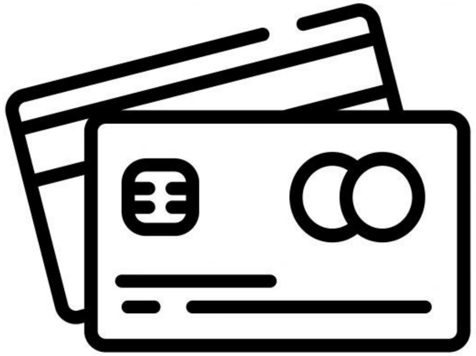
Credit Card Application Evaluator

Team 7: Muhammad Asad Shoaib, Jinghong Peng
Carnegie Mellon University

Value

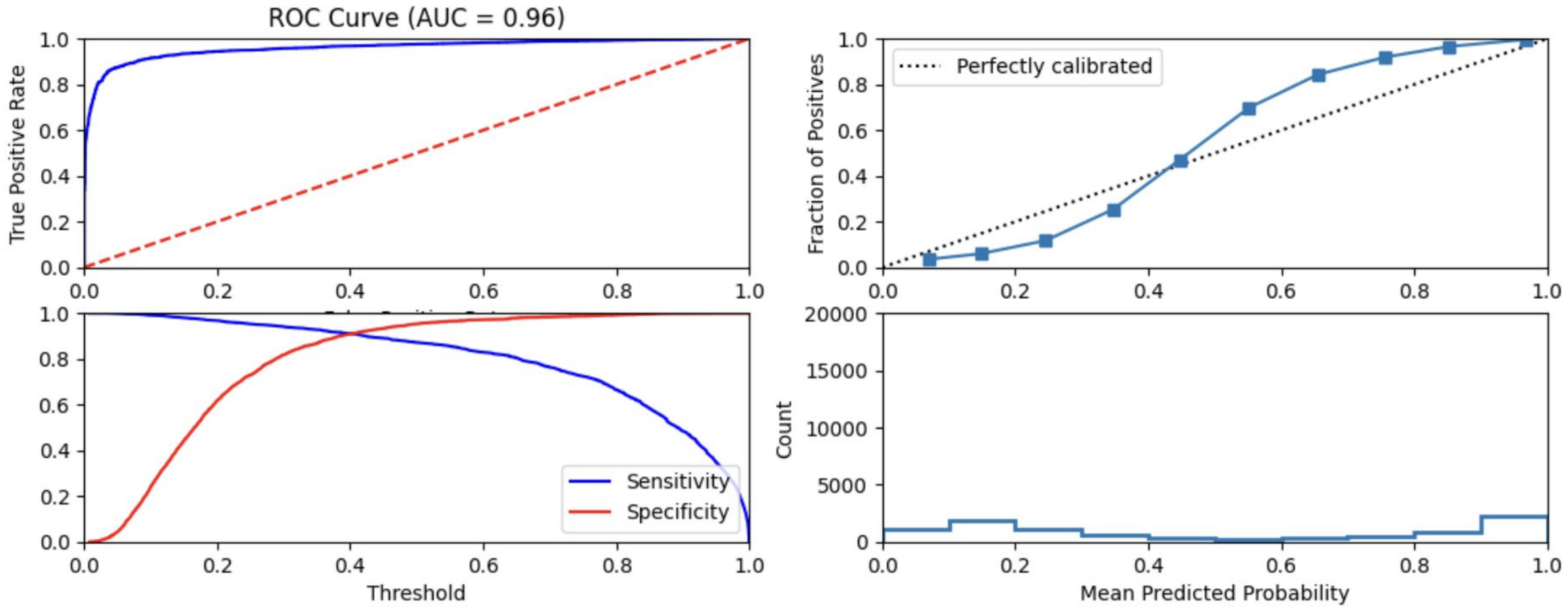


It will benefit financial institutions by improving decision-making processes. It will enhance the efficiency and effectiveness of credit risk assessment in the financial sector.

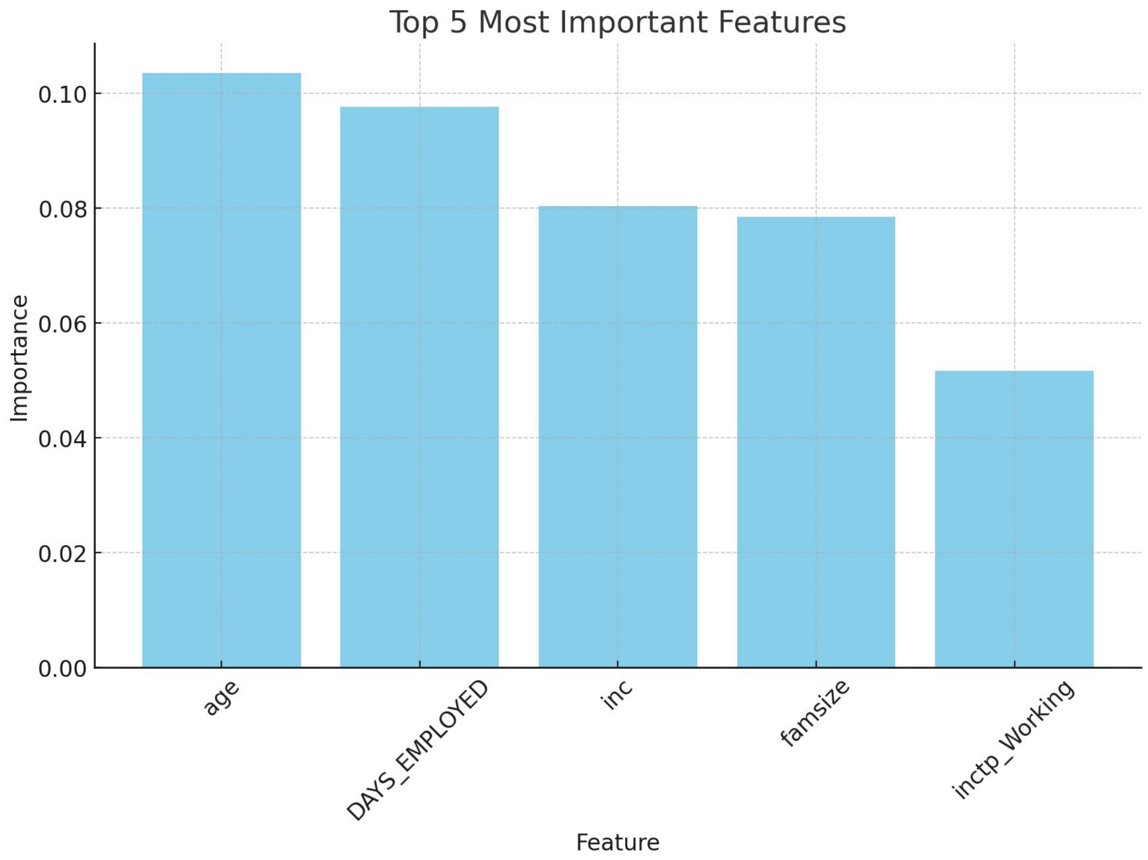
Provide a tool that use machine learning to accurately predict the outcomes of credit card applications.

Results

Our models (RF) can predict credit card rejection with up to 91% precision.



Key Findings - feature importances



age: 0.103

Days_Employed: 0.0976

inc: 0.0803

famsize: 0.0784

Car: 0.0294

inctp_working: 0.0512

Introduction

Problem Statement

We use this definition for machine learning

$P(T, E+ \Delta E) > P(T,E)$,

Where,

P => Precision and AUC possibly area under the receiver operating characteristic curve (AUC-ROC)

T => predict the creditworthiness of credit card applicants

E => the amount of data and the variety of scenarios the predictive model is exposed to during training.

Motivation

- Credit scoring is becoming increasingly vital in financial decisions.
- Forbes reported an average credit card debt of $5,474 per borrower in Q3 2022, totaling $38 billion.
- The intersection of technology and finance, notably in credit evaluation, is rapidly evolving.
- The project aims to utilize machine learning to assess 'good' or 'bad' credit risks, offering insights into improving traditional financial models.
- There are important ethical and regulatory aspects to consider, especially regarding transparency in using machine learning in finance.

Dataset

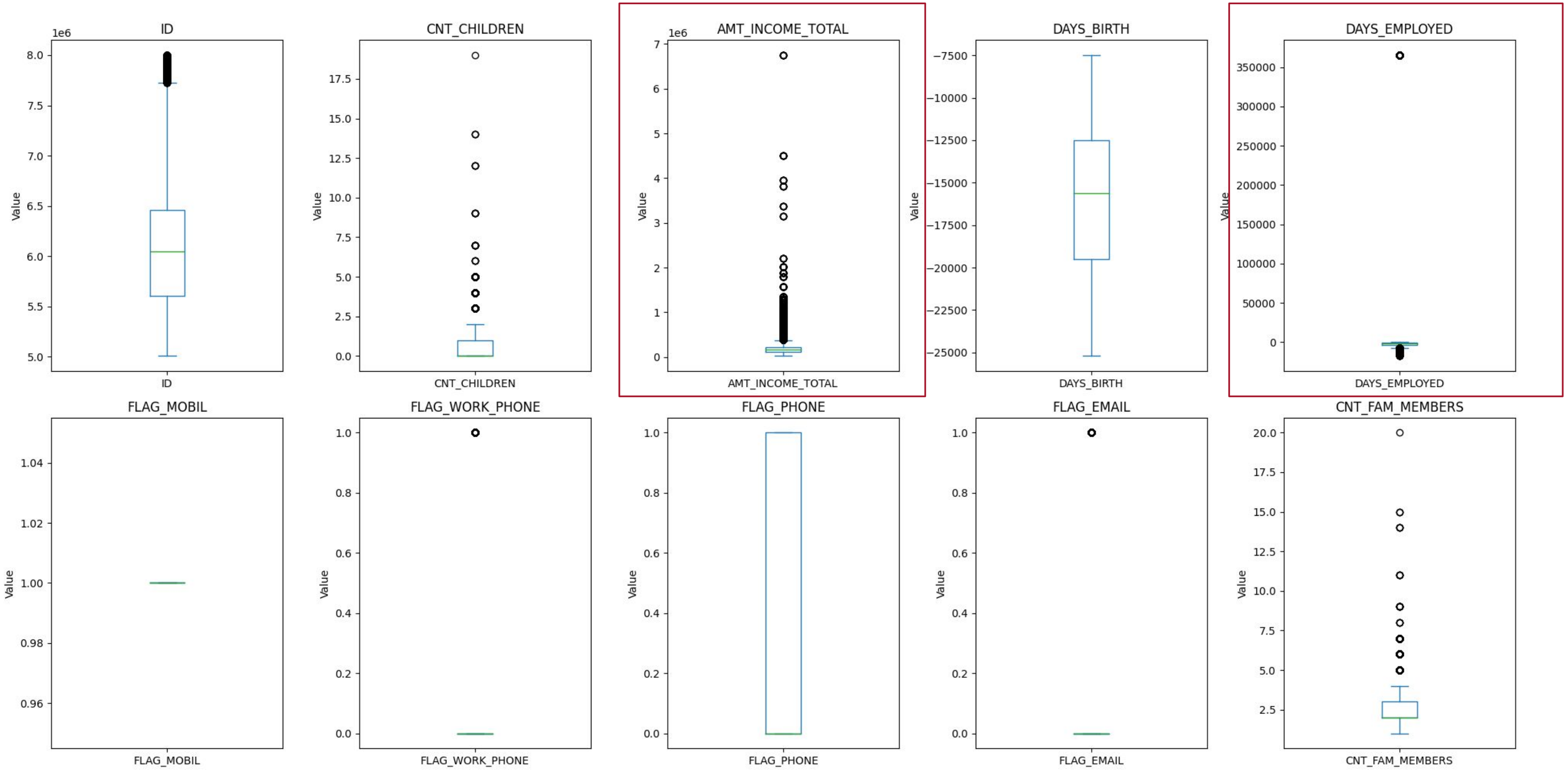
- Project dataset consists of **application_record.csv** and **credit_record.csv**, mergeable via the client number (ID).
    a. application_record.csv includes personal/financial info (gender, car ownership, income, etc.): 17 columns and ~440,000 rows
    b. credit_record.csv tracks monthly credit history, overdue days, and payments: 3 columns and ~1,000,000 rows
- Comprehensive for assessing financial behavior and creditworthiness and suitable for creating a credit scoring predictive model.
- Found during research on credit scoring and finance machine learning on Kaggle.

Data Preparation

Checks

Deduplication	None in the dataset
Sparse Column Identification	None in the dataset

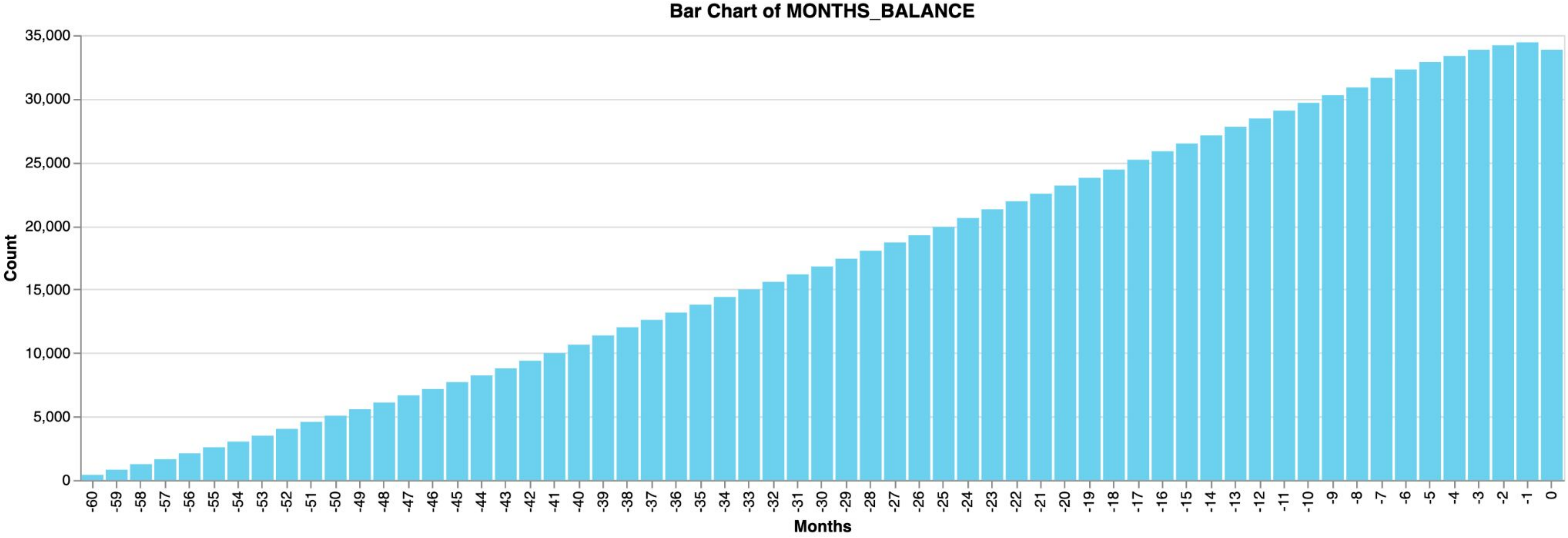
Outlier Handling - Application.csv



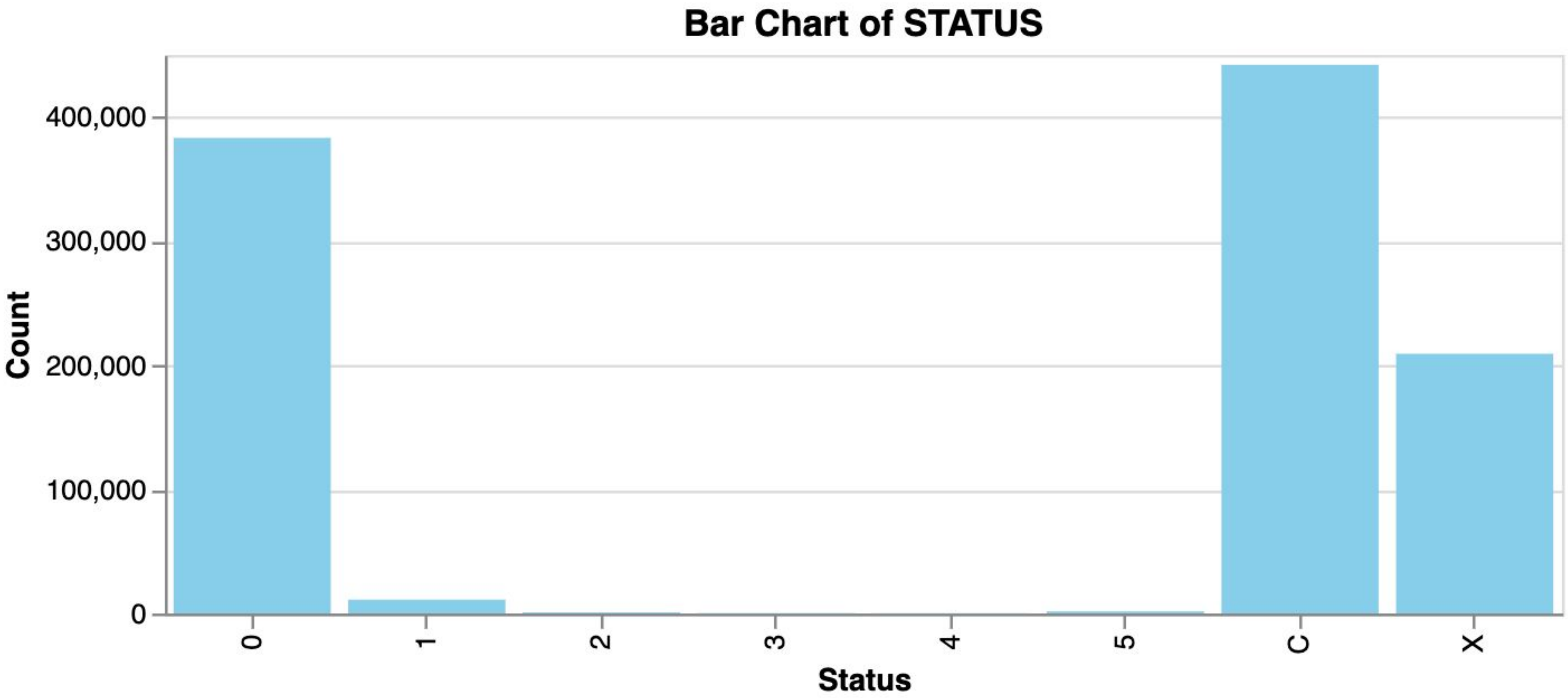
- 6M in income - not an outlier
- 350,000 (~900 years) is just a placeholder value for DAYS_EMPLOYED and hence is not an outlier either.

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Outlier Handling - credit.csv



Outlier Handling - credit.csv



- 0: 1-29 days past due
- 1: 30-59 days past due
- 2: 60-89 days overdue
- 3: 90-119 days overdue
- 4: 120-149 days overdue
- 5: Overdue or bad debts, write-offs for more than 150 days
- C: paid off that month
- X: No loan for the month

Missing Values - Application.csv

ID	0
CODE_GENDER	0
FLAG_OWN_CAR	0
FLAG_OWN_REALTY	0
CNT_CHILDREN	0
AMT_INCOME_TOTAL	0
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
NAME_HOUSING_TYPE	0
DAYS_BIRTH	0
DAYS_EMPLOYED	0
FLAG_MOBIL	0
FLAG_WORK_PHONE	0
FLAG_PHONE	0
FLAG_EMAIL	0
OCCUPATION_TYPE	134203
CNT_FAM_MEMBERS	0

- Occupation_type is missing for a lot of people
- now out of these, 75329 were unemployed so there occupation_type will not be available anyways. We replace them with N/A.
- For the rest, we will deal after we have merged the two datasets

Creating Outcome Variable and merging the datasets

- Dataset includes a monthly status variable with values 0-5, C, or X.
- Goal: Group data to determine credit application approval or rejection.
- Data dictionary definitions:
  - 0: 1-29 days past due
  - 1: 30-59 days past due
  - 2: 60-89 days overdue
  - 3: 90-119 days overdue
  - 4: 120-149 days overdue
  - 5: Over 150 days overdue or bad debts, write-offs
  - C: Paid off that month
  - X: No loan for the month
- Assess payment status: (Variable title: Status_category)
  - Overdue (1-5): **Reject application**
  - Paid on time or slightly overdue (C or 0): **Accept application**
  - No loans taken (X for all months): **Discard as not applicable**

After merging both - Remove not relevant data

ID	0
CODE_GENDER	0
FLAG_OWN_CAR	0
FLAG_OWN_REALTY	0
CNT_CHILDREN	0
AMT_INCOME_TOTAL	0
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATU	0
NAME_HOUSING_TYPE	0
DAYS_BIRTH	0
DAYS_EMPLOYED	0
FLAG_MOBIL	0
FLAG_WORK_PHONE	0
FLAG_PHONE	0
FLAG_EMAIL	0
OCCUPATION_TYPE	134203
CNT_FAM_MEMBERS	0
Status_Category_y	402100

The status_y_category is the variable that we constructed in the credit_record.csv

When we merge with application.csv, we have **402100 mismatches** - i.e., rows for which we don't have the final outcome variable.

Hence, that is not relevant for predicting approval/rejection and hence, we drop the mismatched rows.

We have **33000** rows remaining.

Imputing the missing values with MICE

```
ID                          0
CODE_GENDER                 0
FLAG_OWN_CAR                0
FLAG_OWN_REALTY             0
CNT_CHILDREN                0
AMT_INCOME_TOTAL            0
NAME_INCOME_TYPE            0
NAME_EDUCATION_TYPE         0
NAME_FAMILY_STATUS          0
NAME_HOUSING_TYPE           0
DAYS_BIRTH                  0
DAYS_EMPLOYED               0
FLAG_MOBIL                  0
FLAG_WORK_PHONE             0
FLAG_PHONE                  0
FLAG_EMAIL                  0
OCCUPATION_TYPE         11323
CNT_FAM_MEMBERS             0
Status_Category_y           0
dtype: int64
```

The remaining ~11000 missing values of occupation type were imputed using MICE

Transform Categorical Variables

Perform One-hot encoding on these variables:

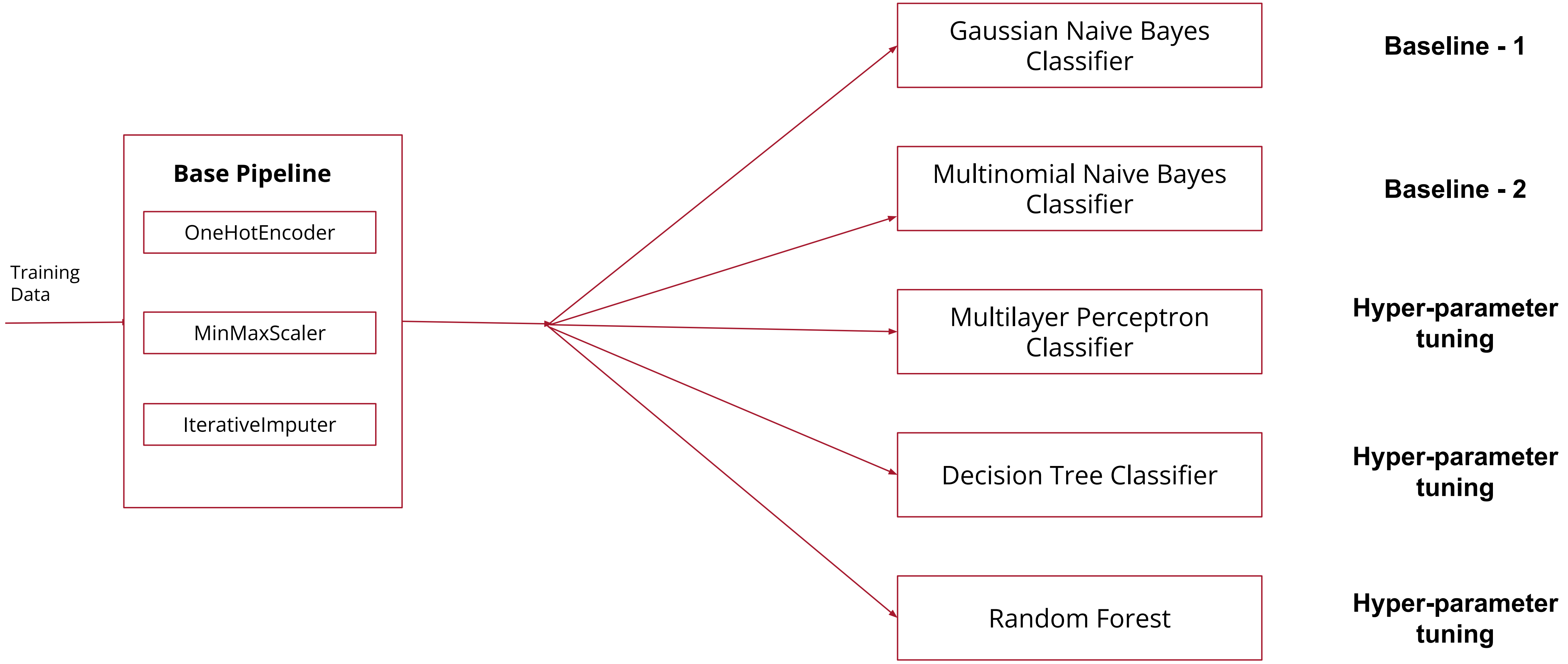
- inctp
- edutp
- famtp
- houtp
- occyp

Balance the datasets

- There are some imbalances in the outcome variable
  - 0: 28819
  - 1:4291
- We use SMOTE to balance the dataset
  - 0: 23015
  - 1: 23015

ML PIPELINE

Architecture

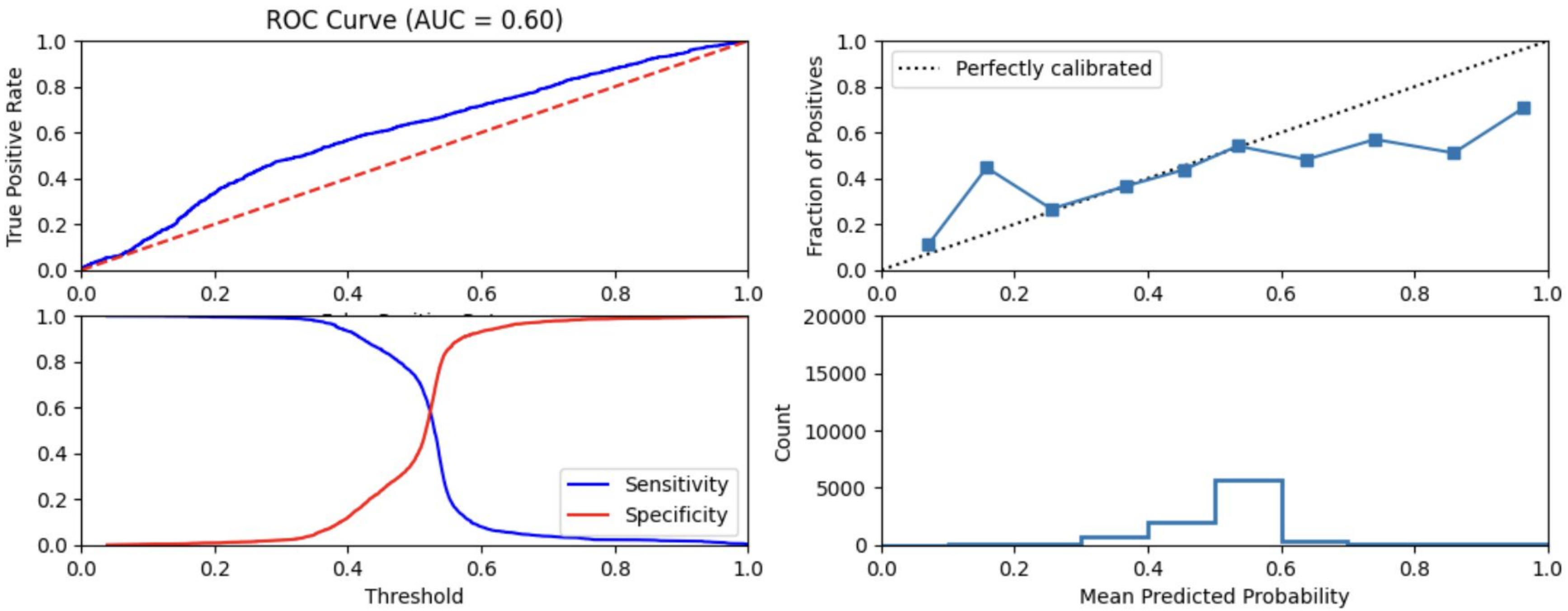


Gaussian Naive Bayes Classifier

Precision: 62%

Accuracy: 0.5896154681729306

	precision	recall
No default	0.5716	0.7106
Default	0.6191	0.4690
accuracy		
macro avg	0.5954	0.5898
weighted avg	0.5954	0.5896



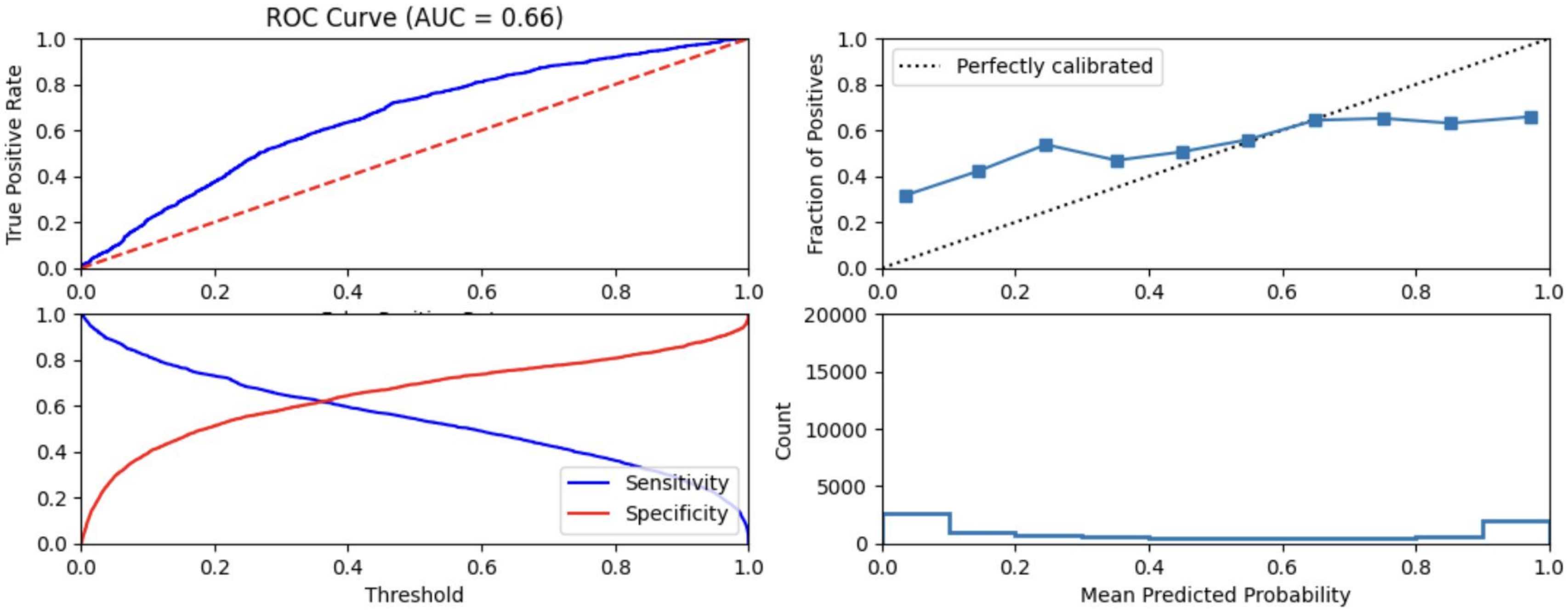
Carnegie Mellon University
Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Multinomial Naive Bayes Classifier

Precision: 60%

Accuracy: 0.6263306539213557

	precision	recall
No default	0.6534	0.5357
Default	0.6076	0.7167
accuracy		
macro avg	0.6305	0.6262
weighted avg	0.6305	0.6263



Multi-layer perceptron

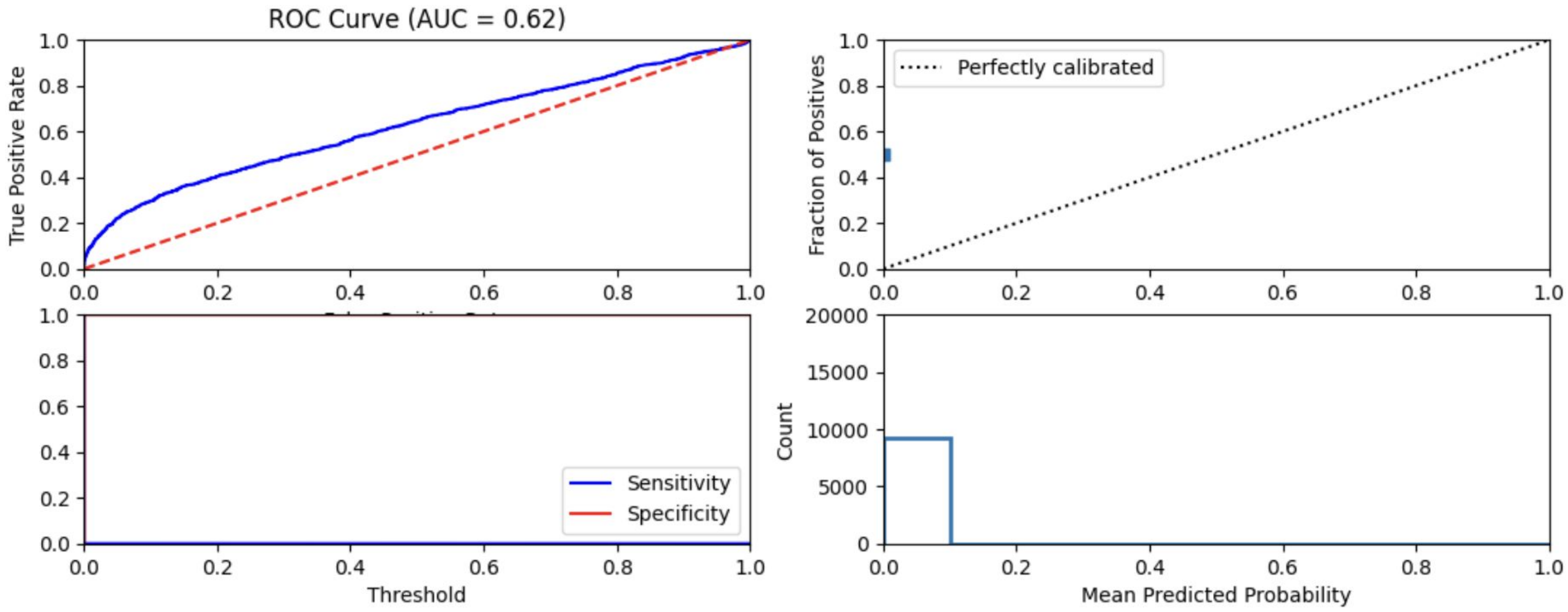
Precision: 71%

Optimal Hyper-parameters through GridSearch:
activation function: relu
max_iterations: 100

```
Accuracy:  0.6055833152291984
                  precision    recall

  No default         0.5702    0.8527
     Default         0.7098    0.3592

    accuracy
   macro avg         0.6400    0.6060
weighted avg         0.6401    0.6056
```



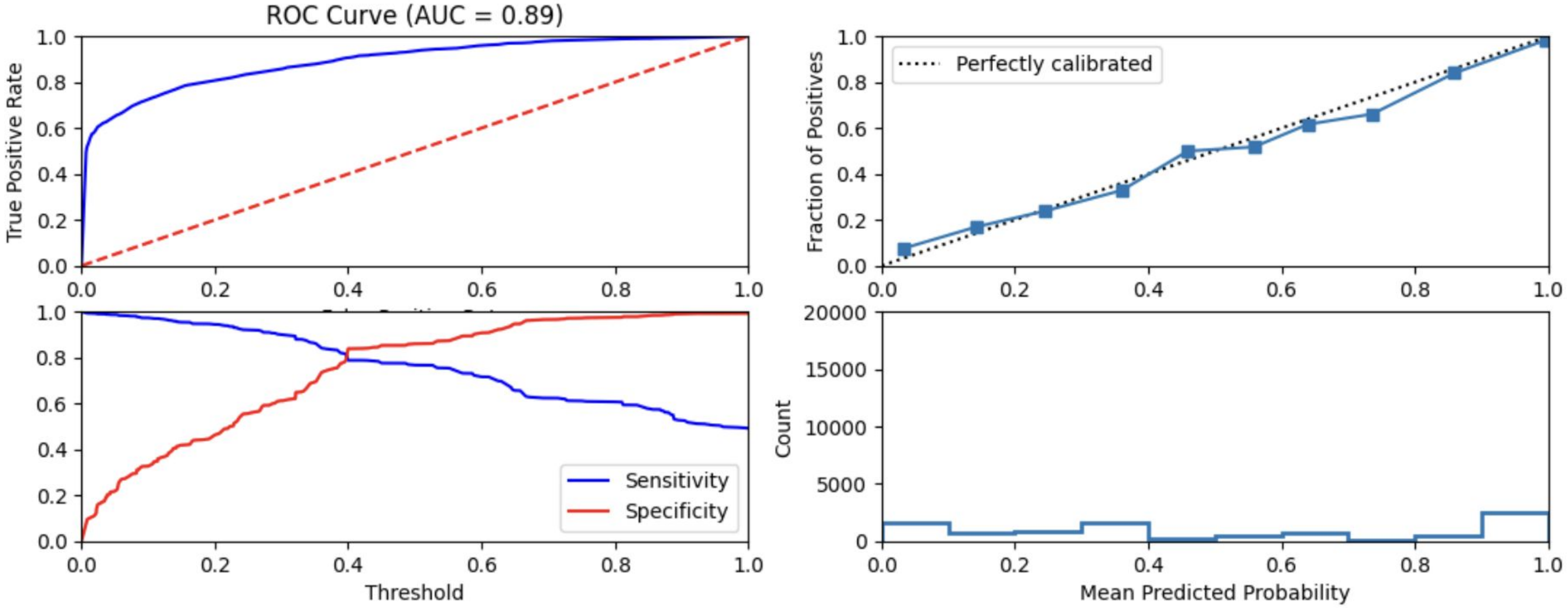
Decision Tree

Precision: 85%

Optimal Hyper-parameters through GridSearch:
loss_criterion: gini
max_depth: 10

Accuracy: 0.8134912013903975

	precision	recall
No default	0.7803	0.8718
Default	0.8553	0.7553
accuracy		
macro avg	0.8178	0.8136
weighted avg	0.8179	0.8135



Random Forest

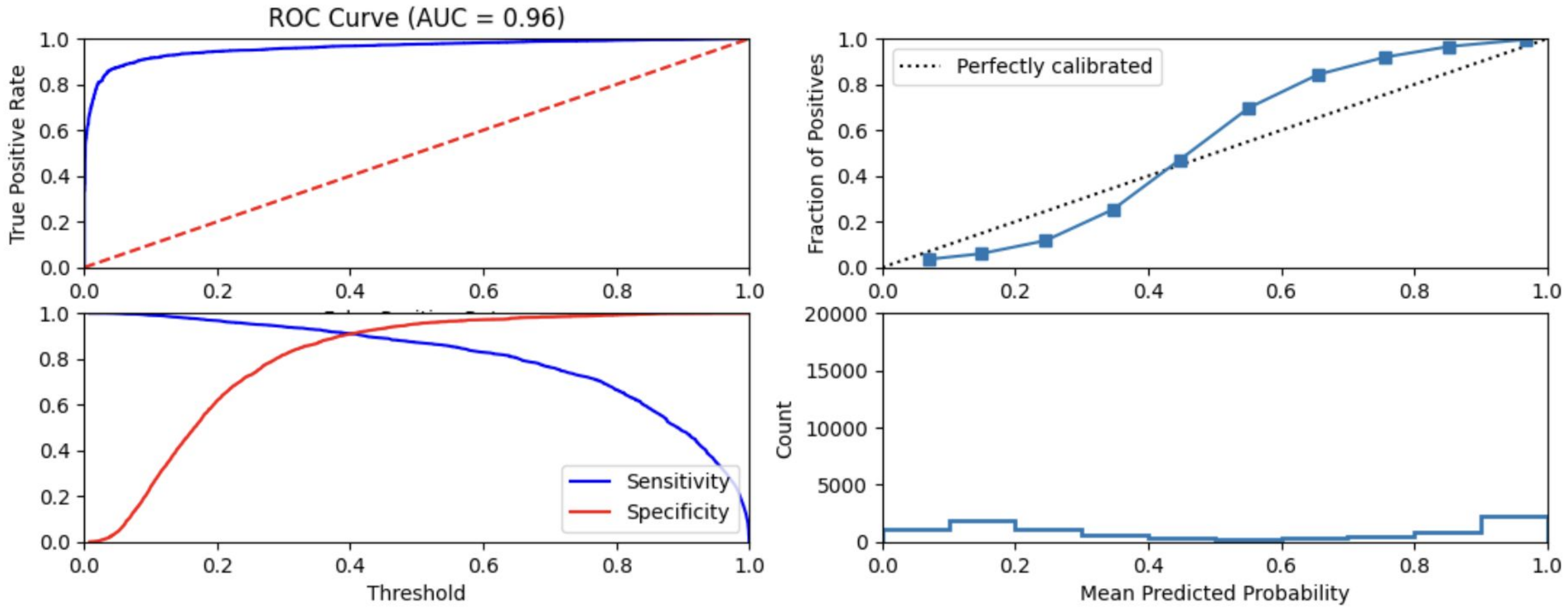
Precision: 91%

Optimal Hyper-parameters through GridSearch:
min_sample_leaves: 3
max_depth: 100
n_estimators: 100

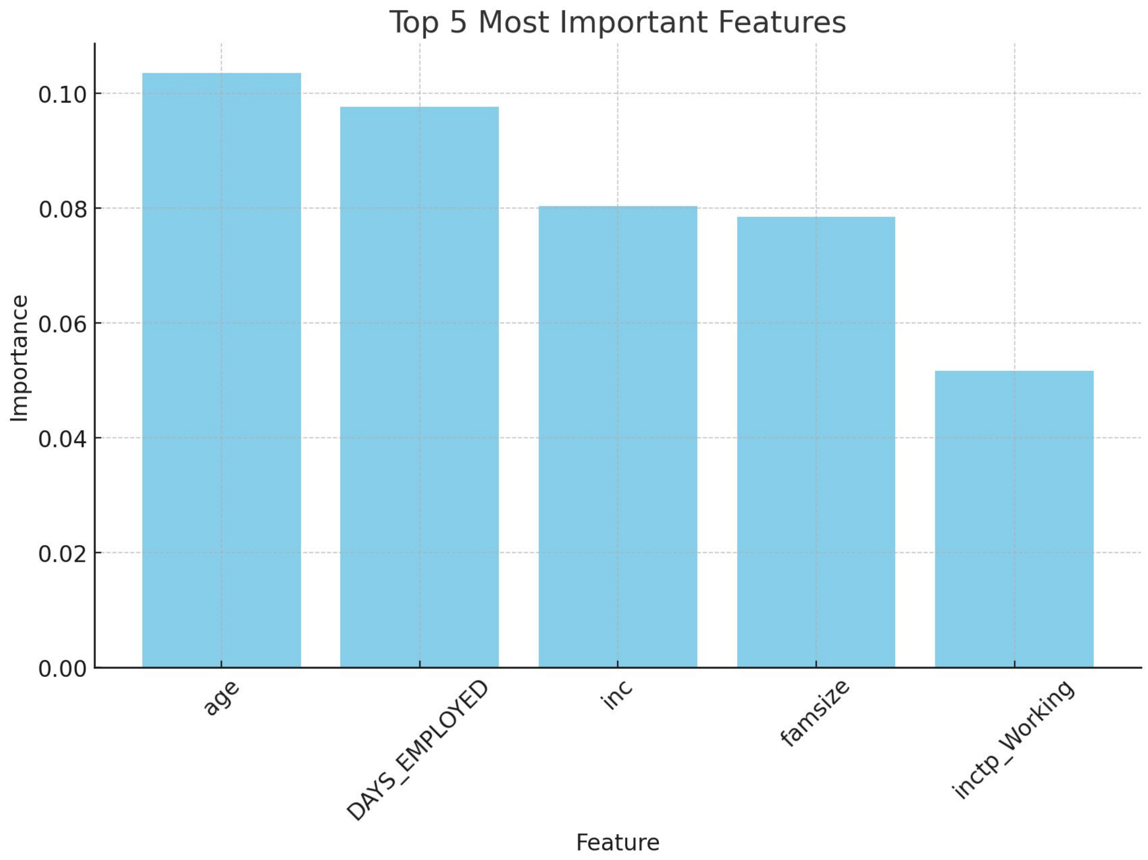
```
Accuracy:  0.9077775363893114
              precision    recall

  No default     0.9144    0.8995
     Default     0.9014    0.9161

    accuracy
   macro avg     0.9079    0.9078
weighted avg     0.9079    0.9078
```



Key Findings - feature importances



age: 0.103

Days_Employed: 0.0976

inc: 0.0803

famsize: 0.0784

Car: 0.0294

inctp_working: 0.0512

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Recommendations and Way Forward

- **Age and employment** critical to approving credit card apps
- Have **tailored strategies** for different age groups and employment categories
- Consider **personalized credit offerings** based on family dynamics

THANK YOU!