Predicting **Delinquency** in Credit Card Payments

Argishti Ovsepyan
Masood Dastan
Gabriela Fichtner
Saamir Shamsie

Improving risk management

By analyzing personal information and historical credit data, the project will estimate the likelihood of payment defaults and facilitate the implementation of effective risk control measures.

Addressing the demand of credit assessment methods that do not require traditional credit checks has potential savings for banks and allows individuals to avoid a credit inquiry on their record.

Agenda

01
Data Cleaning
Process aimed at enhancing data quality

02
EDA
Features and their relationship with delinquency

03
Modeling
Creating and tuning machine learning algorithms

04
Evaluation
Assessing the models performance

05
Limitations
Constraints in the study that can impact our models

01

Data Cleaning

Process aimed at enhancing data quality

Data Source: Kaggle

Predictive Features

- Education Level
- Annual Income
- Occupation
- Days of Employment
- Days from Birth
- Family Status
- Housing Type

(Potential) Target Variables

- Current Delinquency
- 3 Months Delinquency
- 6 Months Delinquency
- 12 Months Delinquency

Enhancing data **quality**

Dropping Features:

Gender, Days Employed, and Days Birth

Generating features:

Age and Years Employed as well as predictive variables

Handling missing values:

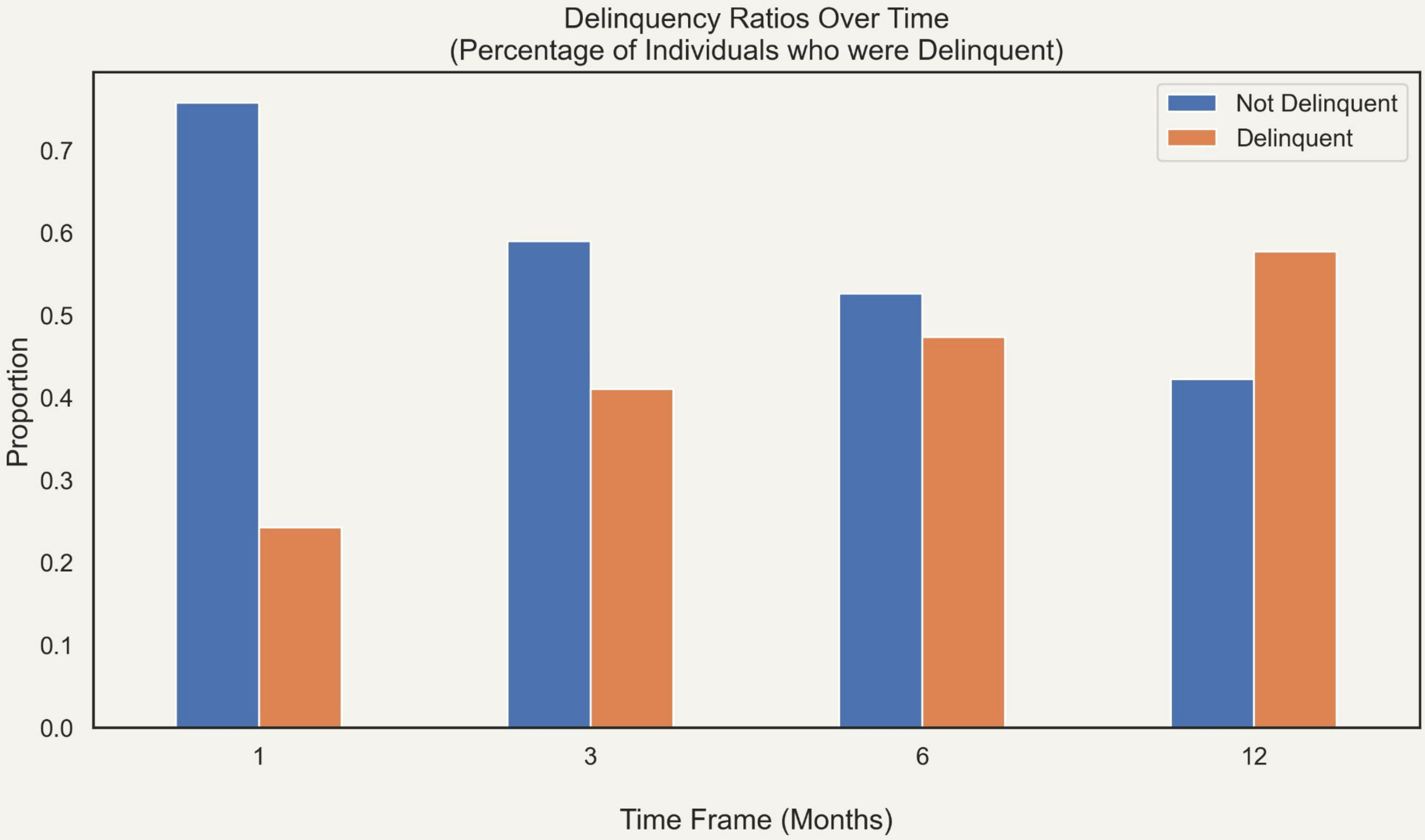
Identified and created a new label for retirees and a separate label 'missing' for the rest

Handling duplicates:

Merged dataset had 47 duplicate ID's with different values and were dropped

Custom functions:

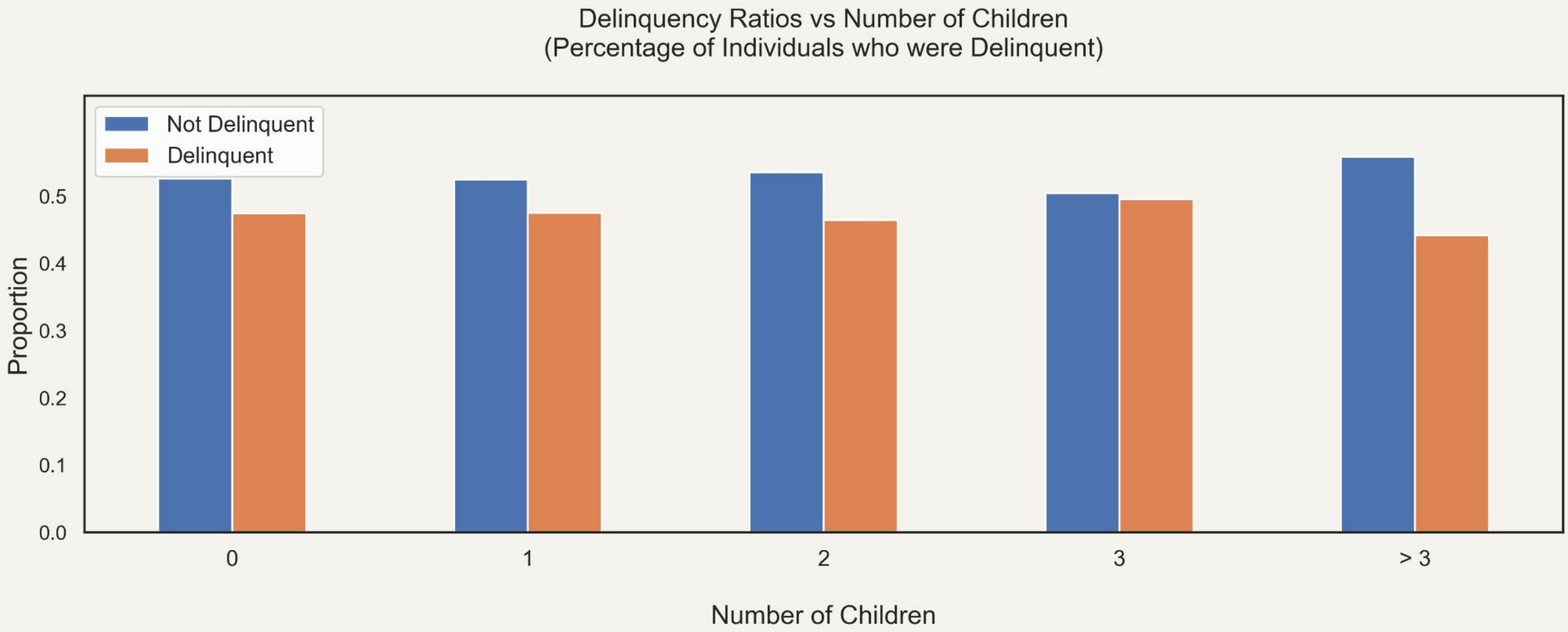
The function 'credit_approval_data_cleaner' was utilized to do all the cleaning on both the training and test data

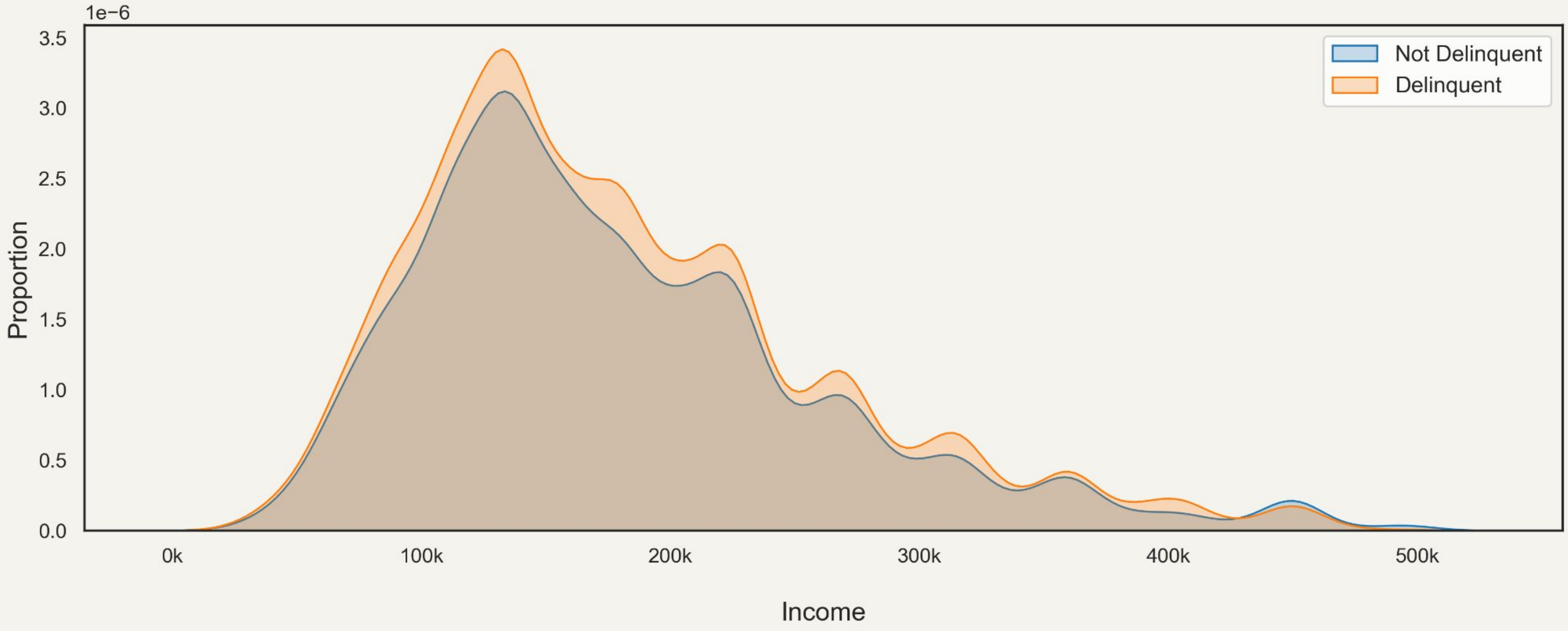

Delinquency Ratios Over Time
(Percentage of Individuals who were Delinquent)
Not Delinquent
Delinquent
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
3
6
12
Time Frame (Months)

02

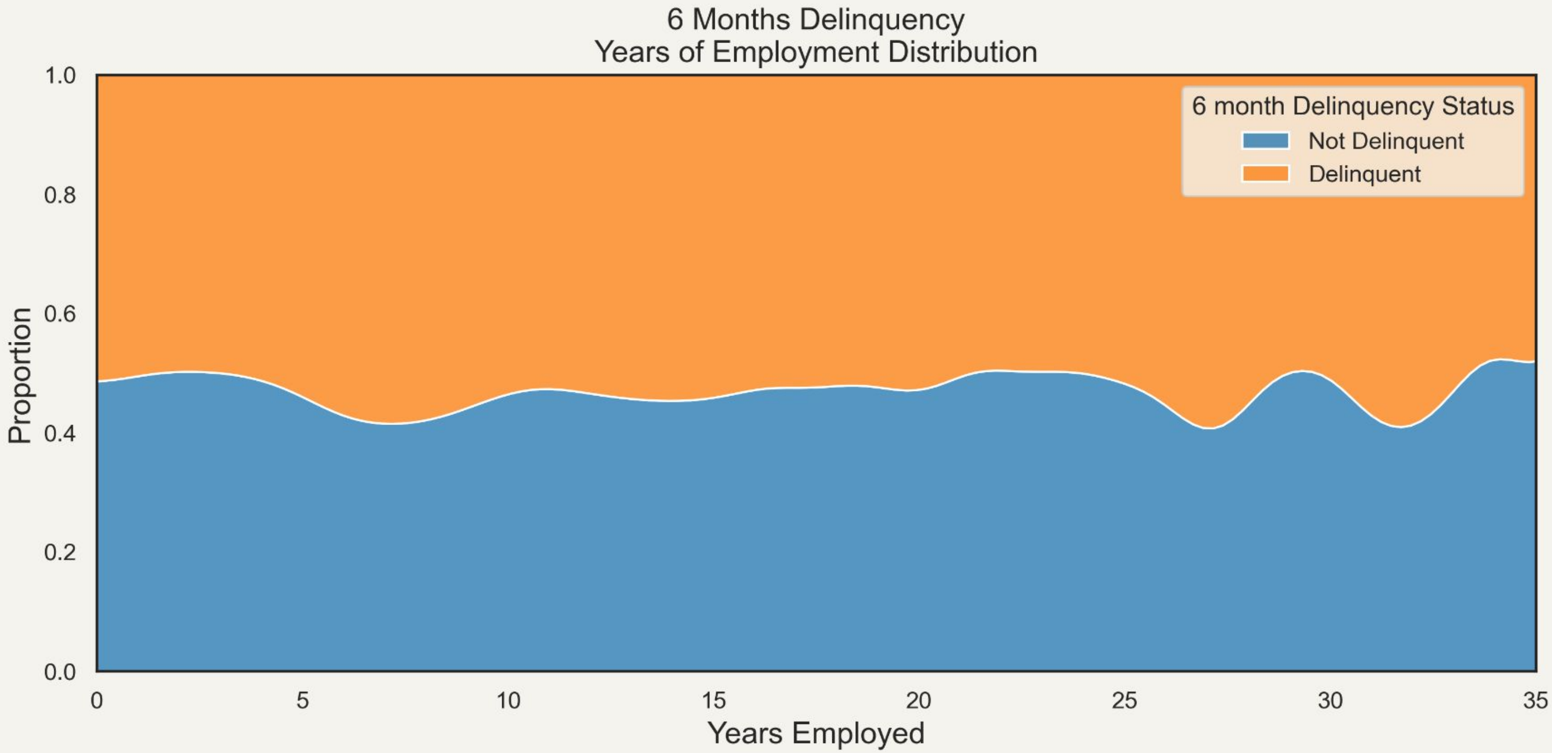
EDA

Features and their relationship with delinquency

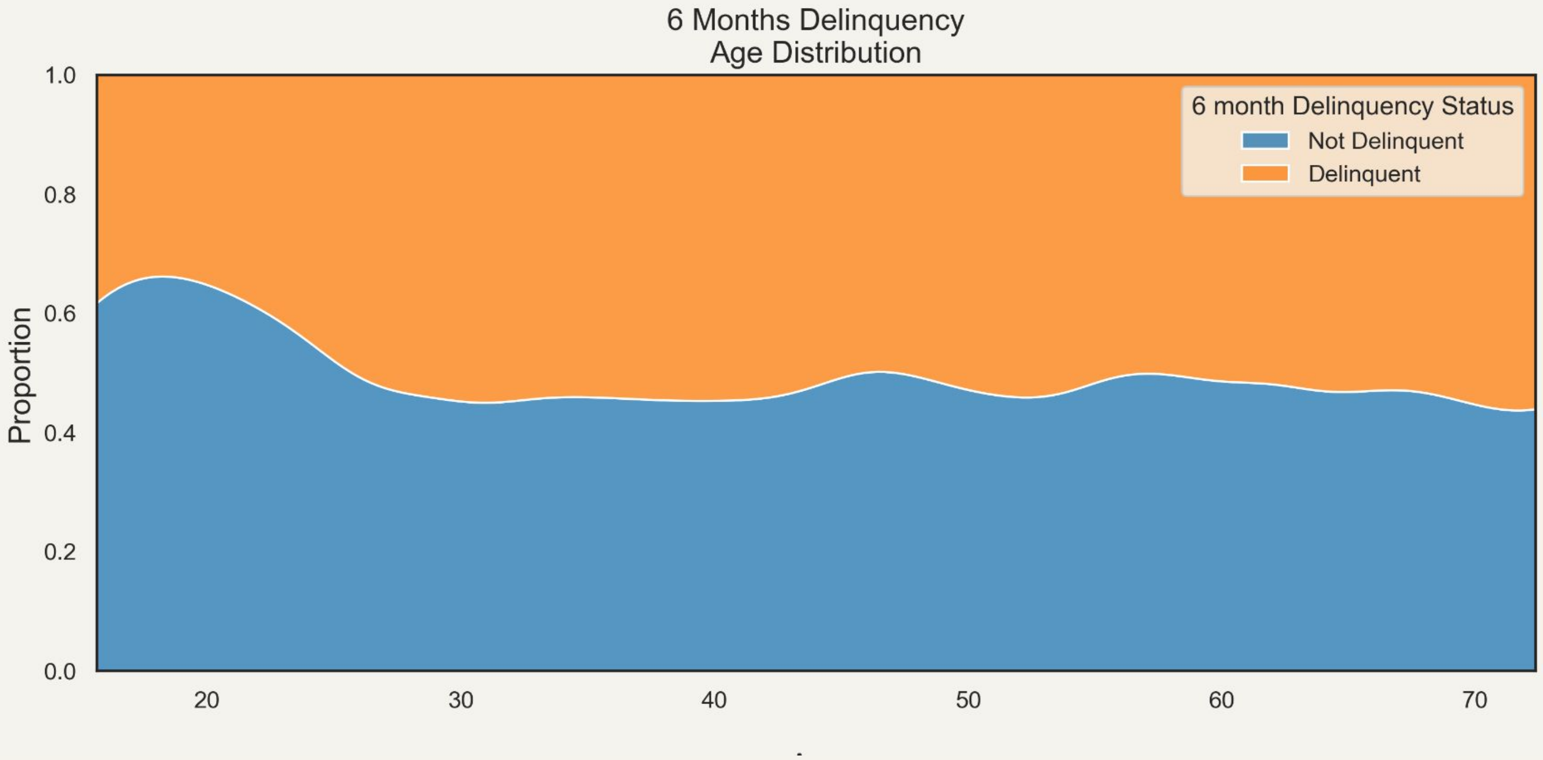

Delinquency Ratios vs Number of Children
(Percentage of Individuals who were Delinquent)
Not Delinquent
Delinquent
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
> 3
Number of Children



1e−6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Proportion
0k
100k
200k
300k
400k
500k
Income
Not Delinquent
Delinquent



6 Months Delinquency
Years of Employment Distribution
6 month Delinquency Status
Not Delinquent
Delinquent
Proportion
Years Employed
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35


6 Months Delinquency
Age Distribution
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
20
30
40
50
60
70
6 month Delinquency Status
Not Delinquent
Delinquent

03

Modeling

Creating and tuning machine learning algorithms

Train-validation-Test Split

We split the data three ways to make sure our predictions are generalizable.

Models

- DNN
- SVC
- Random Forest
- AdaBoost
- Gradient Boosting
- Logistic Regression

Scores on 6-month delinquency

	Accuracy
DNN	62%
Logistic Regression	75.5%
SVC	76.7%

	Accuracy
Gradient Boost	79.05%
Random Forest	79.06%
Ada Boost	79.3%

04

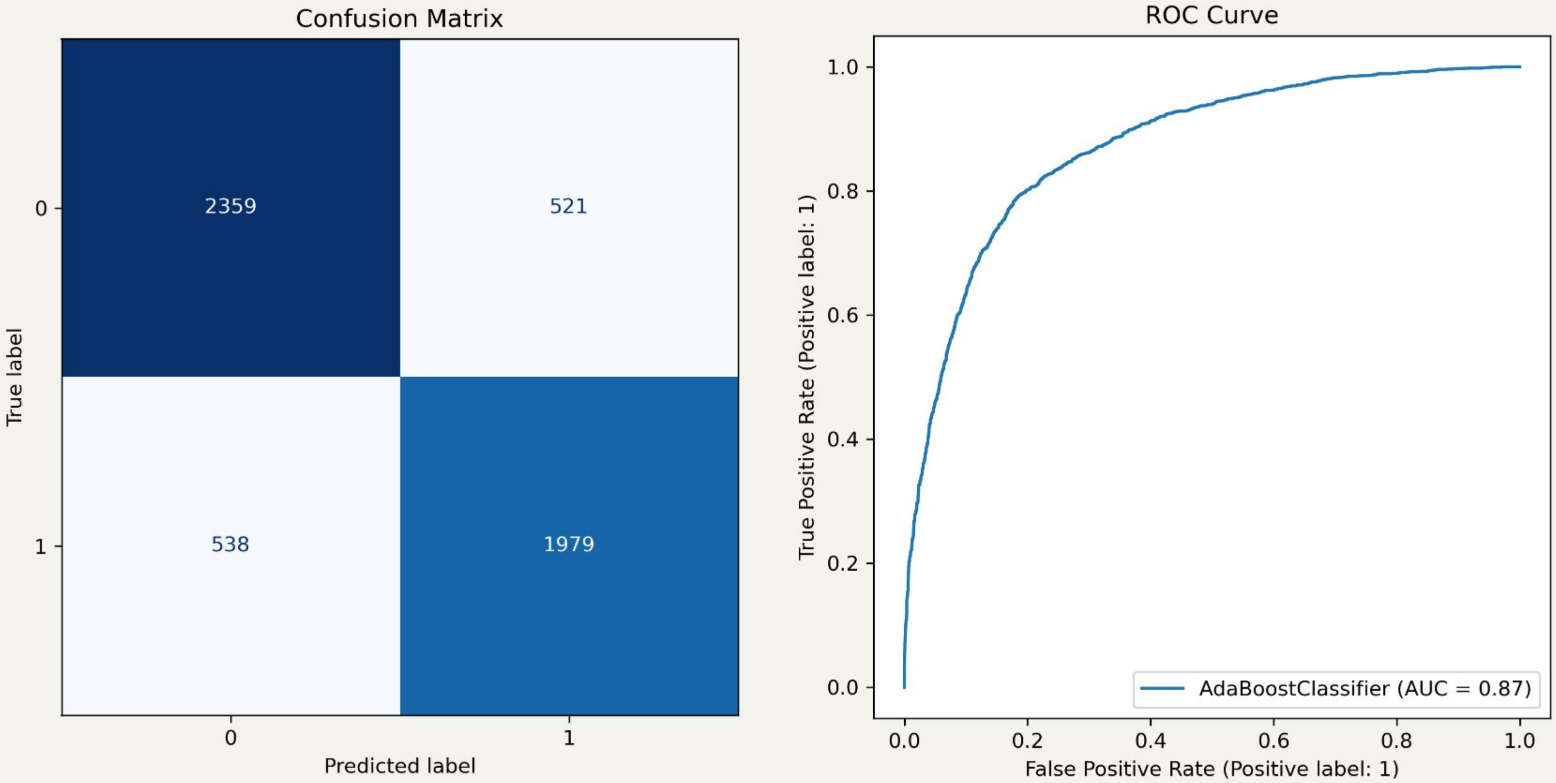
Evaluation

Assessing the models performance

Best Model: AdaBoost with GridSearchCV

- ’Number of estimators’: 300
- ’learning rate’: 2.25
- ’max depth’: None
- ’max features’: ’auto’

Model Performance

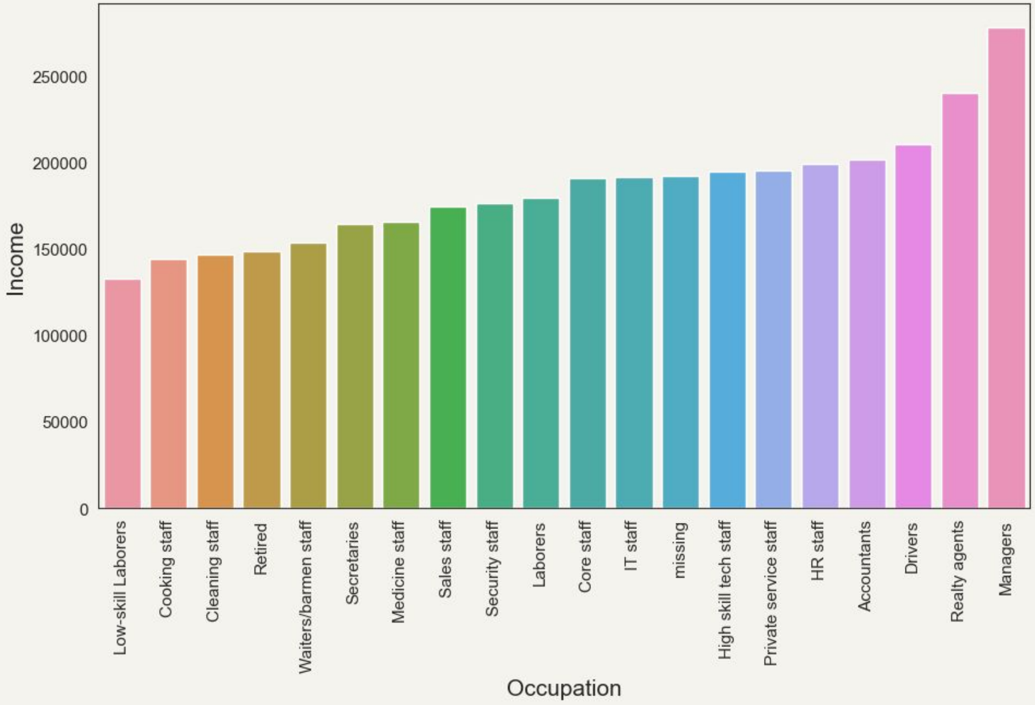


05

Limitations

Constraints in the data that can impact our models

Data Limitations



Thanks!

Do you have any questions?