PREDICTING BANK CREDIT RATING

SATURN'S RINGMASTERS Consulting (Team 2)
19 May 2023

OUTLINE

- Introduction
- Data
- Methods
- Variable selection
- Results
- Limitations & suggestions
- Conclusion
- Questions

OBJECTIVES

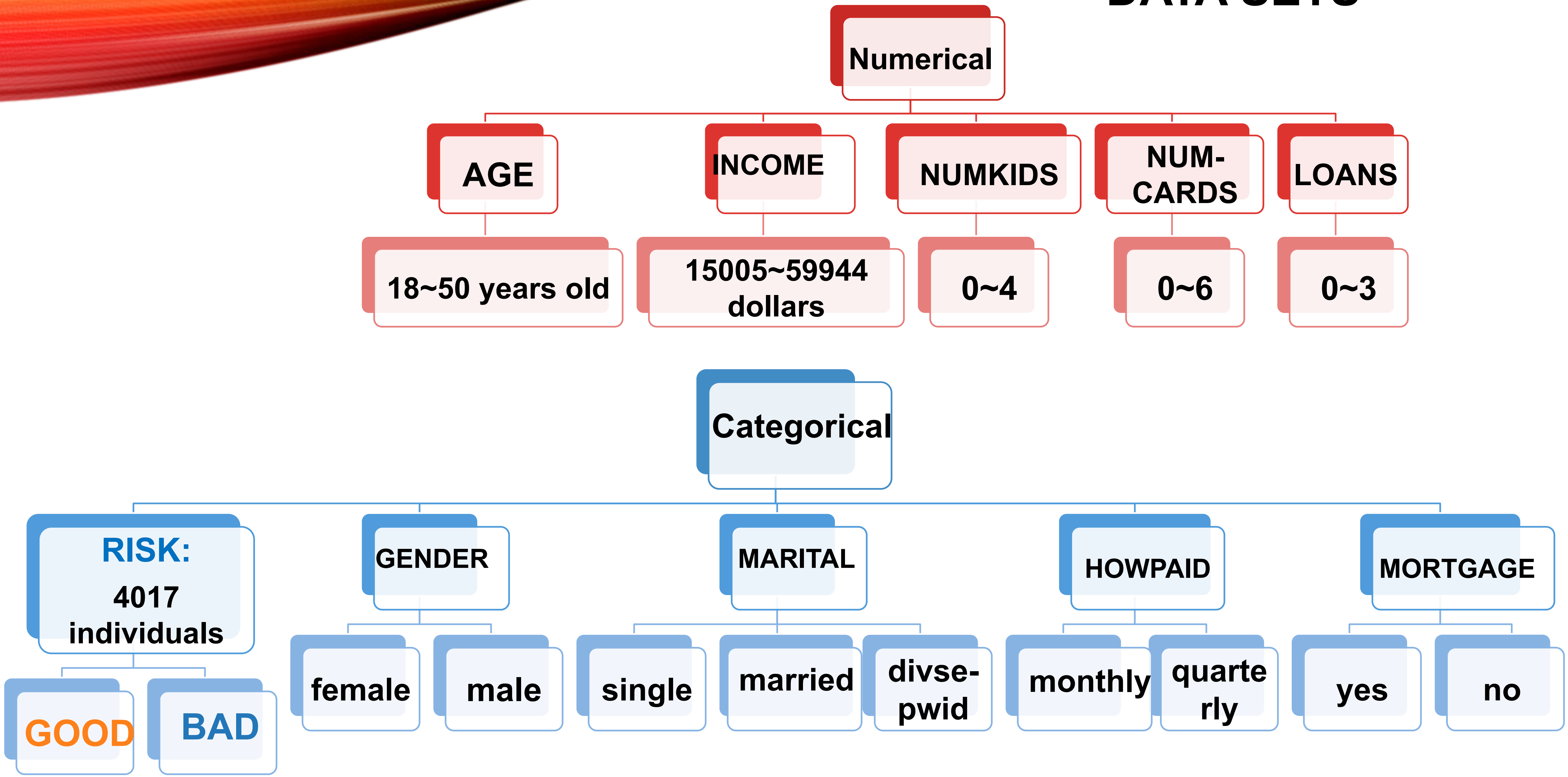
Effectiveness of the credit rating detection

Predict an individual's credit rating

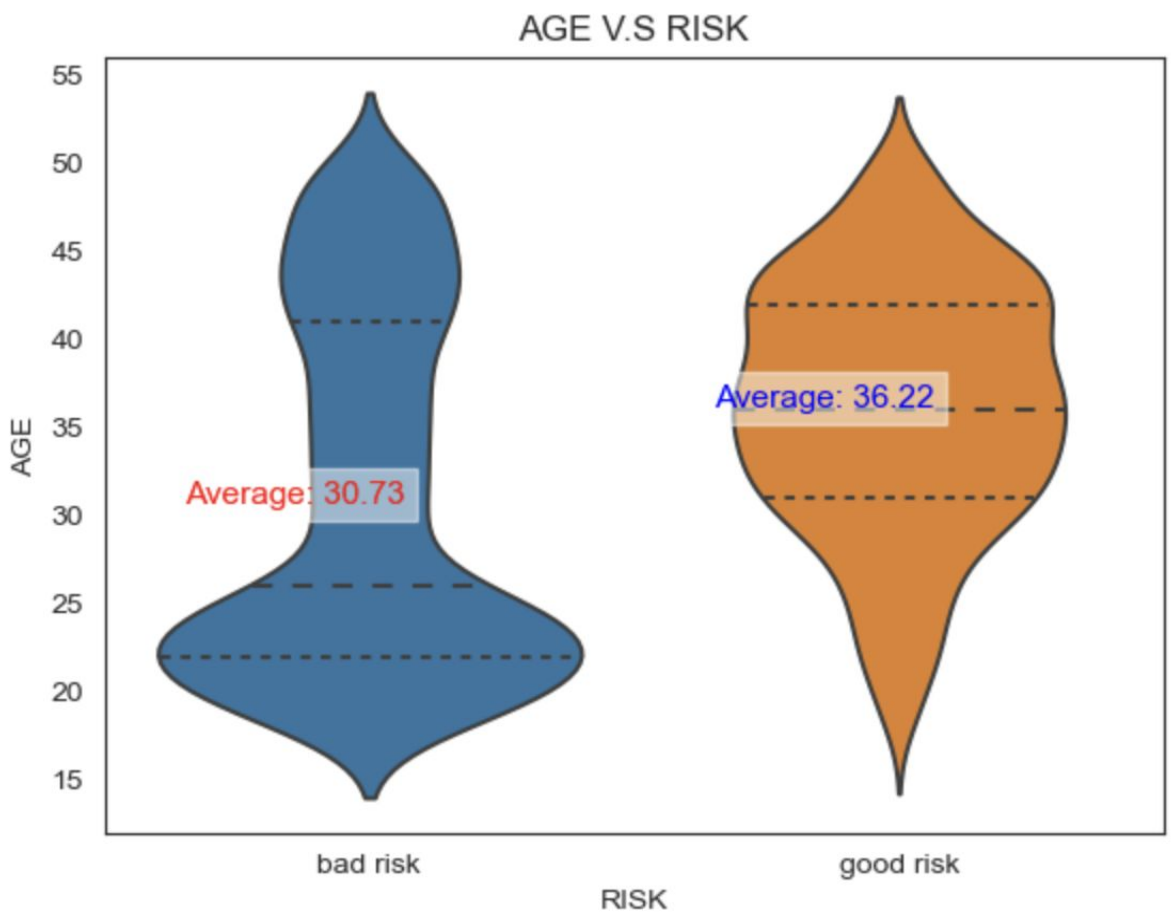
VARIABLES

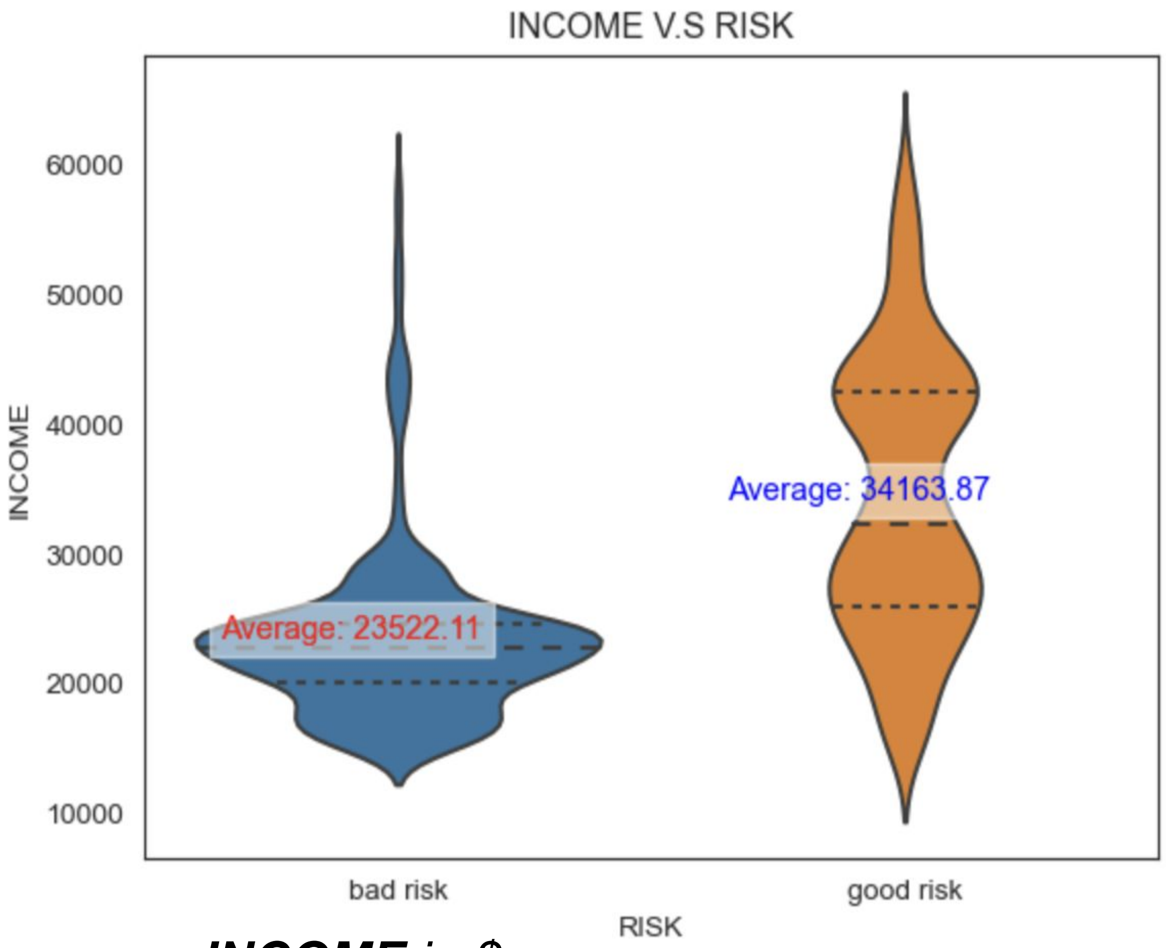
- **AGE:** age in years
- **INCOME:** annual income (in dollars)
- **GENDER:** gender ('female' or 'male')
- **MARITAL:** marital status ('single', 'married' or 'divsepwid')
- **NUMKIDS:** number of kids
- **NUMCARDS:** number of cards
- **HOWPAID:** how paid ('monthly' or 'quarterly')
- **MORTGAGE:** mortgage ('yes' or 'no')
- **LOANS:** number of existing loans

DATA SETS


Numerical
AGE
18~50 years old
INCOME
15005~59944 dollars
NUMKIDS
0~4
NUM-CARDS
0~6
LOANS
0~3
Categorical
RISK:
4017 individuals
GOOD
BAD
GENDER
female
male
MARITAL
single
married
divse-pwid
HOWPAID
monthly
quarterly
MORTGAGE
yes
no

AGE AND INCOME





AGE *in years*

Bad risk: 20～30

Good risk: 35~40

INCOME *in $*

Bad risk:

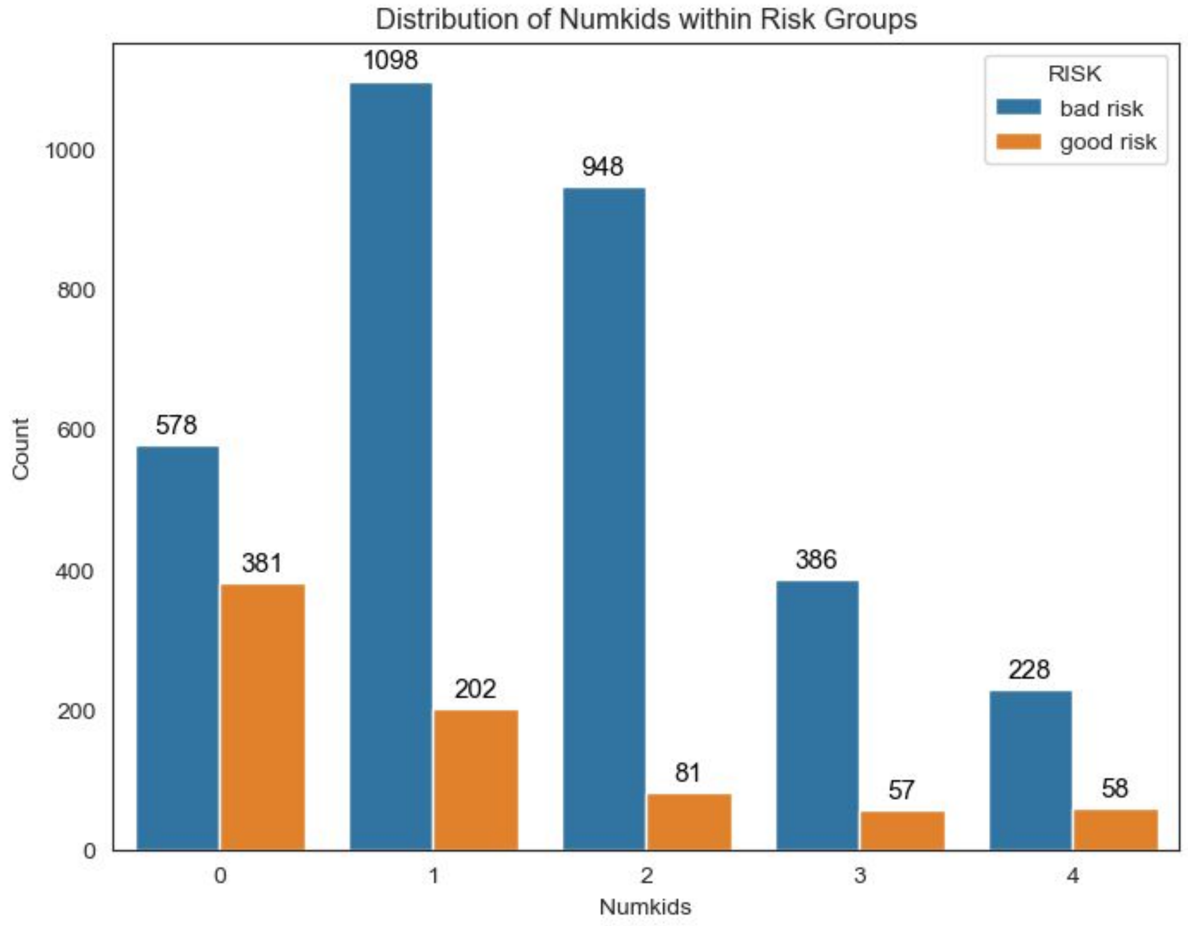
20000 to 25000

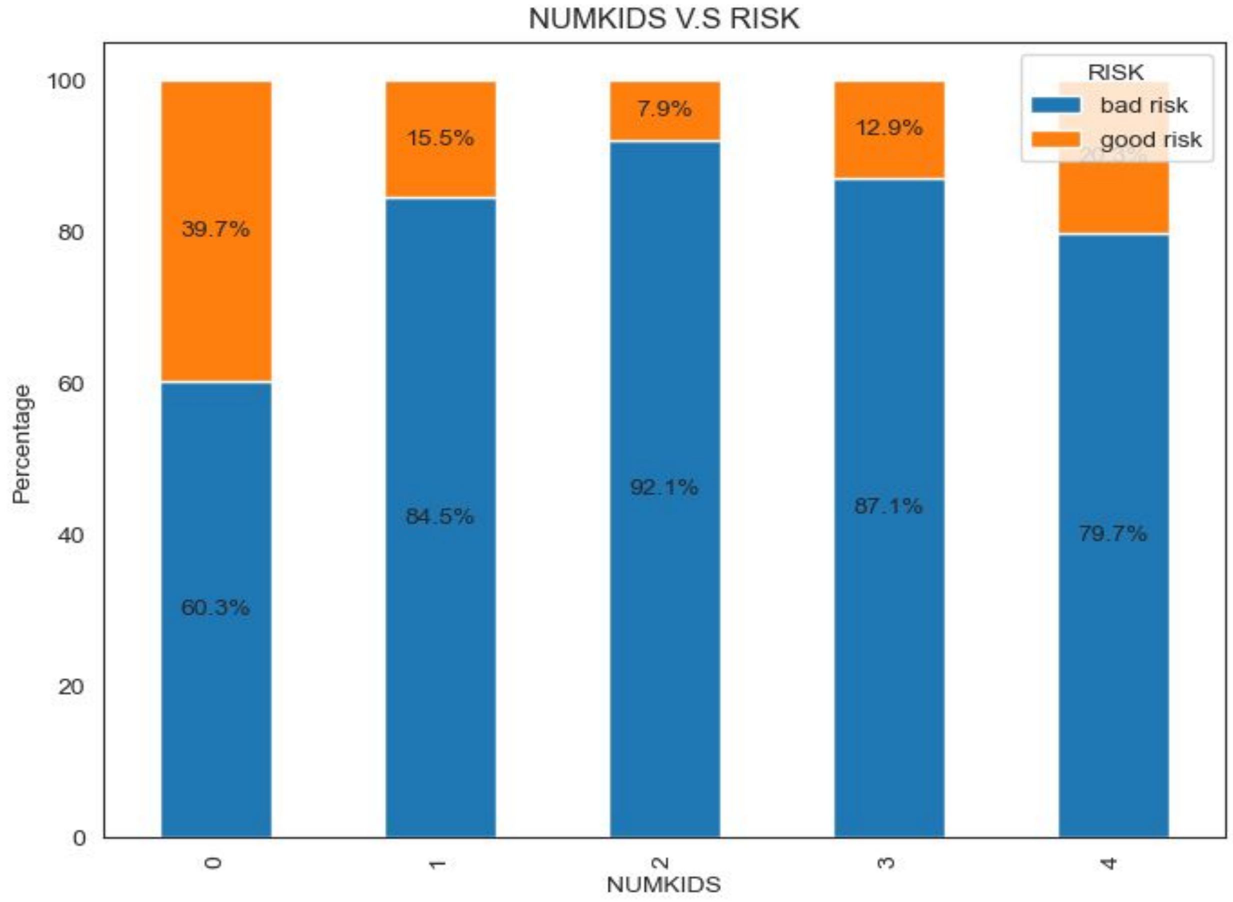
Good risk:

25000~30000

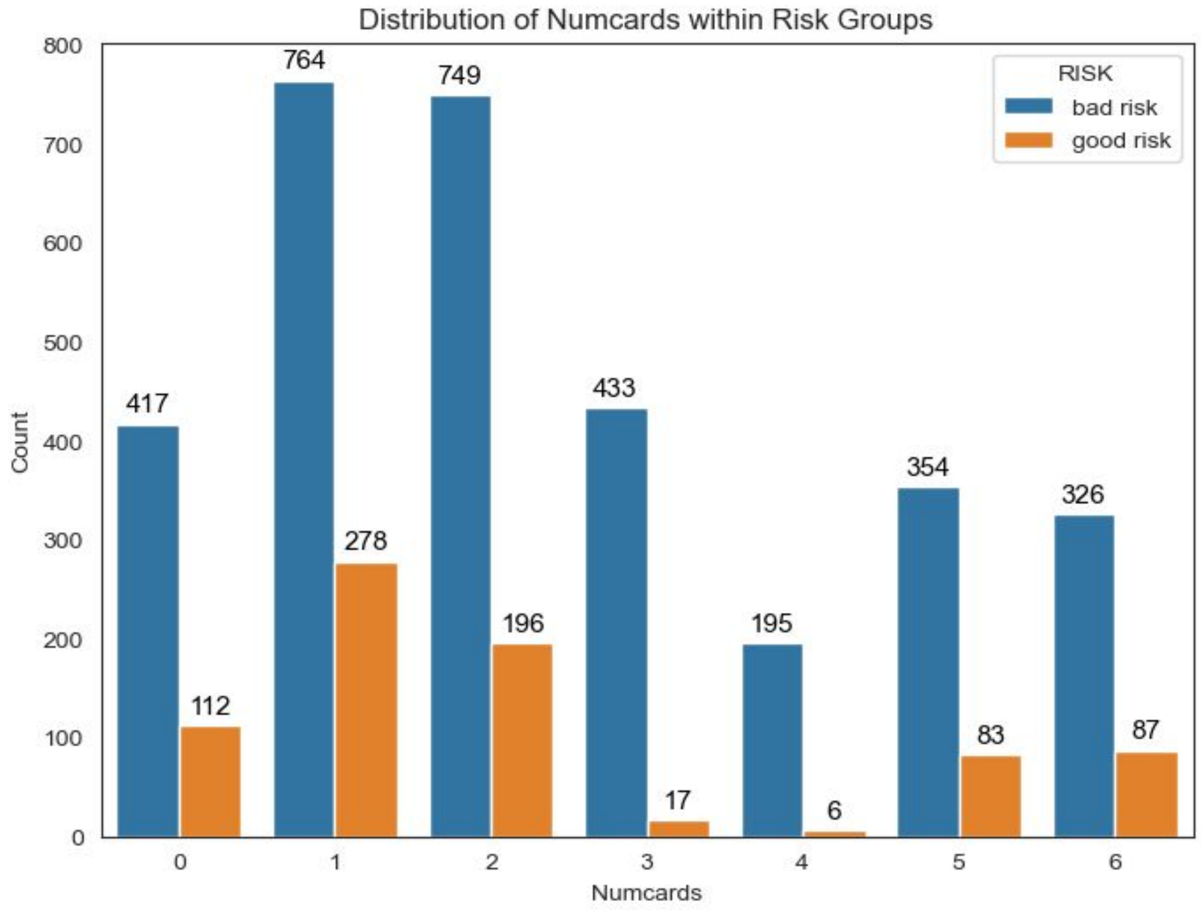
or 40000~45000

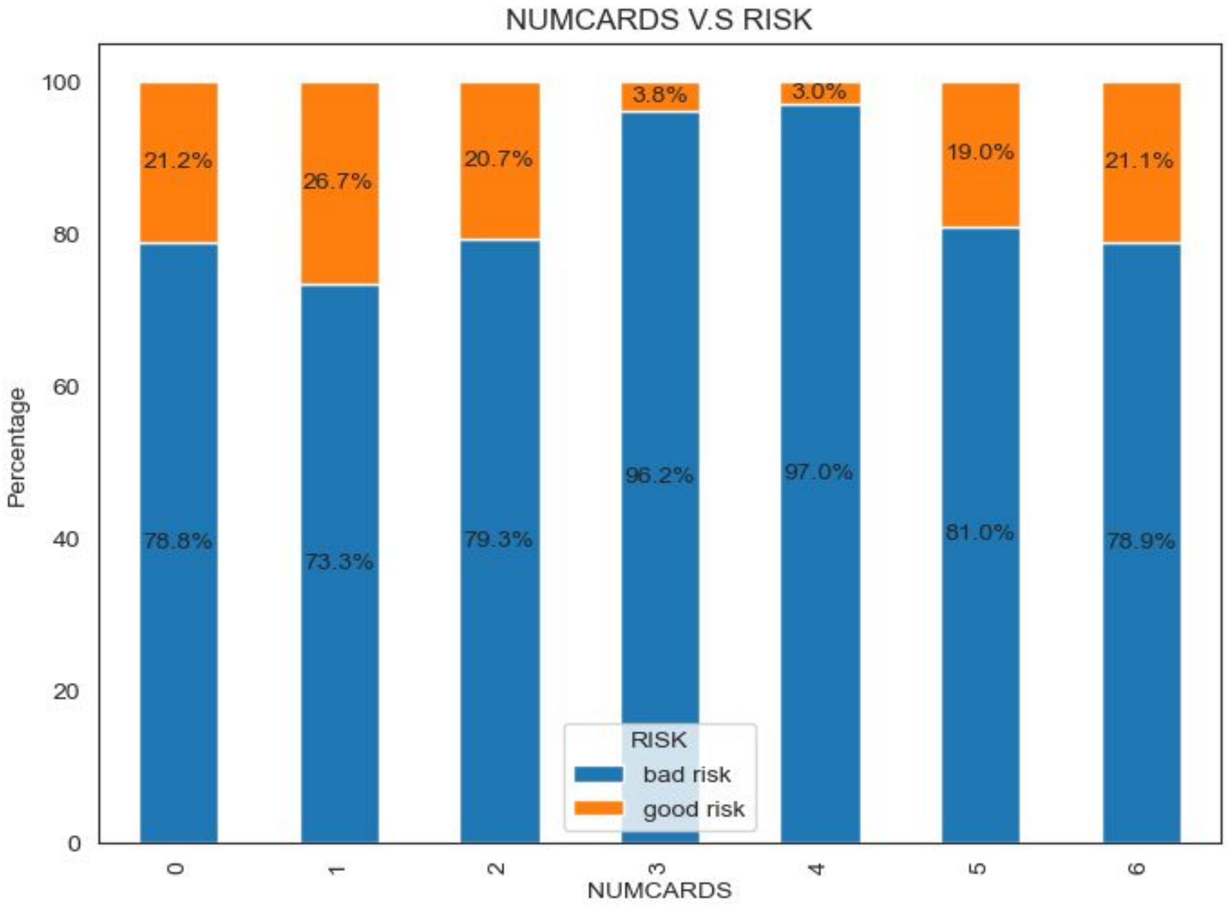
NUMBER OF KIDS



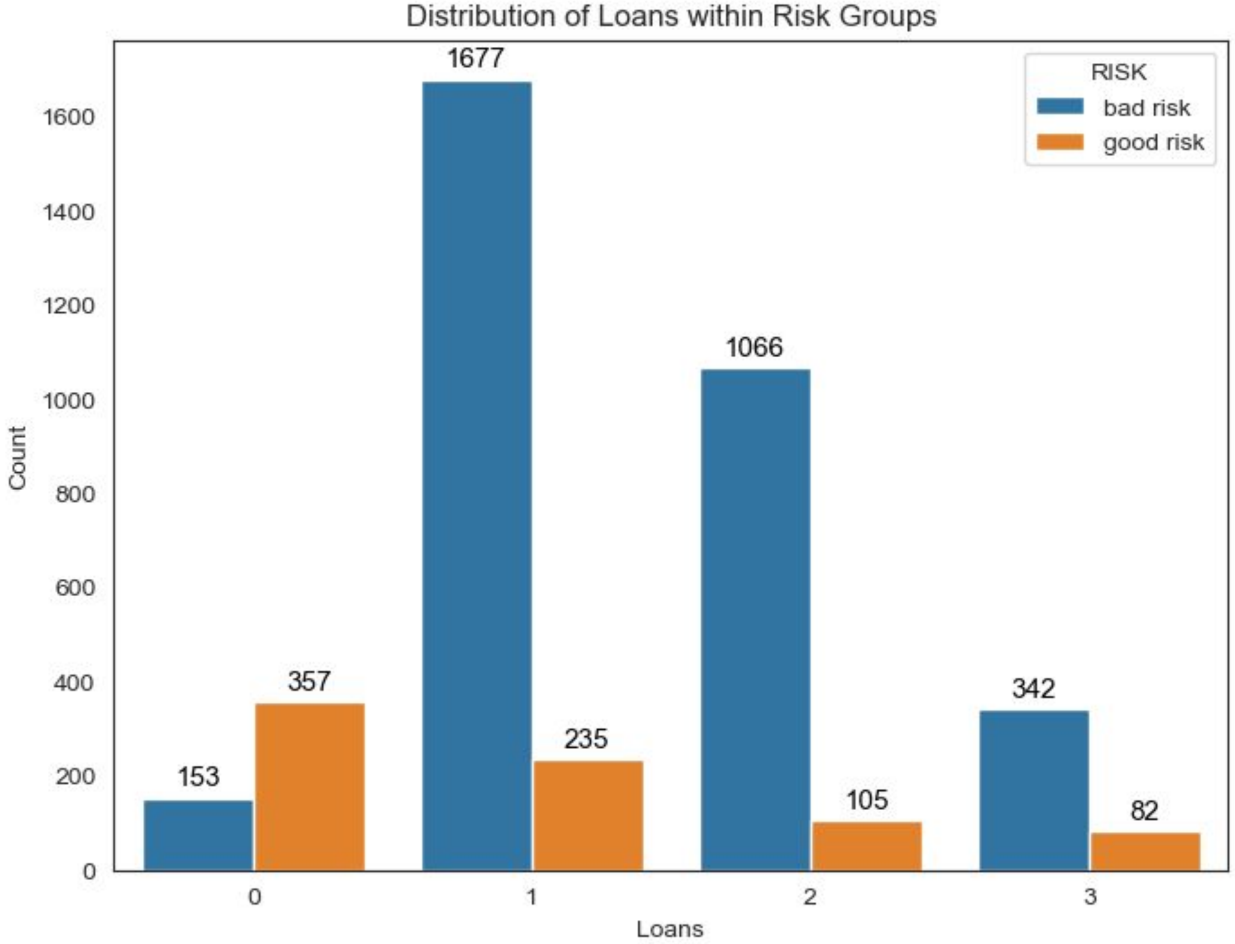


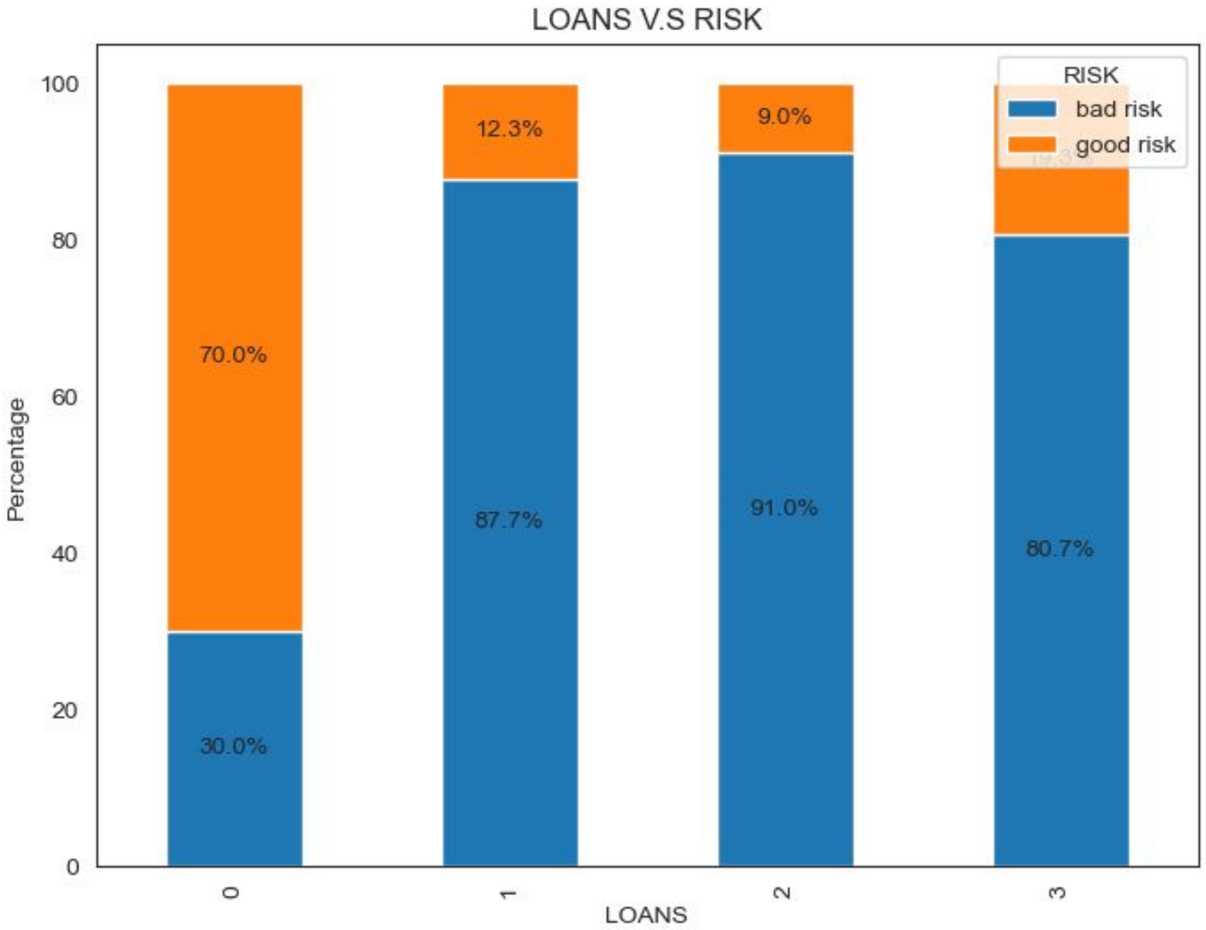
NUMBER OF CARDS



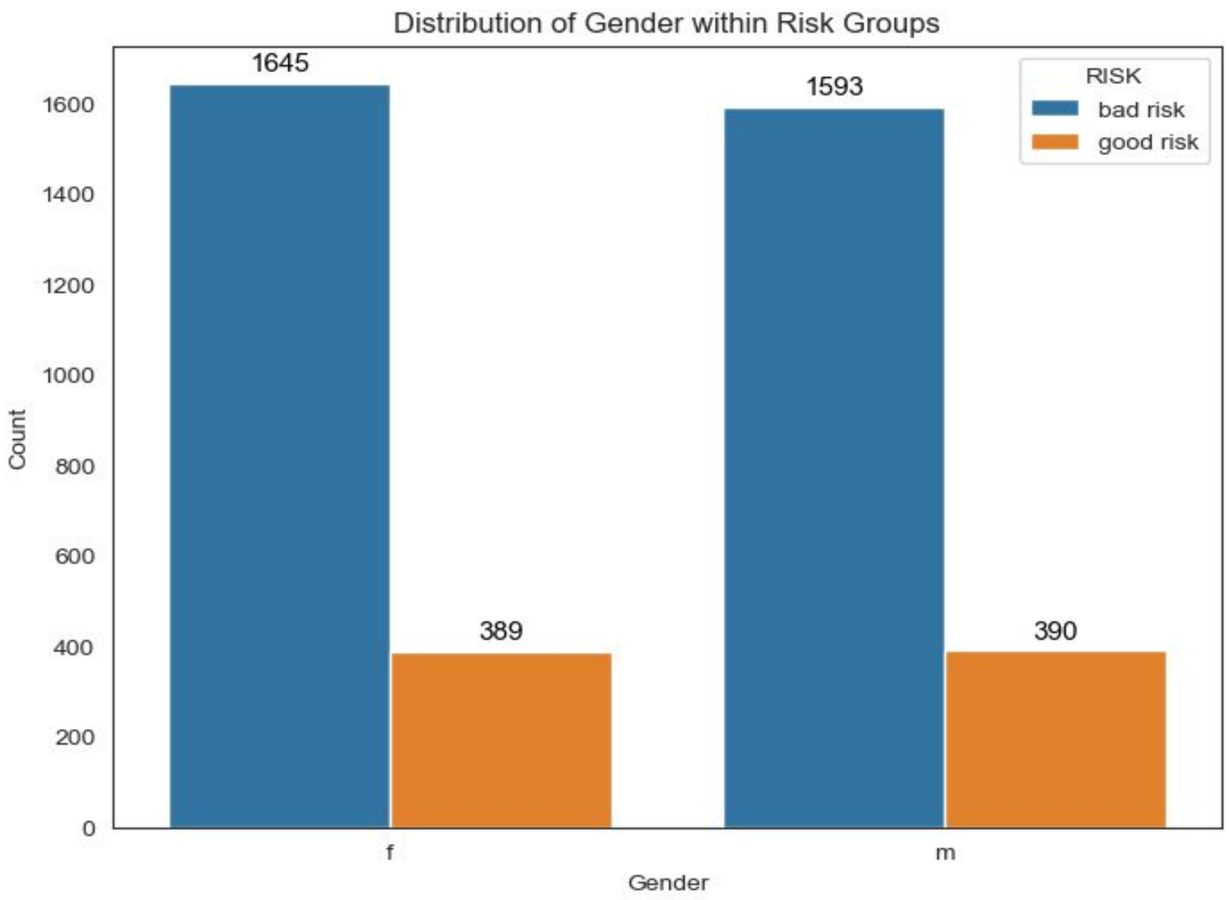


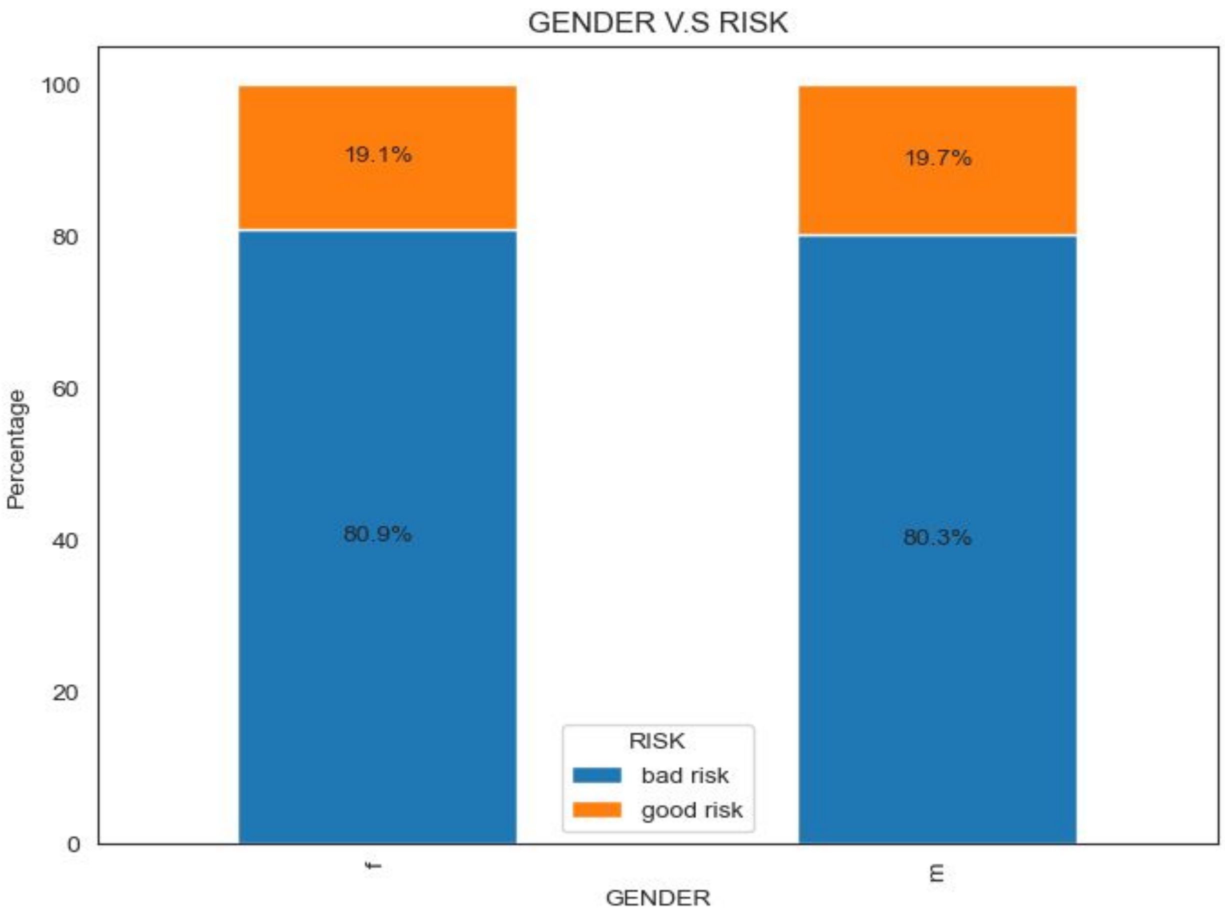
NUMBER OF EXISTING LOANS



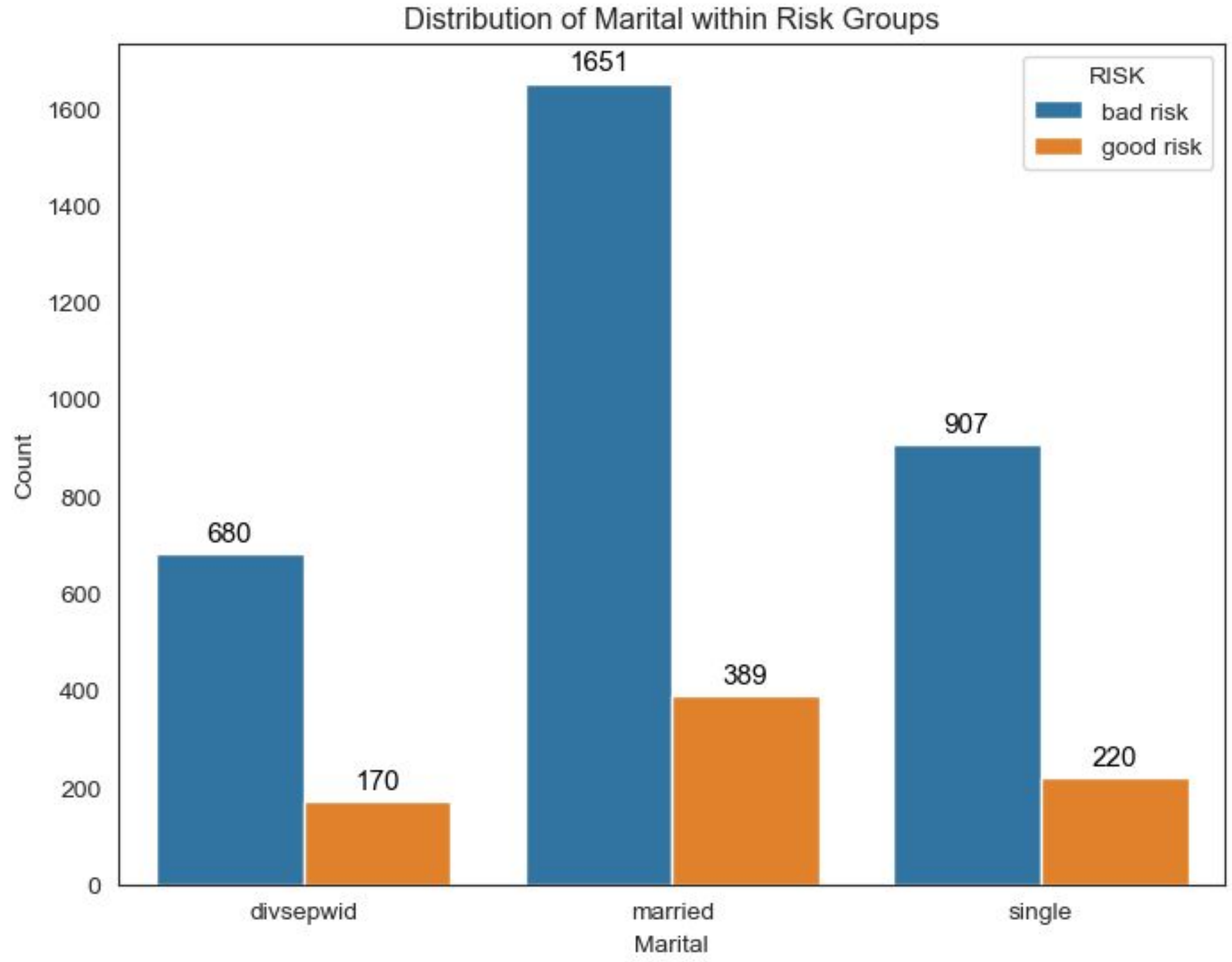


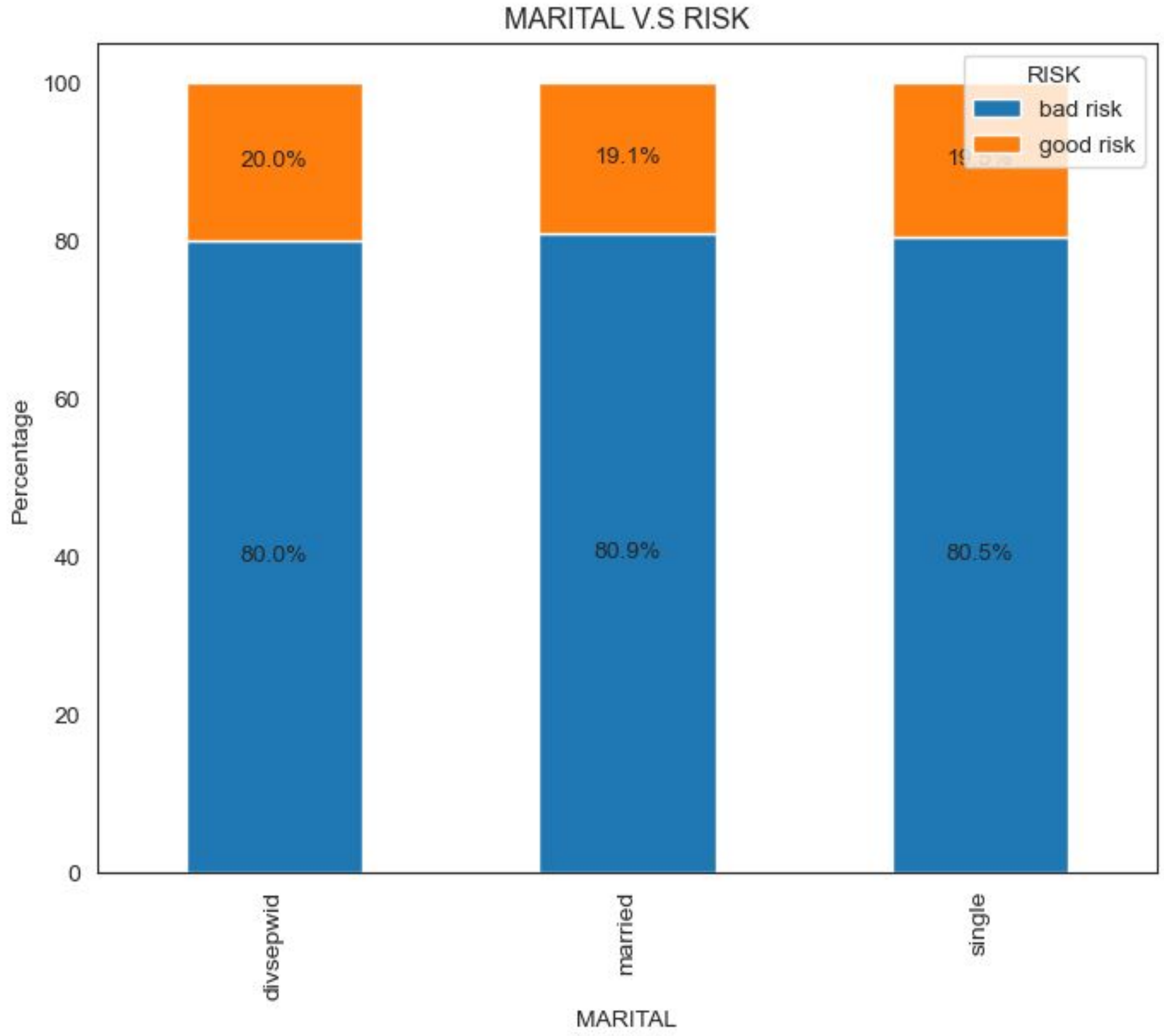
GENDER


Distribution of Gender within Risk Groups
RISK
bad risk
good risk
1645
389
1593
390
Count
0
200
400
600
800
1000
1200
1400
1600
f
m
Gender

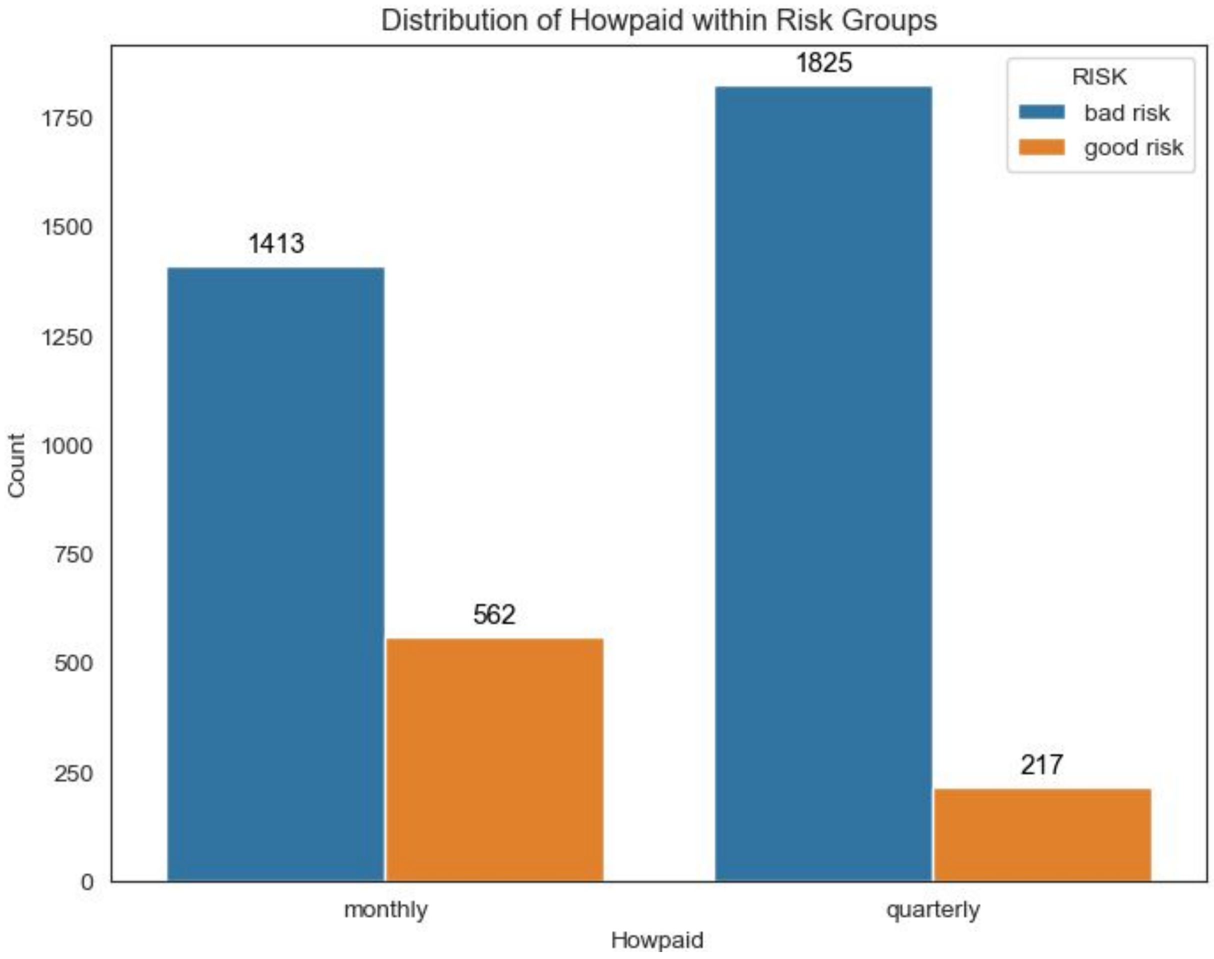

GENDER V.S RISK
19.1%
80.9%
19.7%
80.3%
RISK
bad risk
good risk
Percentage
0
20
40
60
80
100
f
m
GENDER

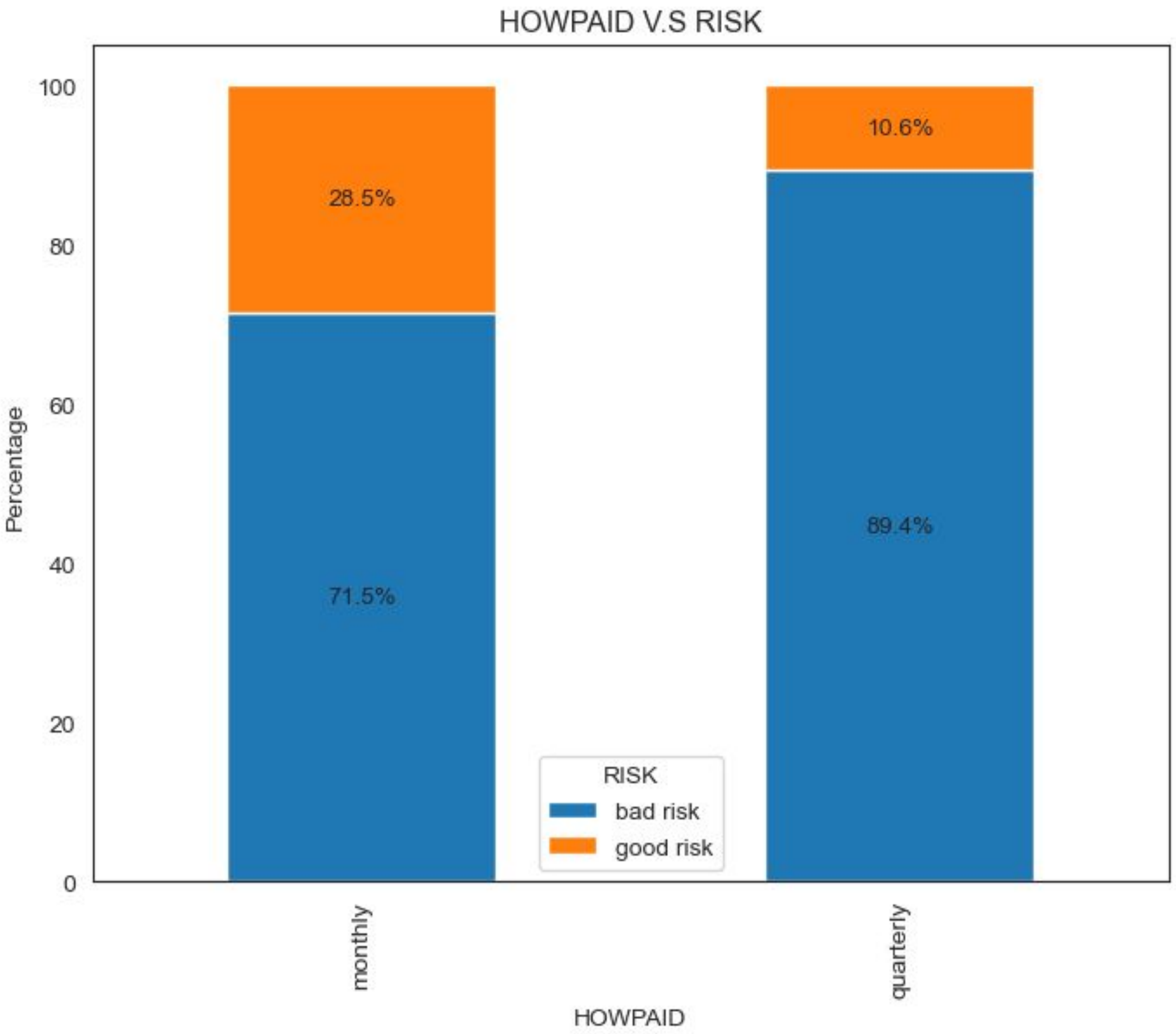
MARITAL STATUS



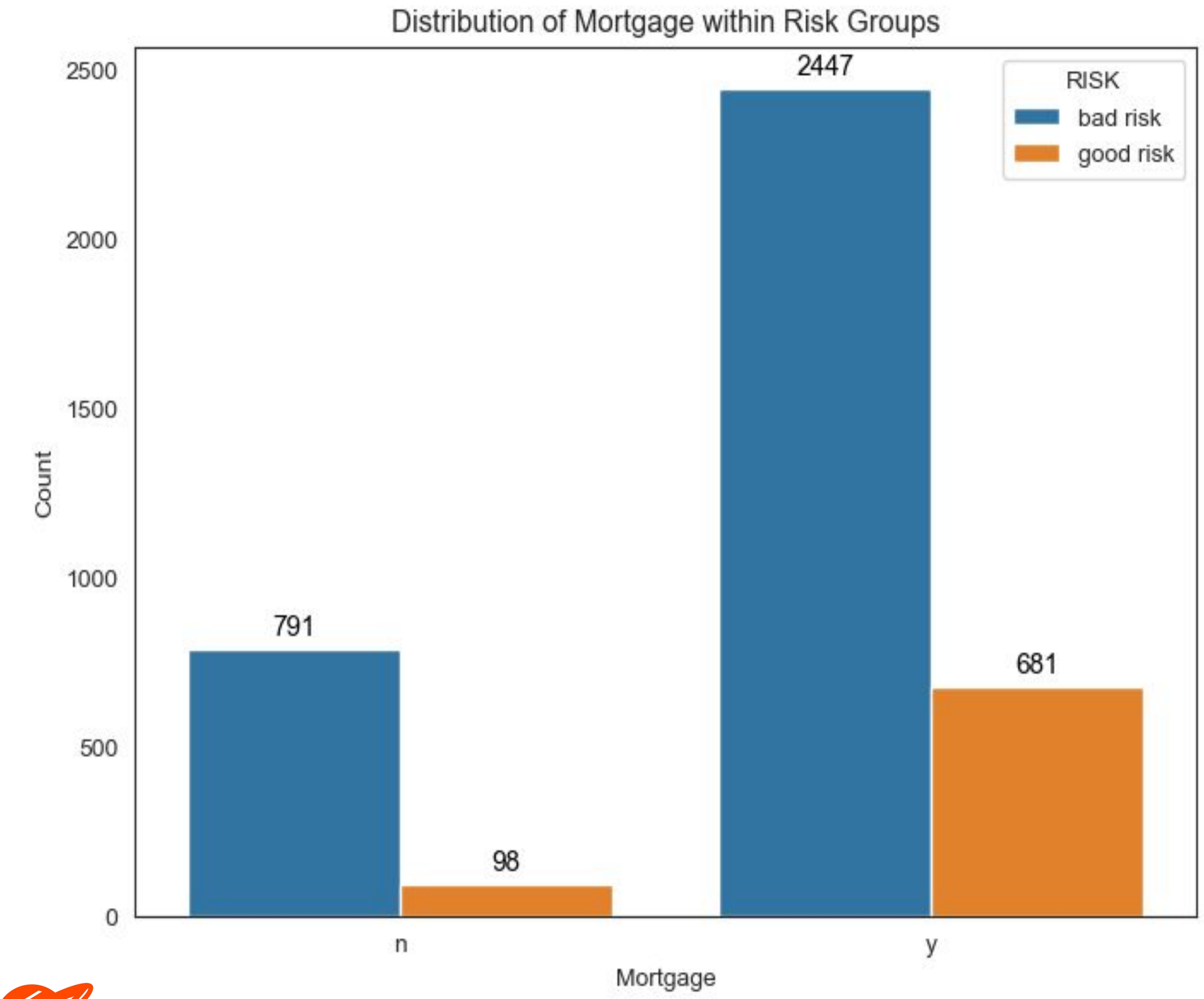


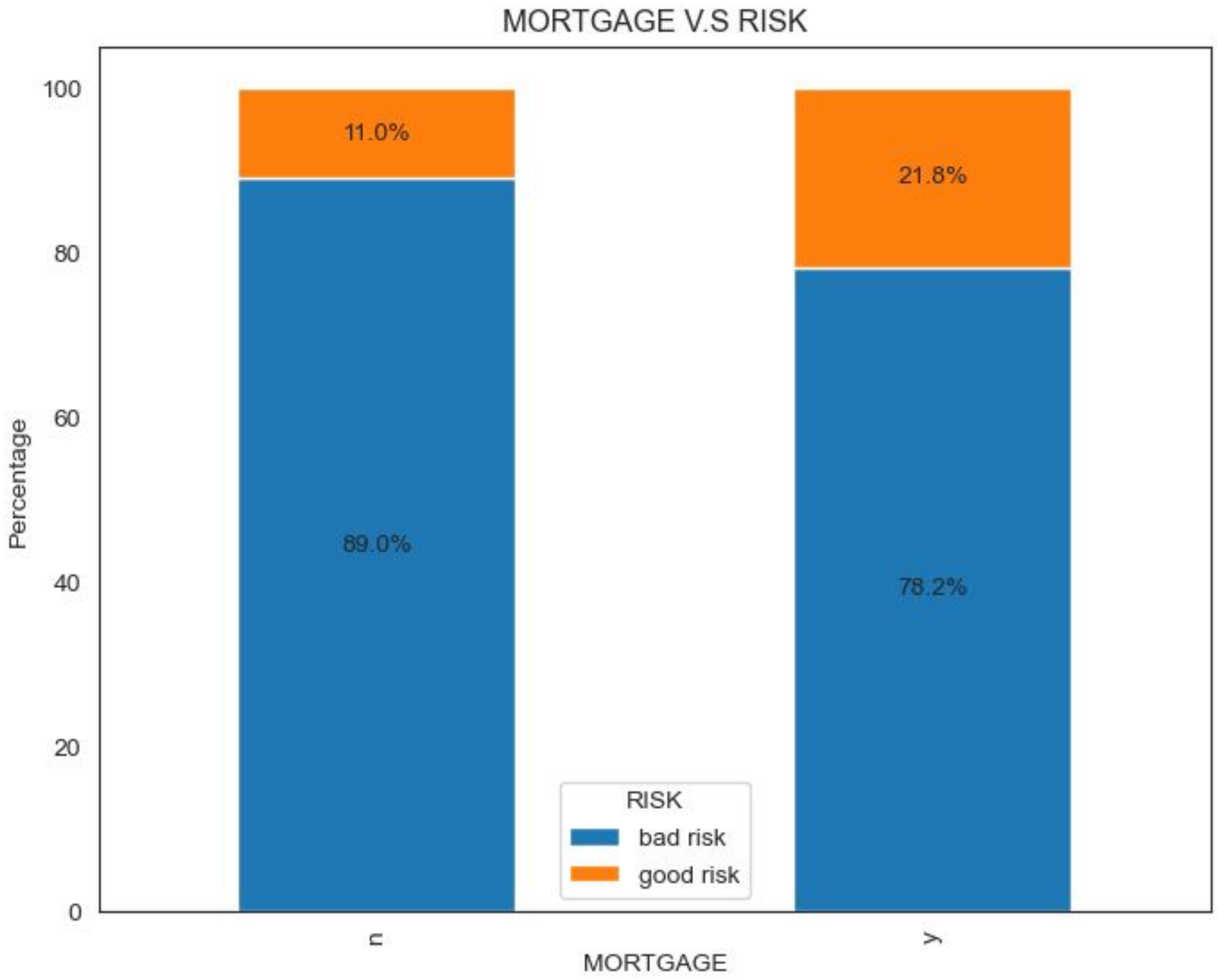
HOW PAID


Distribution of Howpaid within Risk Groups
Count
1750
1500
1250
1000
750
500
250
0
1413
562
1825
217
RISK
bad risk
good risk
monthly
quarterly
Howpaid

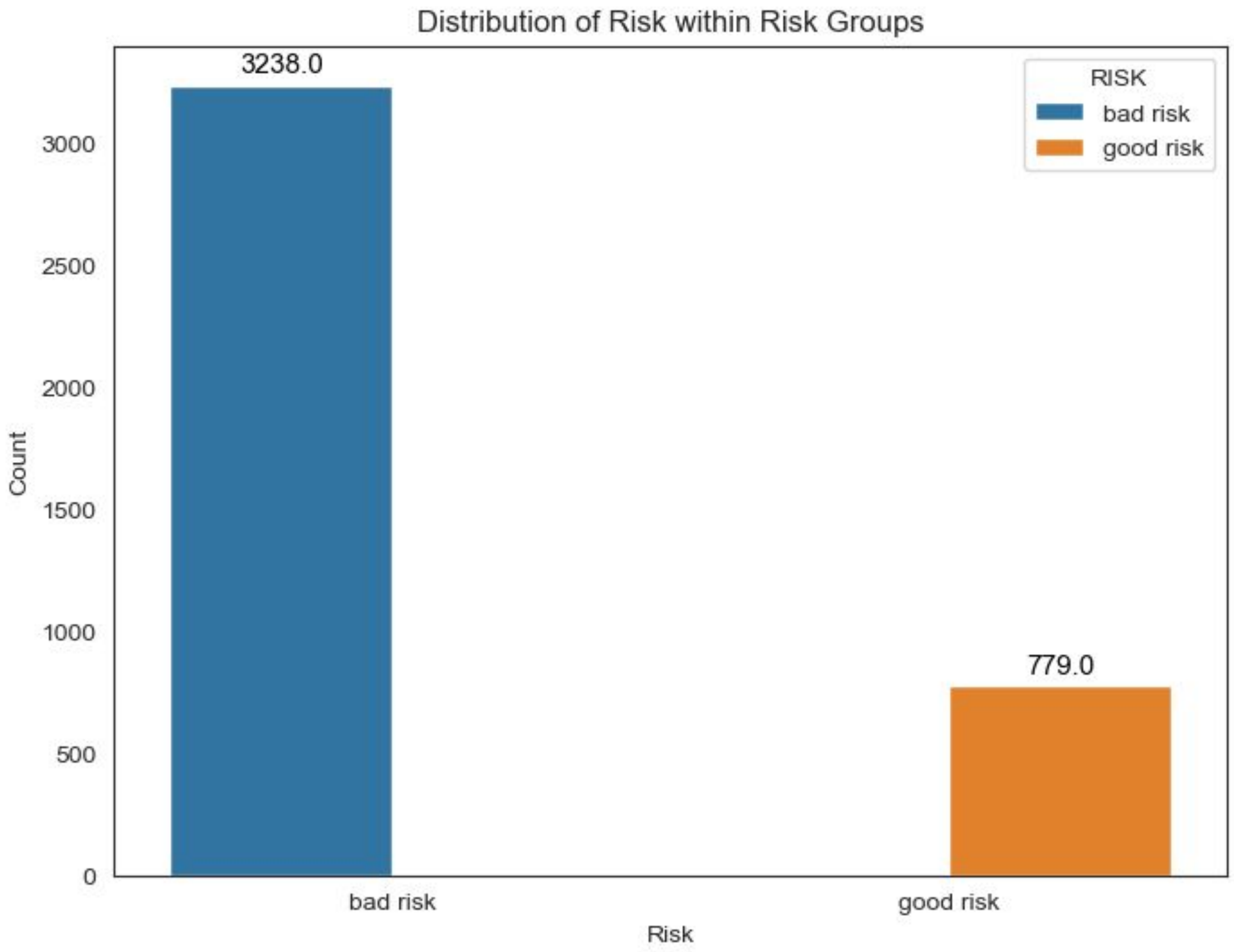

HOWPAID V.S RISK
Percentage
100
80
60
40
20
0
28.5%
71.5%
10.6%
89.4%
RISK
bad risk
good risk
monthly
quarterly
HOWPAID

MORTGAGE


Distribution of Mortgage within Risk Groups
RISK
bad risk
good risk
2447
791
681
98
Count
2500
2000
1500
1000
500
0
n
y
Mortgage


MORTGAGE V.S RISK
11.0%
21.8%
89.0%
78.2%
RISK
bad risk
good risk
Percentage
100
80
60
40
20
0
n
y
MORTGAGE

OUTCOME VARIABLE



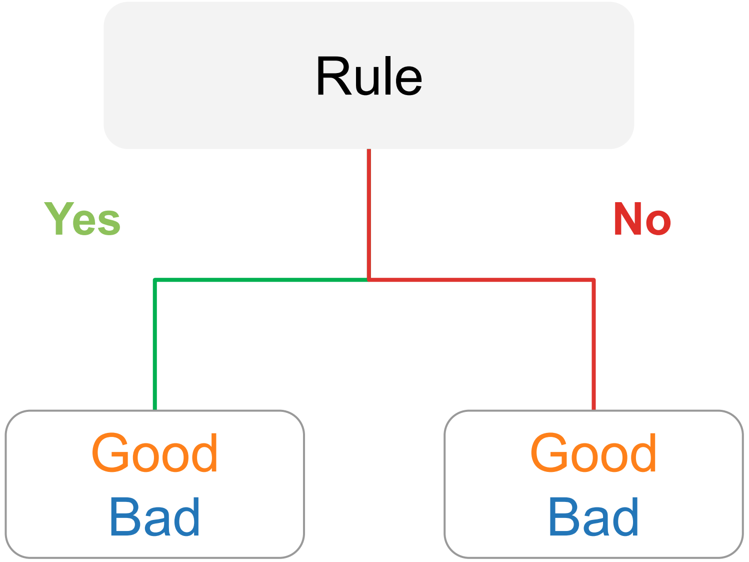
FIRST METHOD

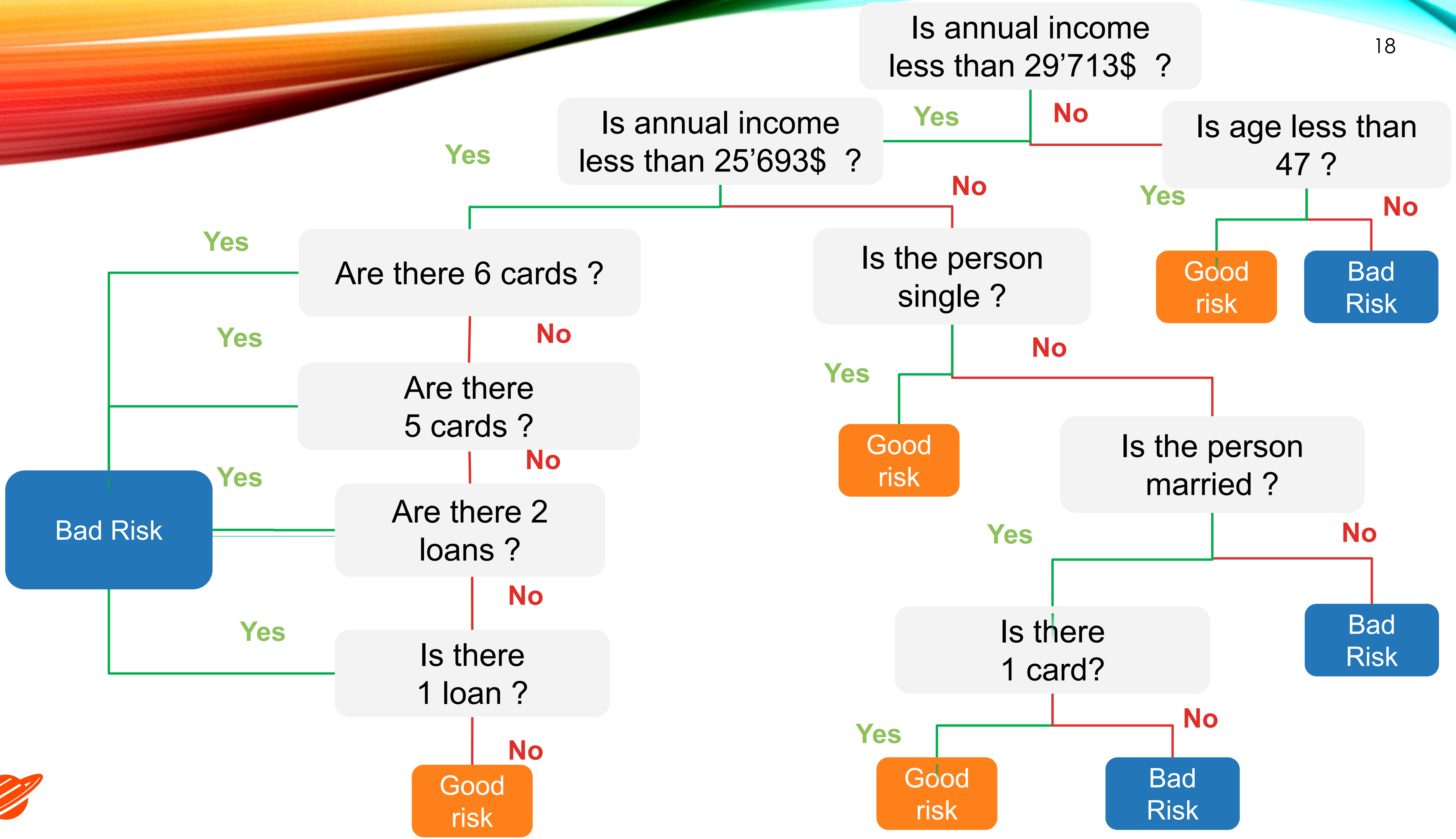
METHOD 1: DECISION TREE

Make predictions based on how a previous set of questions were answered.

METHOD 1: DECISION TREE

Example:





Is annual income less than 29'713$?
Yes
No
Is annual income less than 25'693$?
Is age less than 47 ?
Yes
No
Yes
No
Good risk
Bad Risk
Yes
Are there 6 cards ?
Is the person single ?
No
Yes
Yes
No
Are there 5 cards ?
Good risk
Is the person married ?
No
Yes
Bad Risk
Are there 2 loans ?
Yes
No
No
Yes
Is there 1 card?
Bad Risk
Is there 1 loan ?
Yes
No
No
Good risk
Good risk
Bad Risk

METHOD 1: DECISION TREE

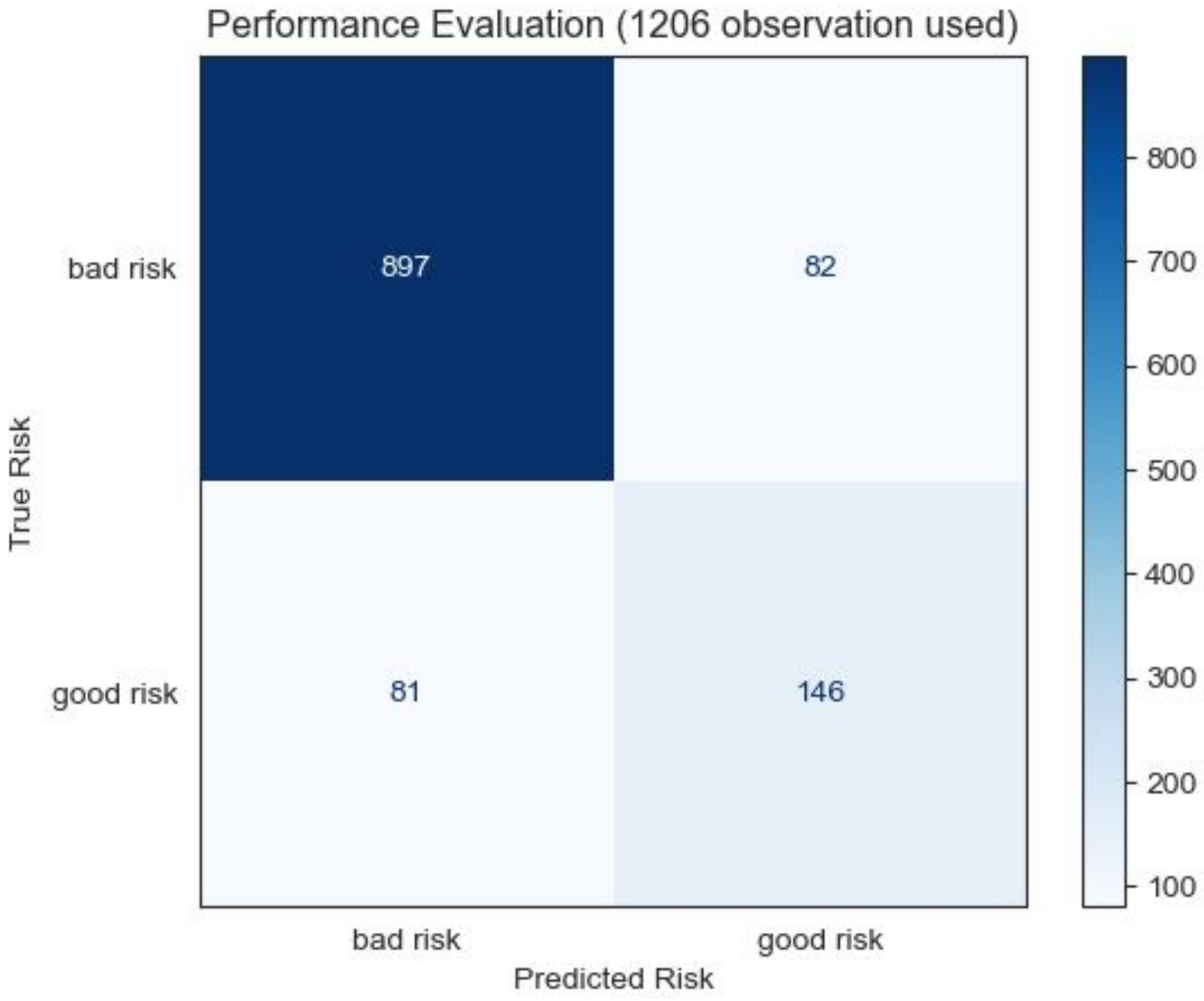
Performance evaluation

For ***bad risk*** category:

- high accuracy
- low misjudgement

For ***good risk*** category:

- high accuracy
- some misclassification



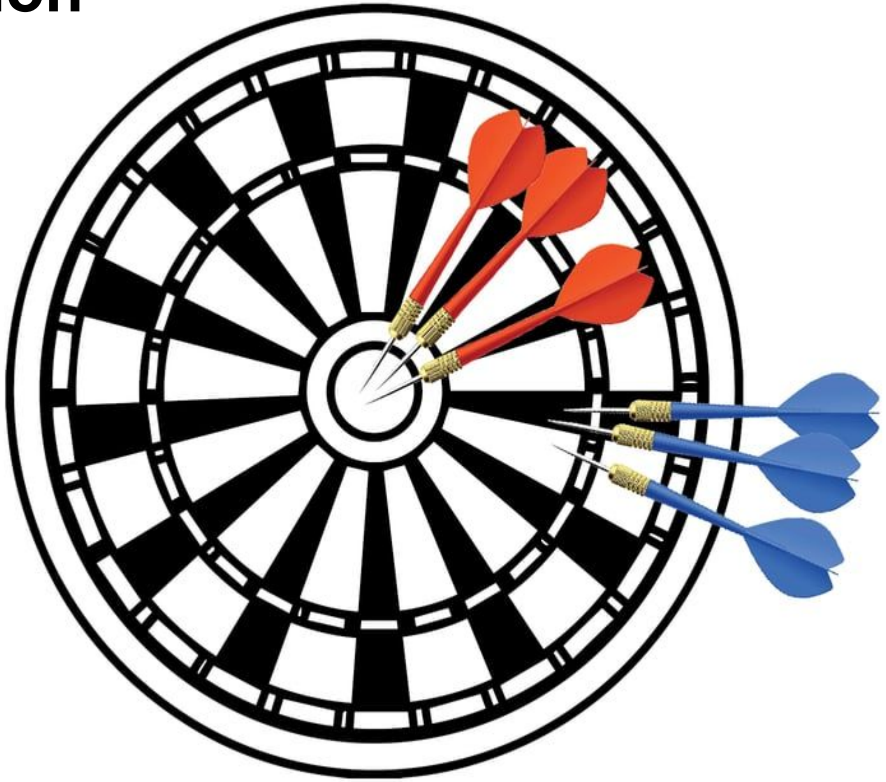
METHOD 1: DECISION TREE

Accuracy vs. Precision

Assume the objective is to hit the bullseye.

The **blue** dart is **precise**, but not **accurate**.

The **red** dart is both **precise** and **accurate**



Overall Accuracy: 86%

Precision:

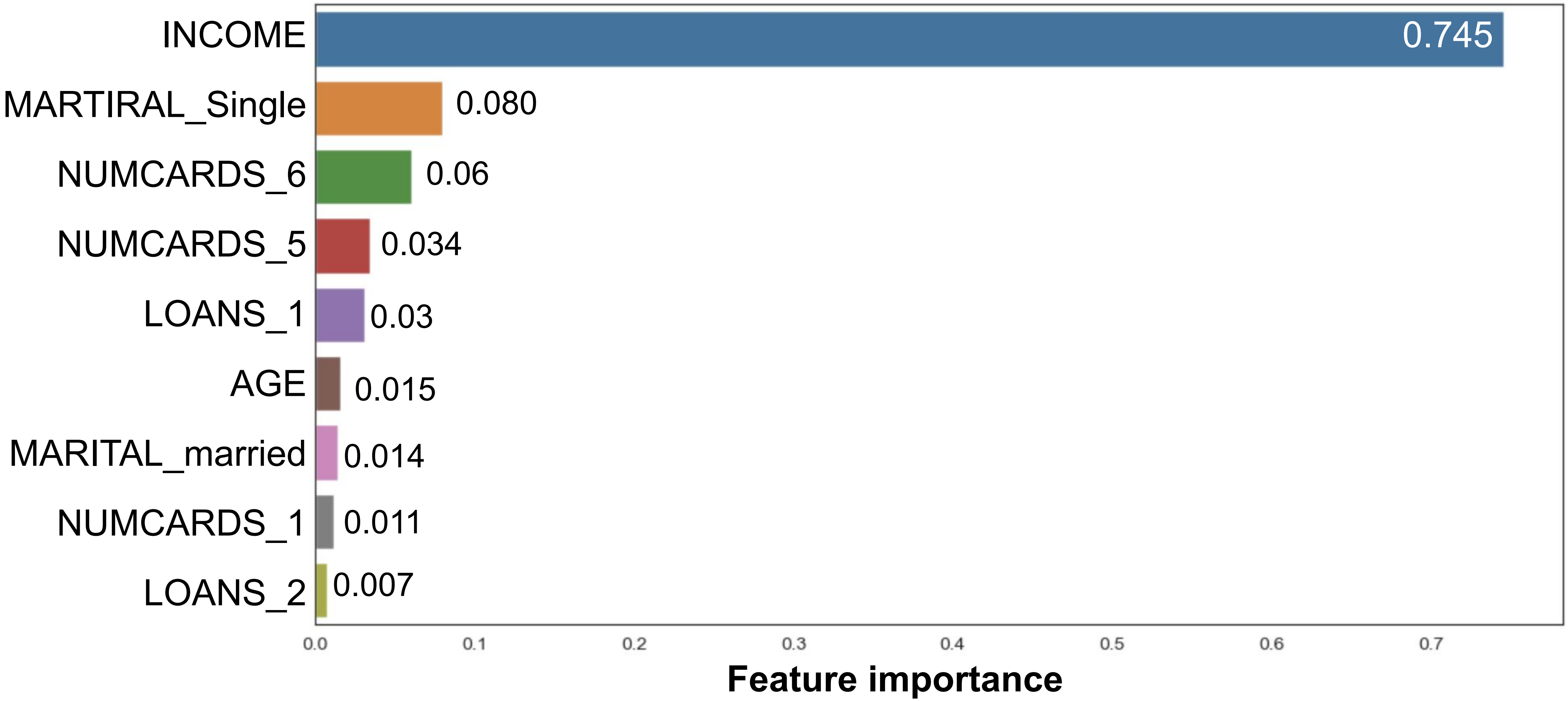
Bad risk: 92%

Good risk: 64%

METHOD 1: DECISION TREE

Due to a class imbalance in the data
there are more *bad risk* 3238 than *good risk* 779

METHOD 1: DECISION TREE



Is there room for improvement ?

MODEL IMPROVEMENT

METHOD 2: BOOSTED DECISION TREE

Creating series of trees, where each tree is built to improve the mistakes of the previous tree.

METHOD 2: BOOSTED DECISION TREE

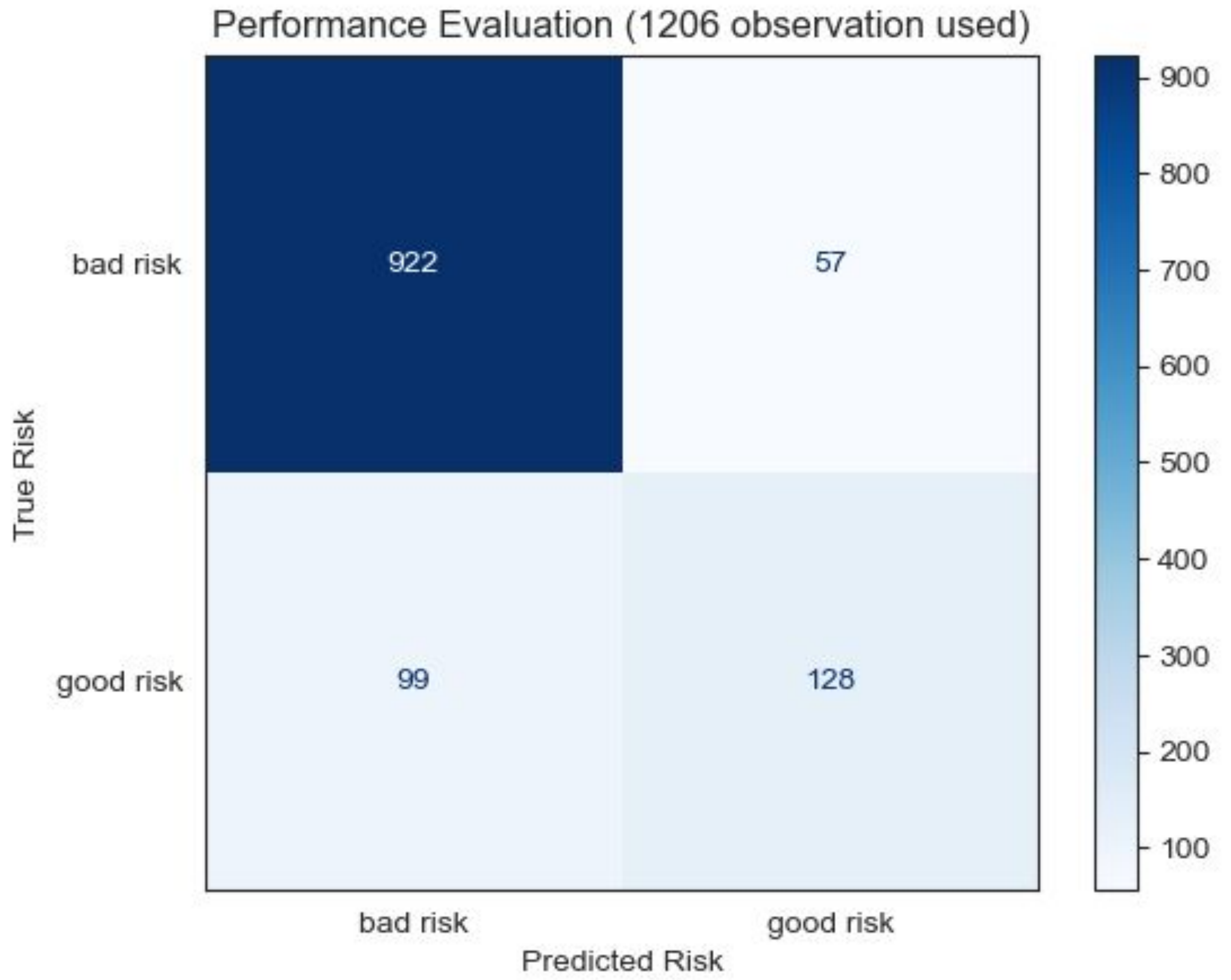
Performance evaluation

For **bad risk** category:

- Small decrease in precision
- Still very good performances

For **good risk** category:

- Improved performances



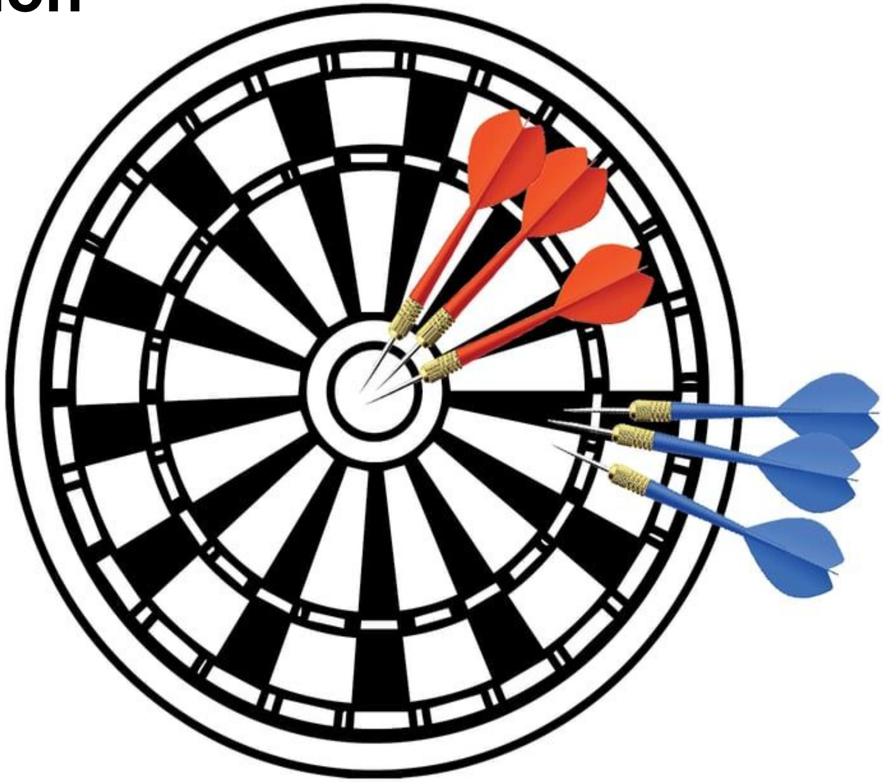
METHOD 2: BOOSTED DECISION TREE

Accuracy vs. Precision

Assume the objective is to hit the bullseye.

The **blue** dart is **precise**, but not **accurate**.

The **red** dart is both **precise** and **accurate**



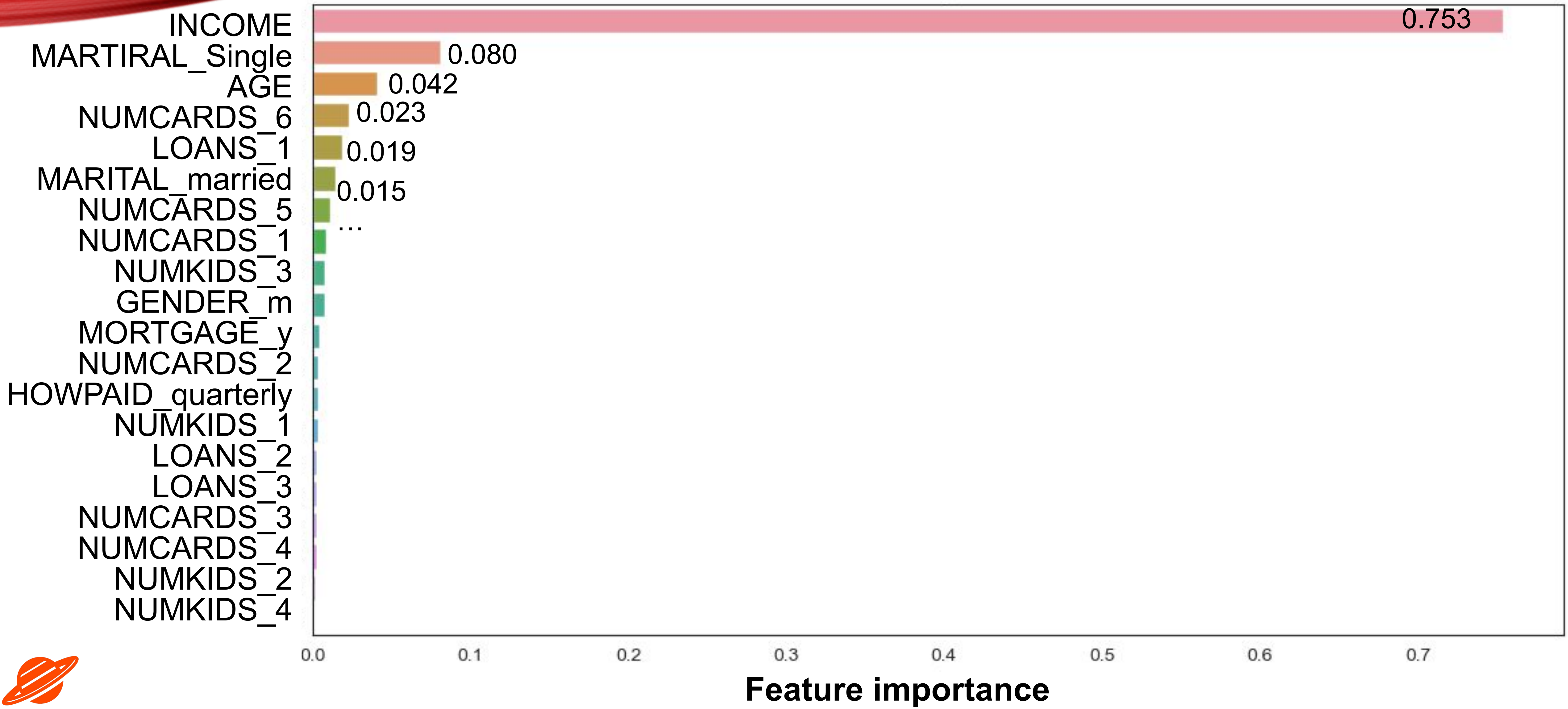
Overall Accuracy: 87%

Precision:

Bad risk: 90%

Good risk: 69%

METHOD 2: BOOSTED DECISION TREE



RESULTS

Which model should we use ?

RESULTS

- Decision Tree Model if we are interested in predicting the bad risks.

- Boosting Model if we interested in predicting the good risks **(with reservations)**

RESULTS

Effectiveness of the credit rating detection:

- Performed **well** in the **bad risk** category, with **high precision**.
- Performed **poorly** in the **good risk** category, with relatively **low precision**.

Decision rules:

- It should include variables related to: *Income*; *Age*; *Marital status*; *Loans*; *Cards*

LIMITATIONS

- The sample of customers in the *good risk* category was small which lead to bias the model towards predicting the majority class (*bad risk*).
- Sample collection method
- Is the sample representative of the true population (the bank)?
- The feature selections
- Economic conditions

SUGGESTIONS

- Increase the sample data
- Introduce more features
- Consult our team for your next data collection

CONCLUSION

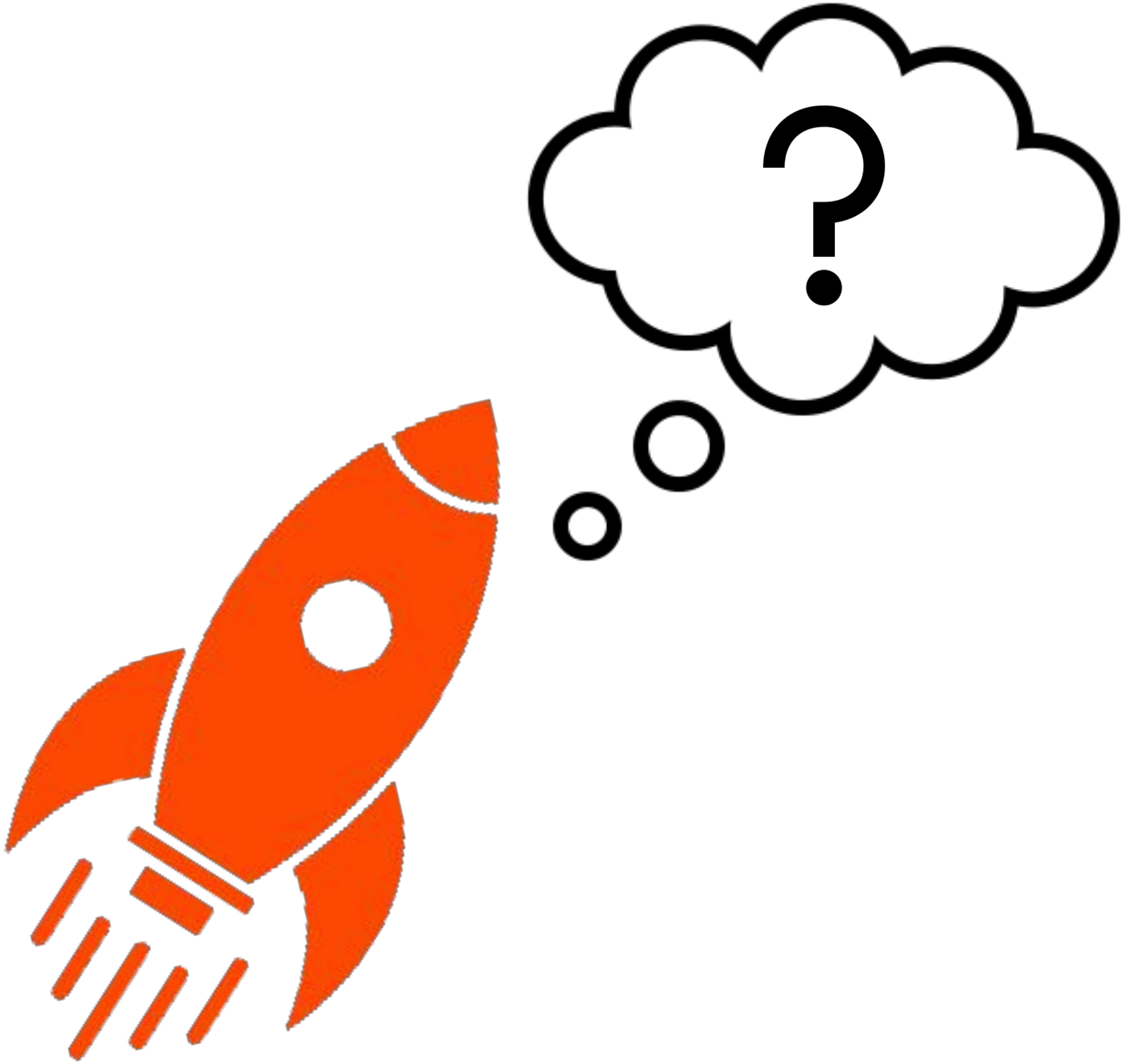
Effectiveness of the credit rating detection:

We focused on bad risk customers and we are able to detect credit rating: precision 92%

Predicting credit rating:

We rely essentially on:

I. Income
II. Number of cards
III. Marital status
IV. Loan
V. Age



Thank you for working with

SATURN'S RINGMASTERS

Consulting