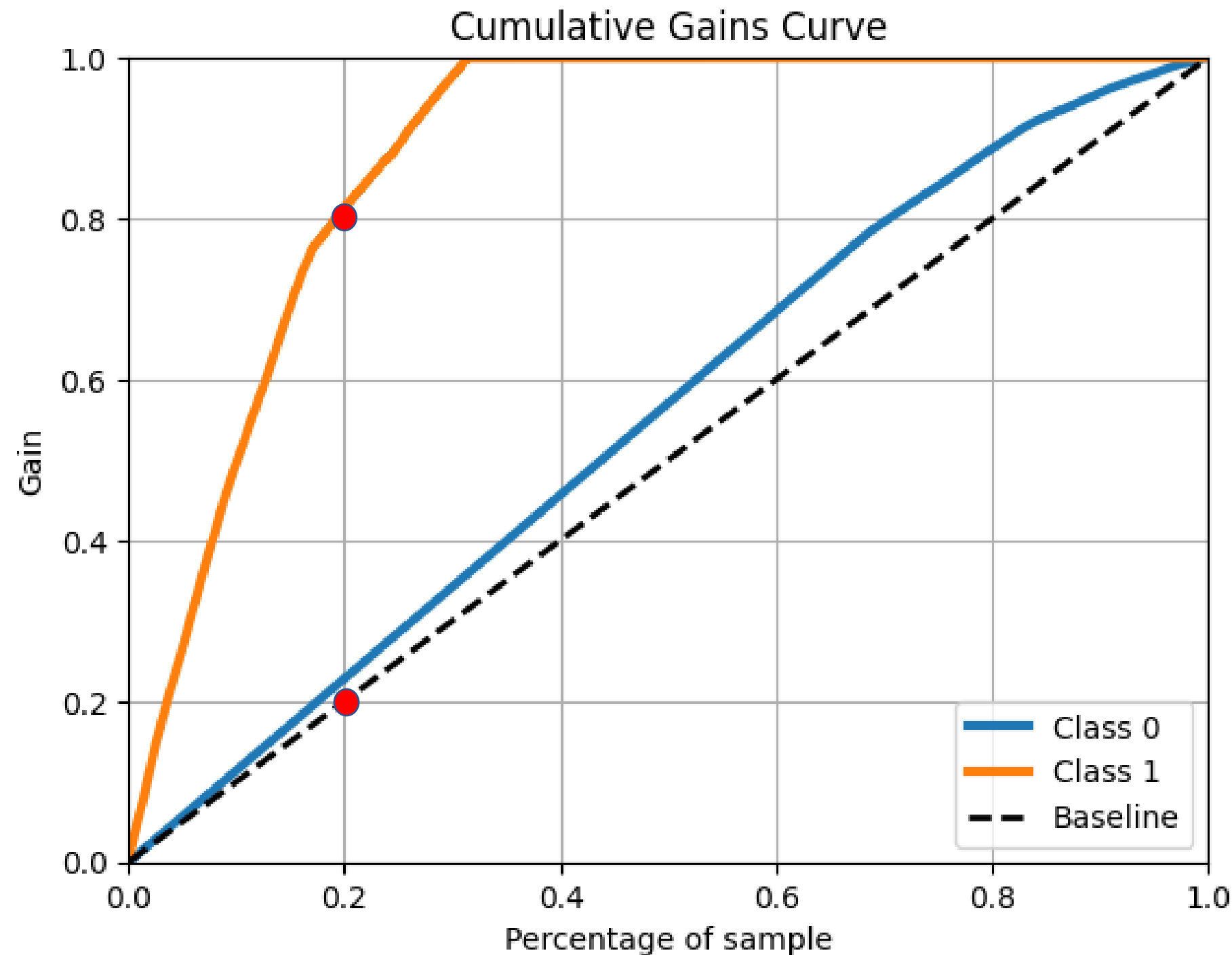
Based on the barplot above, age, income, `current job years, and experience have a high effect on the risk flag. And age has the highest effect.

# Business Case

The background of the image consists of several overlapping US dollar bills, including \$100 and \$10 bills, which are slightly out of focus. The bills are scattered across the frame, with some showing the portrait of the president and others showing the back of the bills with the Great Seal.

**"How much will this model  
impact the company?"**

# To find out we will use **Cumulative Gain Curve**



We can say that the results of the `plot_cumulative_gain` are good, for example, if we take a sampling of 0.2, then we have got in Class 1 is 0.8, which means 4x better performance than we do not use the model (baseline).

notes : The baseline on this curve is not the model baseline, but a prediction made without the model.



# Final summary

**Debtors who are unmarried, live in a rented house, and do not own a car are very risky to finance.**

**Variables that have a high effect on credit risk are customer age, married status, home ownership and car ownership.**

**In this case, the model using Random Forest with Accuracy evaluation matrix has the best performance, with an evaluation of 88.69%.**

**With a high level of evaluation, of course this prediction also has a high impact on business, at least it can help reduce losses > 88%.**



# Business Recommendations

**Avoid financing debtors who are unmarried, live in a rented house, do not own a car, aged between 20s years old.**

**However, if you still want to finance debtors with these categories, you may be able to submit additional data such as additional income, proof of ownership of other assets, or a guarantor from the family who is able and willing to take financial responsibility.**



The background of the image consists of several US dollar bills, including a \$100 bill and a \$10 bill, which are slightly out of focus. A semi-transparent dark blue rectangle is centered over the bills.

# Appendix



# Here are some of the methods I used to find the best model

1. - One Hot Encoding for Married, home ownership, car ownership  
- Frequency Encoding for Profession, city, state
1. - One Hot Encoding for Married, car ownership  
- Weight of Evidence (WoE) Encoding for Profession, city, state, home ownership
1. - One Hot Encoding for Married, home ownership, car ownership  
- Frequency Encoding for Profession, city, state  
- Do standard scaller for all numerical data  
( this methode has the best model )
1. - One Hot Encoding for Married Married, car ownership  
- Weight of Evidence (WoE) Encoding for Profession, city, state, home ownership  
- Do standard scaller for all numerical data

Do the method repeatedly for balance and imbalance data, and also use other matrix evaluations such as Accuracy, F1 score, Recall and Precision on each method to see which evaluation matrix is better.

# Comparison between baseline and model after data processing

Jenis Model	Baseline				After Processing Data			
	Accuracy	F1 Score	Recall	Precision	Accuracy	F1 Score	Recall	Precision
Logistic Regression	0.8759	0.0	0.0	0.0	0.5071	0.2245	0.5751	0.1395
Decision Tree	0.8814	0.5441	0.569	0.5219	0.8713	0.6157	0.8306	0.4891
KNN	0.8897	0.5338	0.5089	0.5612	0.8610	0.5002	0.5605	0.4516
Random Forest	0.8960	0.5642	0.5421	0.5881	0.8869	0.6339	0.7892	0.5296
Gaussian Naive Bayes	0.8759	0.0	0.0	0.0	0.2662	0.2249	0.8581	0.1294
Gradient Boosted Tree	0.876	0.0019	0.0009	0.6667	0.6444	0.2634	0.5126	0.1773



# References

1. <https://medium.com/responsibleml/visualize-ml-model-bias-with-dalex-b63f182cd649>
2. <https://medium.com/@pararawendy19/memahami-metrik-pada-pemodelan-klasifikasi-29cd5b738ee7>
3. <https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f>
4. <https://www.ceicdata.com/id/indicator/indonesia/non-performing-loans-ratio>
5. <https://www.idscore.id/faq/detail/apa-itu-non-performing-loan-npl-serta-dampak-negatif-bagi-lembaga-keuangan>

# Thank you very much!

Final Project Report  
Data Science Bootcamp Batch 18  
[dibimbing.id](https://dibimbing.id)

