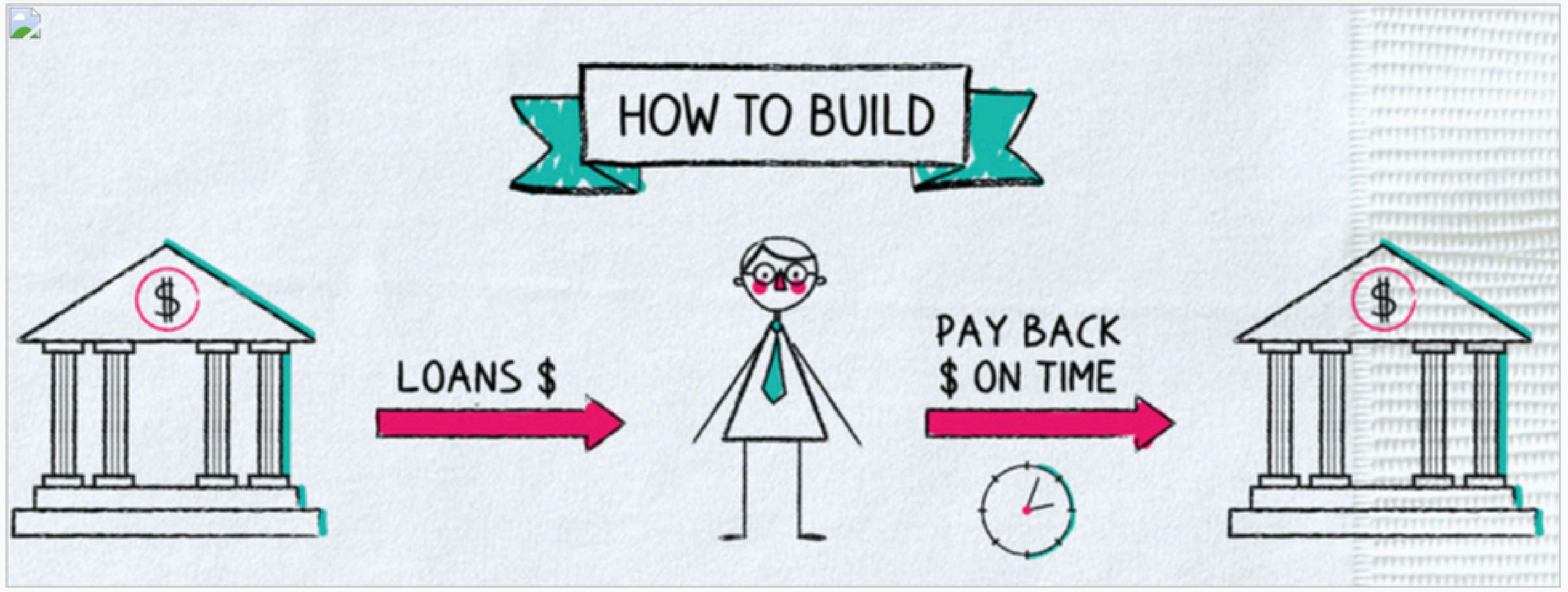
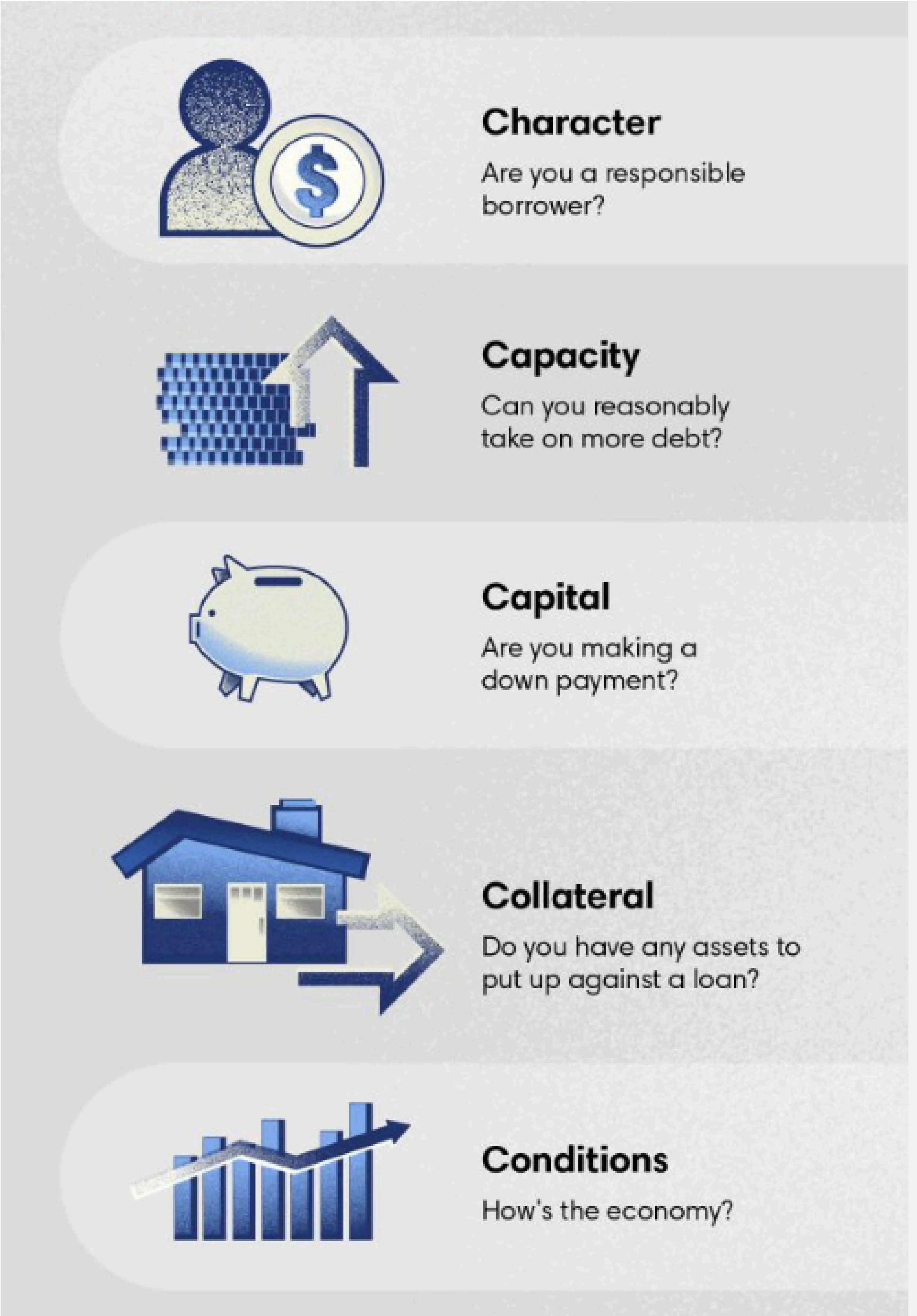
Smart Credit Predictions

Shuteng, Dean, Leila

In general, how do banks decide if you're a good credit risk for a loan?

They use the 5 C's of credit





Our goal: What features cause people to have overdue credit payments?

Features Explored:

- **Employment & Income**
  - Years employed
  - Total income

- **Personal Information**
  - Age
  - Education Level
- **Family and assets**
  - Family/Marital Status
  - Properties Owned

Prediction:

We will be predicting whether a credit card applicant is a "**good**" or "**bad**" client, based on historical data and features derived from their application

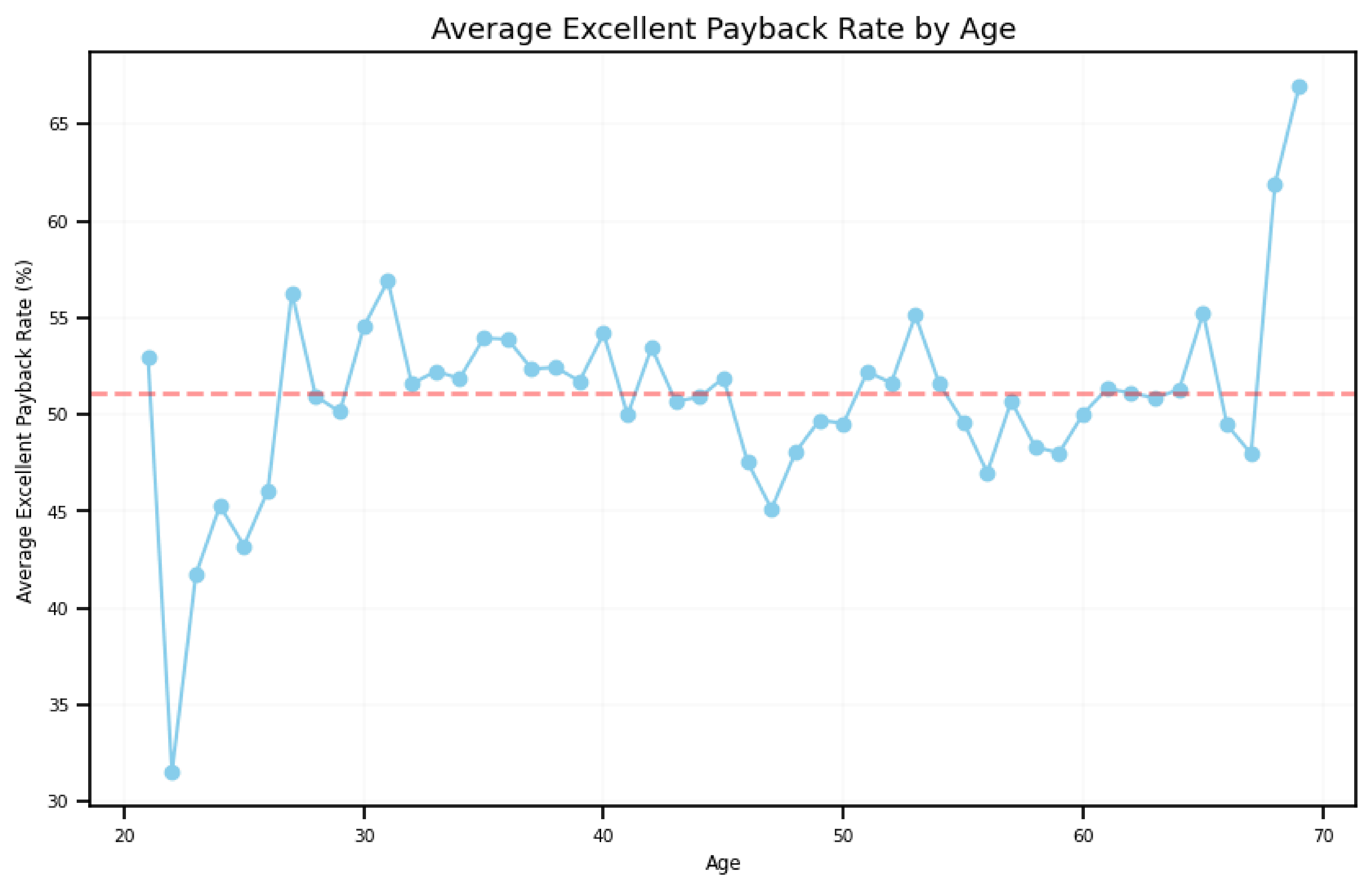
Our Data

ID	Income	Age	Marital Status	Student	On Time
43583	121500	45	Married	0	0
281472	360000	30	Single	1	1
570404	126000	26	Single	1	0
648705	180000	58	Married	0	1
760580	107500	63	Married	0	0

01

Data Exploration

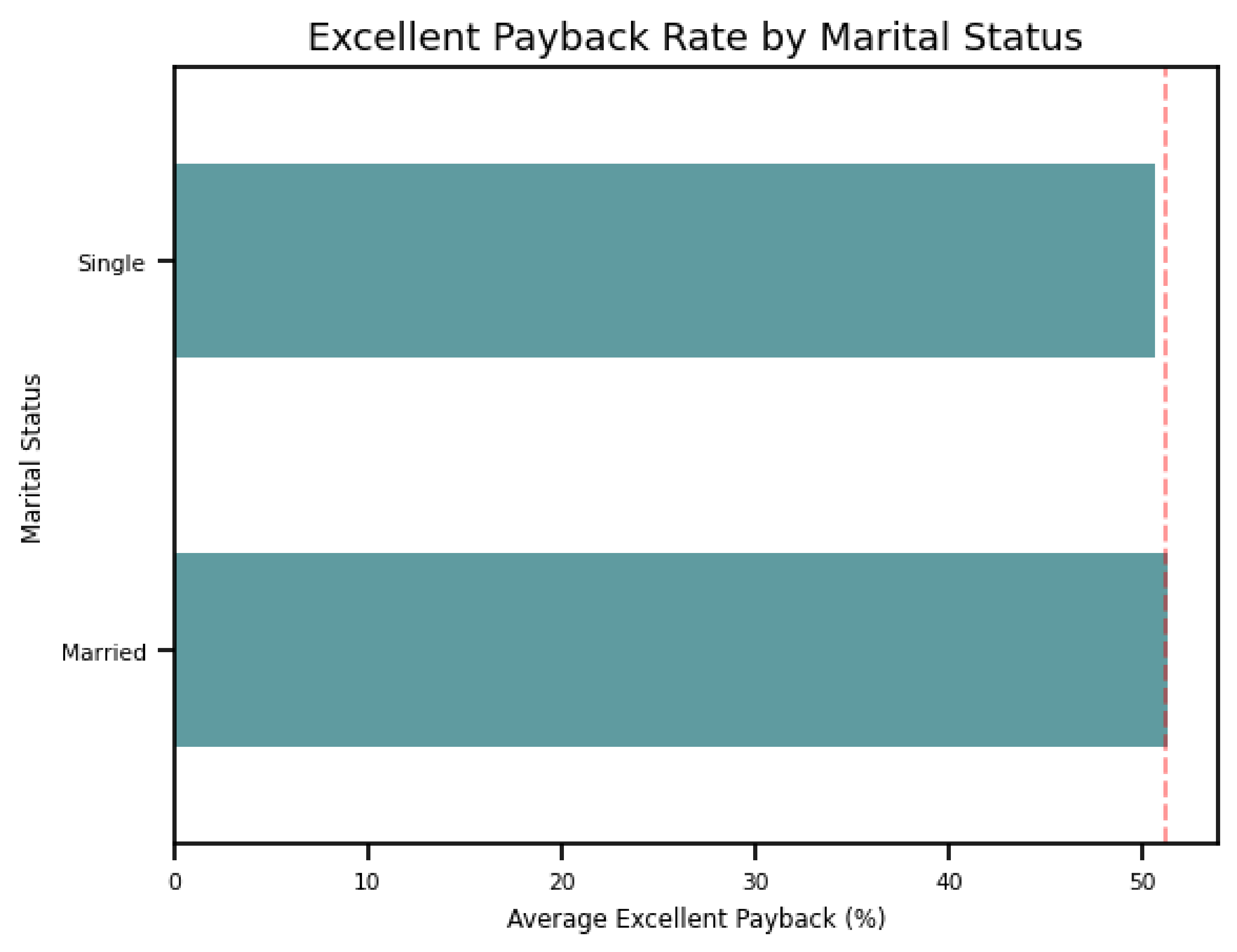
Age



Average: 51%

- **< 30** lower payback rate
- **30-70** hover around the same
- **> 70** highest payback rate

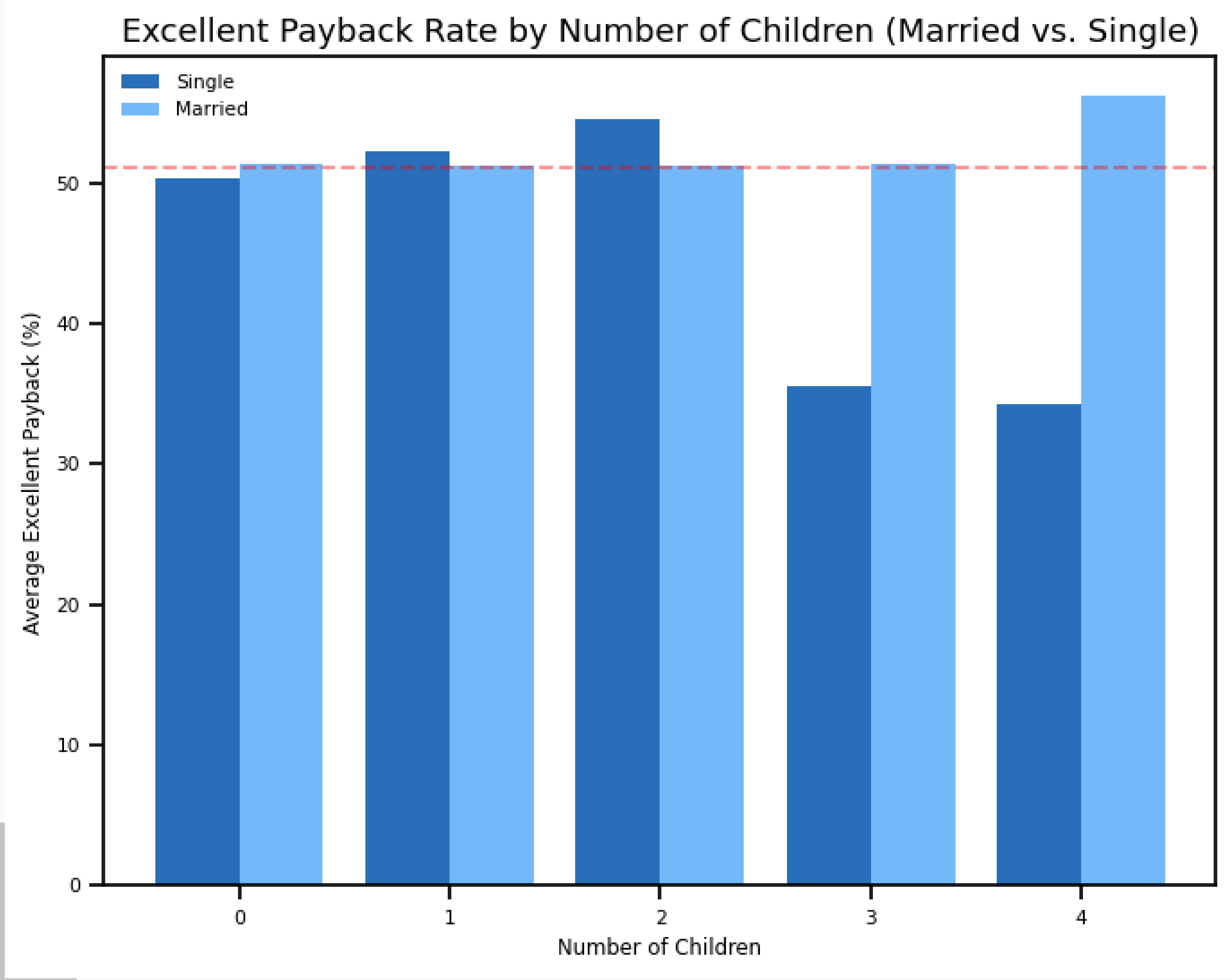
Family Status



Average: **51%**

- Overall, **married & single** individuals similar payback rate
    - However, **single** slightly below average

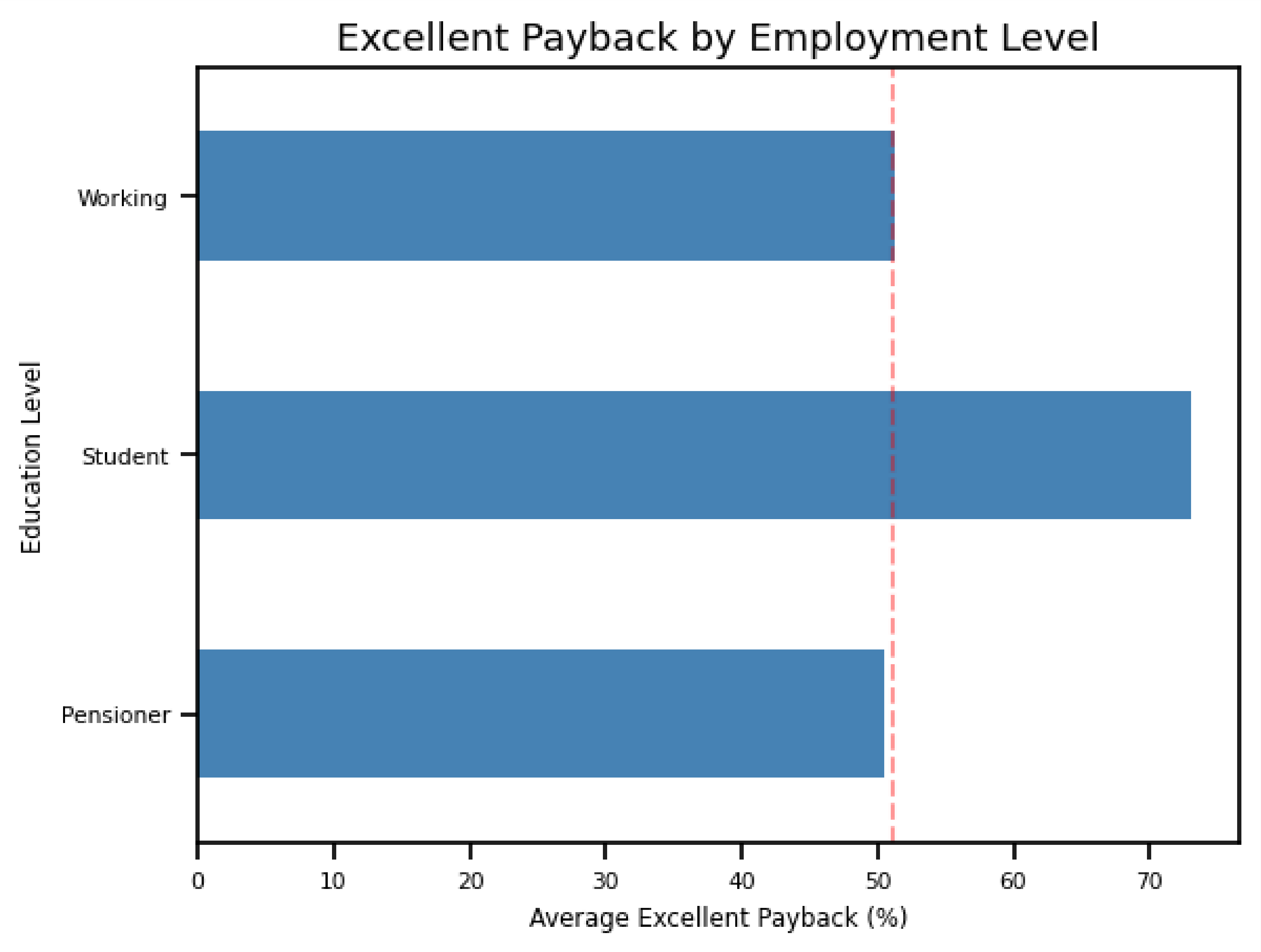
Family Status



Average: **51%**

- **single** individuals:
  - significantly lower payback rate 3+ children
- **married** individuals: average or above
  - dual-income household

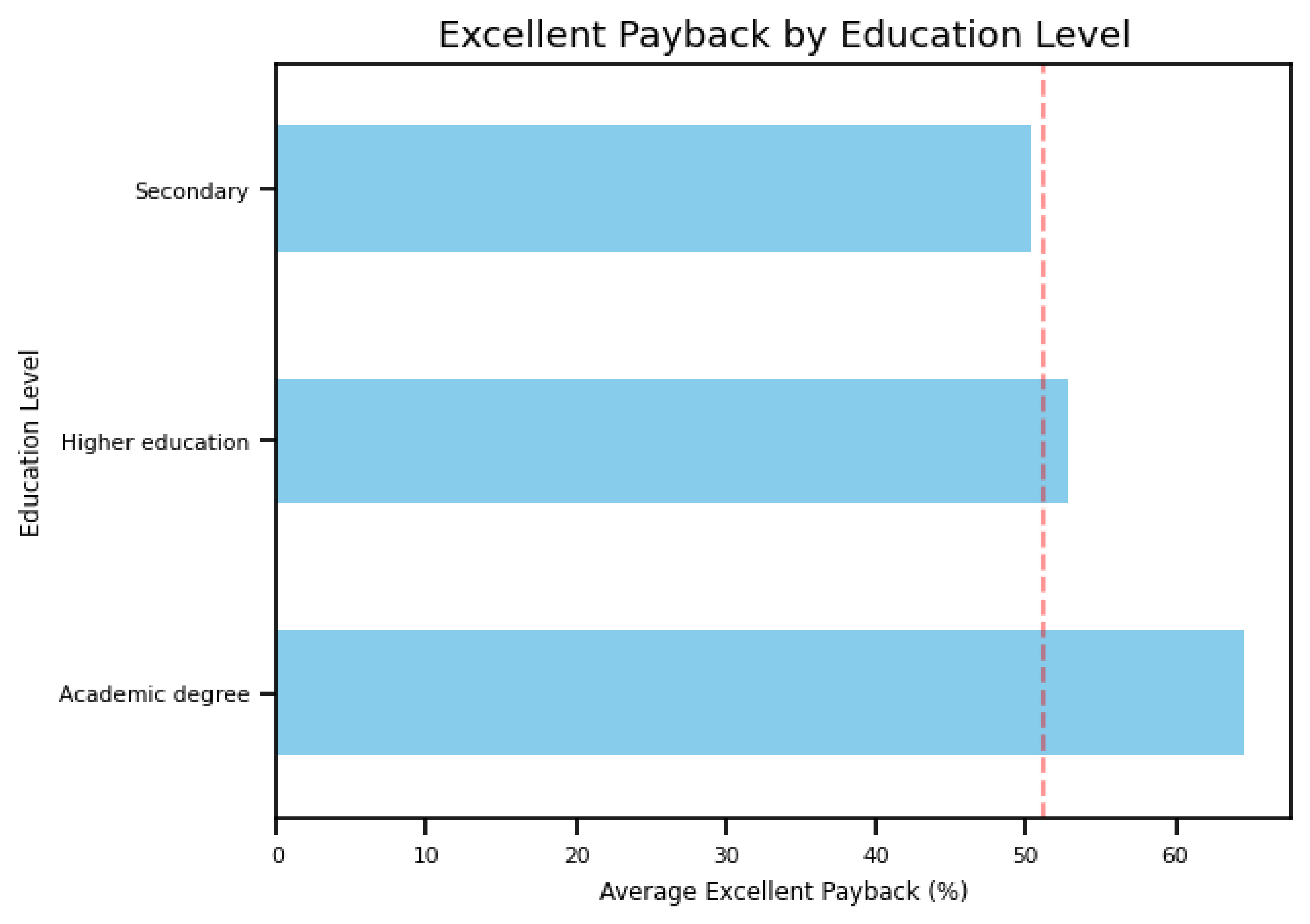
Employment Type



Average: **51%**

- Less **students** in data
  - shorter credit history
- **Working** & **pensioner** around average

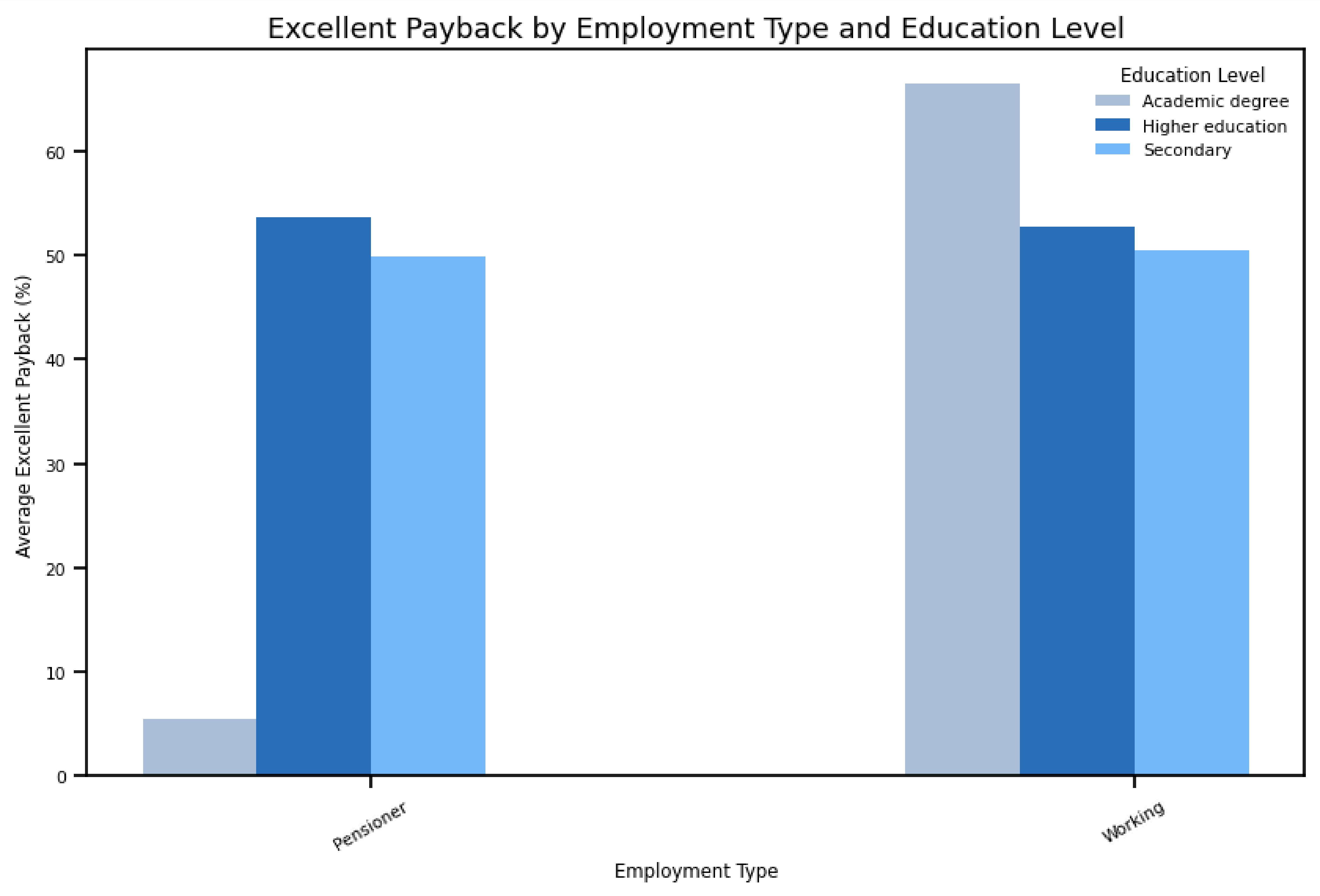
Education Level



Average: **51%**

- With college degree higher payback rate
  - **Higher education**: undergradute degee
  - **Academic degree**: postgraduate degree
- **Secondary**: completion of high school/middle school

Employment Type & Education Type



Average: 51%

- **working** with **postgraduate degree** highest payback rate

Data Preparation

Dataset Merge

- Combined two datasets using **inner join** to ensure only applicants with complete data were included

Created Target Variable

- Caculated each applicant's **on-time** payment rate
- Defined applicants as **"Good Clients (1)**" if their on-time payment rate **≥ 75%** and "**Bad Clients (0)**" otherwise

Data Preparation

Processed Categorical Variables

- Used **one-hot encoding** to convert **categorical** features (e.g., education level, employment type, marital status) into **binary** (0/1) columns

Feature Scaling

- Applied **StandardScaler** to standardize numerical features like years employed, total income, and age

Handled Class Imbalance

- Used **SMOTE** (Synthetic Minority Over-sampling Technique) to balance the dataset by oversampling the minority class

Definitions:

Models

Decision tree

predictive model that splits data into **branches** to make decisions based on features

Random Forest

group of decision trees that predicts the class (or value) of a data point by **aggregating** the results of multiple trees

Definitions:

Confusion Matrix

Simple table that shows how well a classification model is performing by comparing its **predictions** to the **actual** results

True \ **Predicted**	Bad (0)	Good (1)
Bad (0)	**TN**	**FN**
Good (1)	**FP**	**TP**

02

Decision Tree

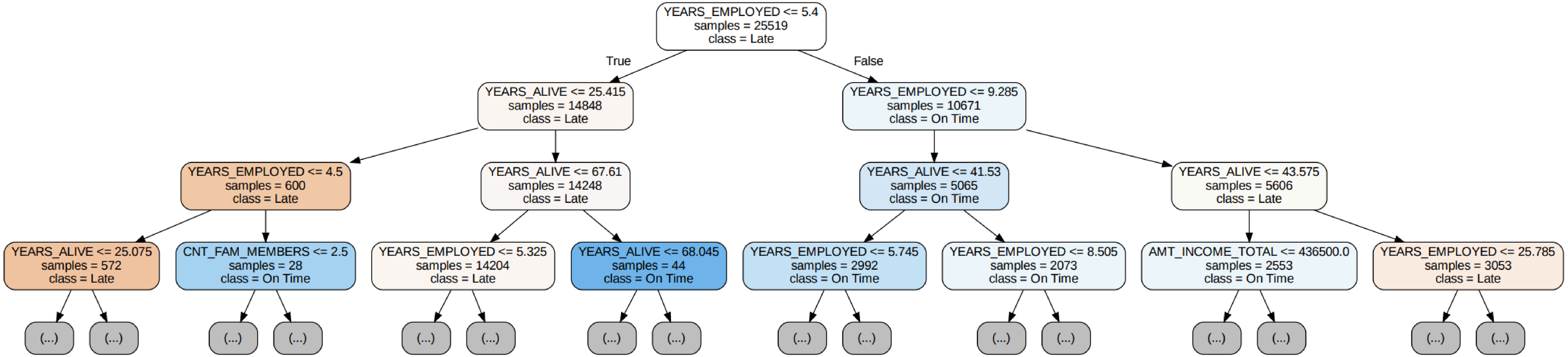
Decision Tree

Columns selected: Car Ownership, House Ownership, Total Income, Number of Children, Number of Family Members, Years of Employment, Age

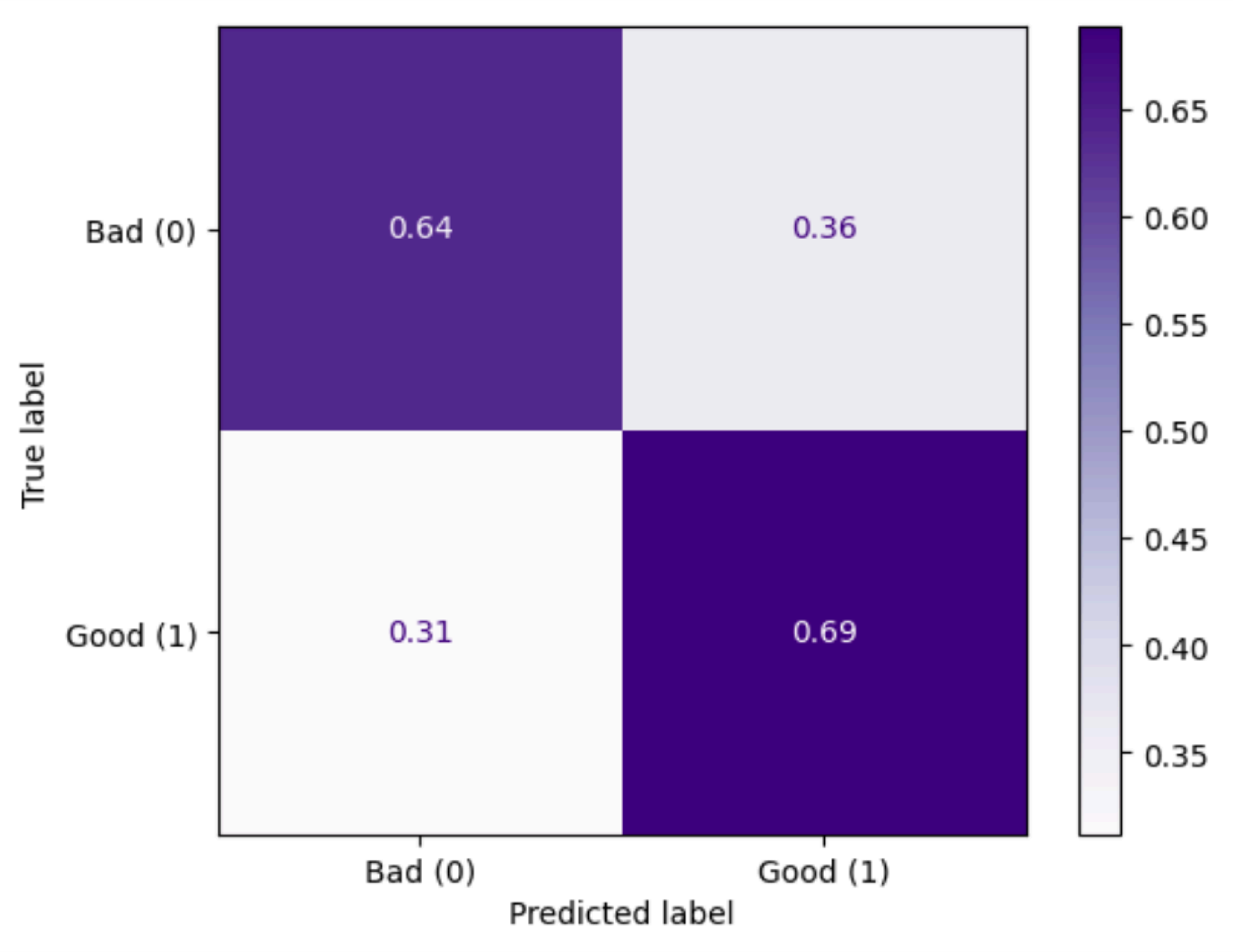
Chosen based on their **relevance to financial responsibility**

Predicting: On time payments

Decision Tree



Confusion Matrix



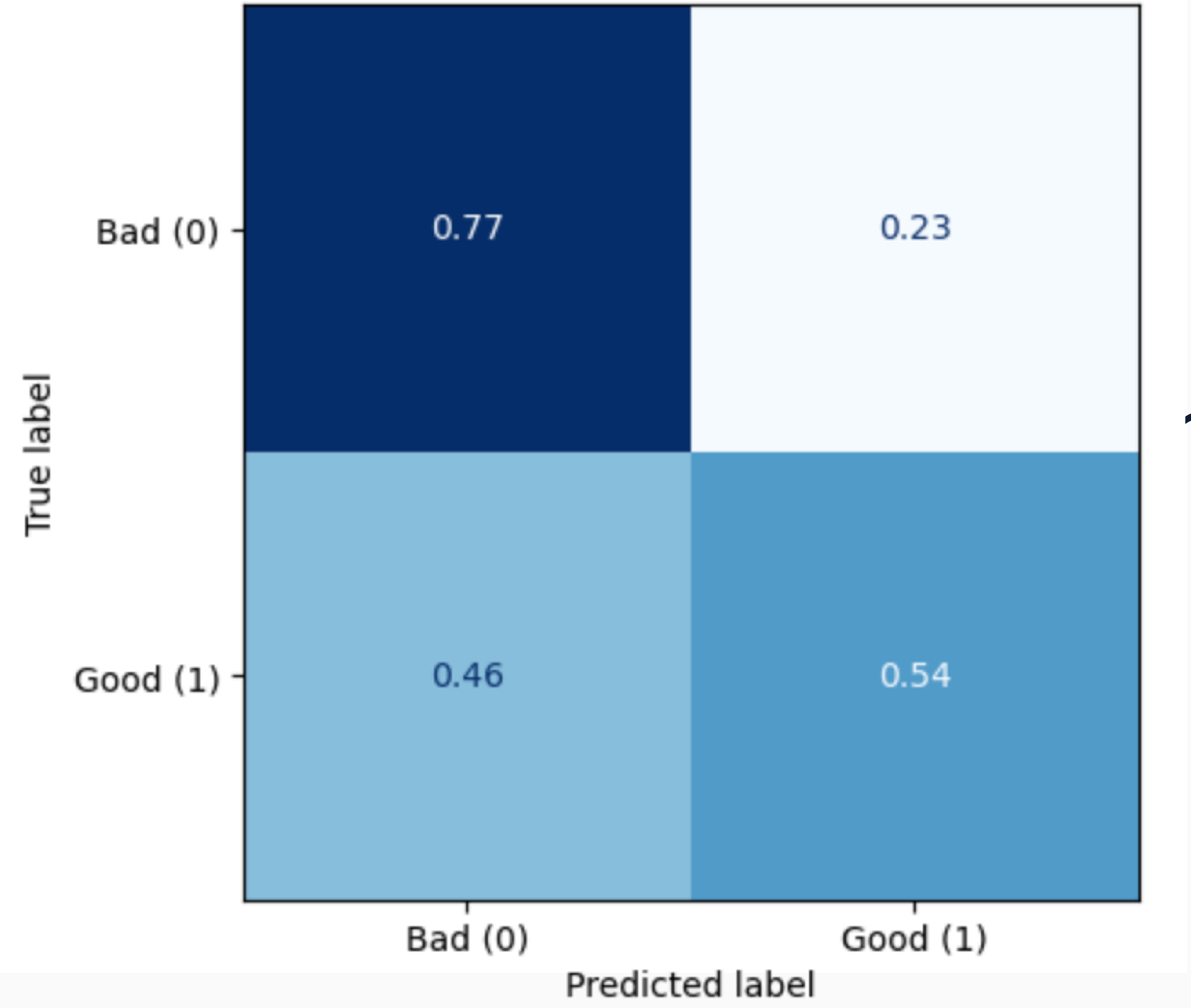
~ 67% accurate

03
Random Forest

Random Forest

Columns selected: Car Ownership, House Ownership, Total Income, Number of Children, Number of Family Members, Years of Employment, Age, **Gender, Education Level, Marital Status, Housing Status, Employment Type**

Confusion Matrix



~ 70% accurate

Model Comparison

Metrics	Decision Trees	Random Forest
Accuracy	0.67	0.70
True Positive Rate	**0.69**	0.54
True Negative rate	0.64	**0.77**
False Positive Rate	0.36	0.23
False Negative Rates	0.31	0.46
Precision	0.48	0.54

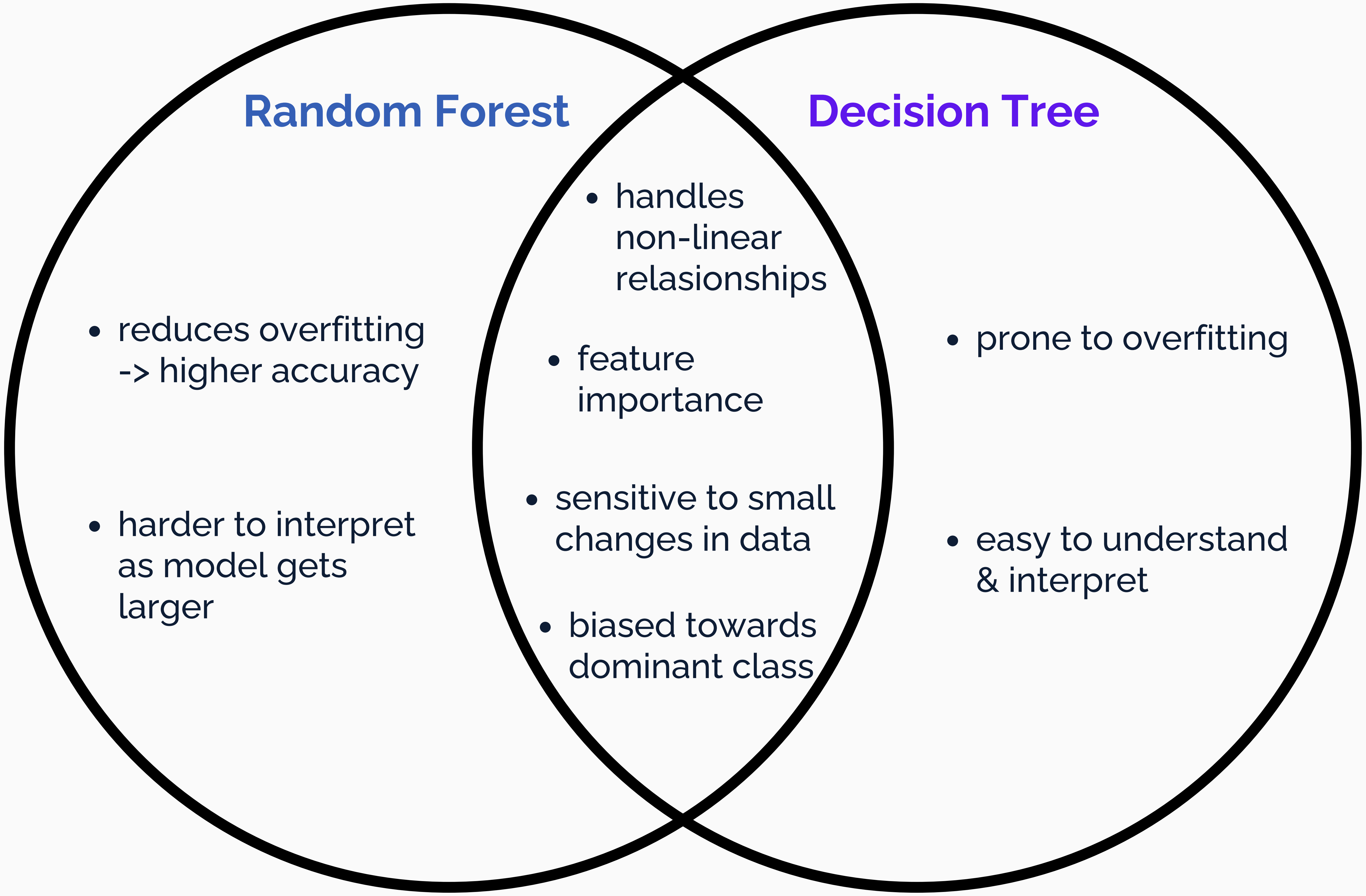
Random Forest Model:

- True Negatives:
  - 71% only completed high/middle school
  - > 80% of applicants have at least one child

Decision Trees Model:

- True Positives:
  - 70% of people had no children
  - 67% of people owned their home

Model Comparison



Conclusion

Analysis

- higher education = **higher** on-time payback rate
- more children = **lower** on-time payback rate
- home ownership = **higher** on-time payback rate

Models

- **Decision Tree**: can identify specific patterns for **true positives** (paying on time)
- **Random Forests**: more robust, better handling **true negatives** (not paying on time) due to their ensemble nature & ability to generalize

Thank You!