1994

CENSUS BUREAU

INCOME

Mannheim University

Miguel Mendes,2179726

João Ferreira, 2179738

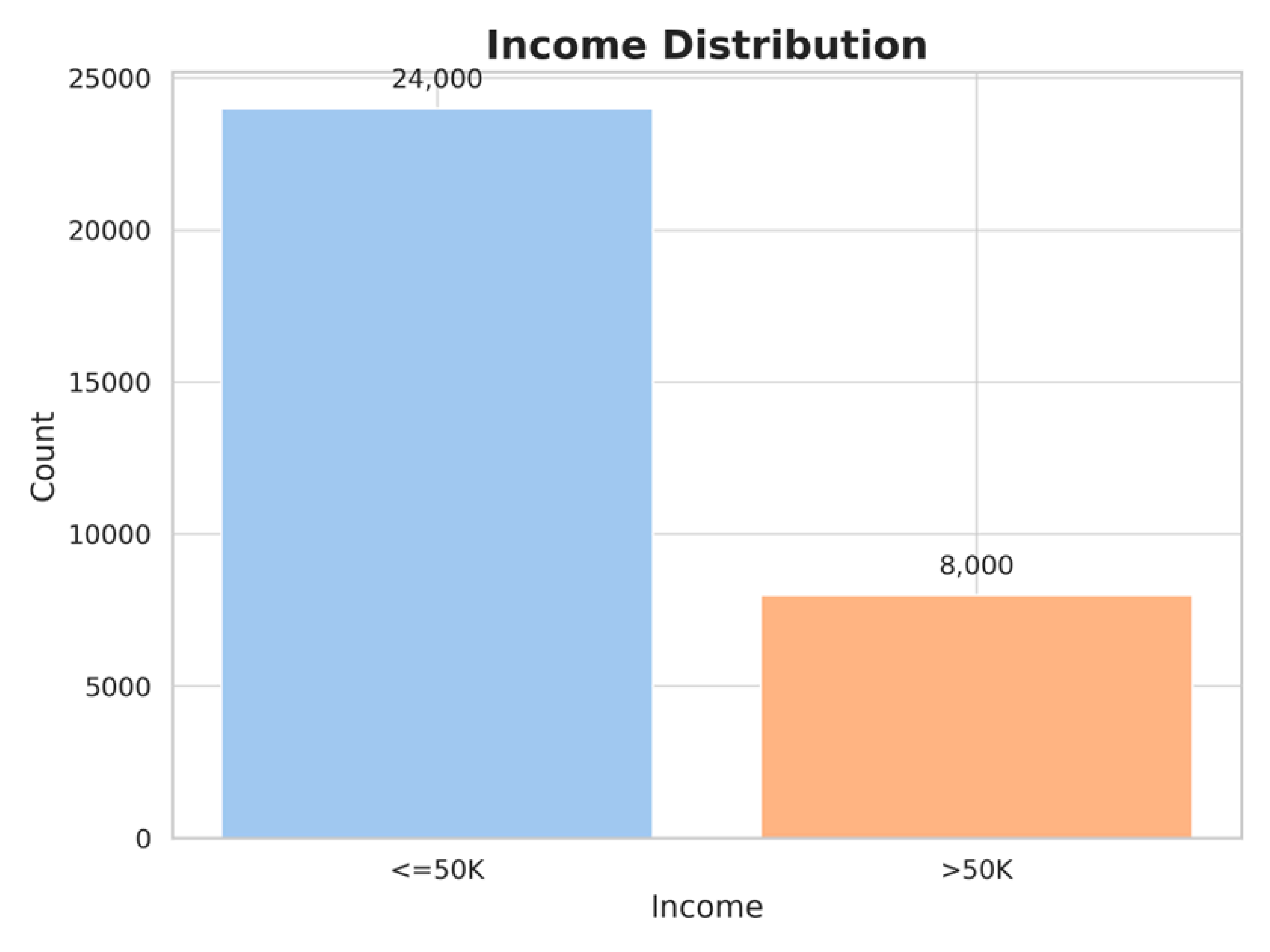
Maria Beili Mena, 2177377

Paola Tomorri, 2033630

Klea Hoxha, 1961755

Sueda Sogutlu, 1978962

What Brought Us Here



GOAL

Predict whether an individual earns more or less than $50k per year

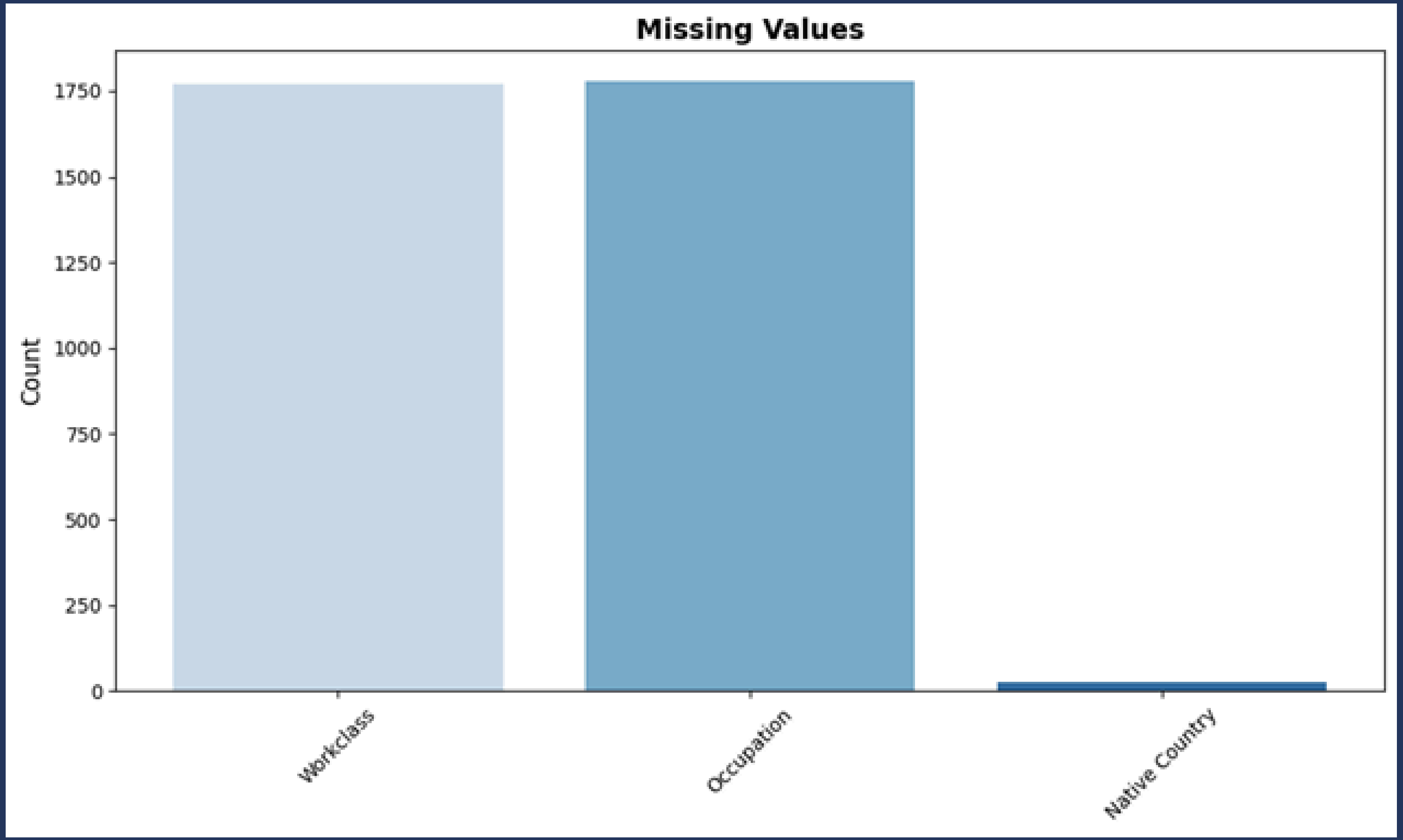
Data Profile and analysis

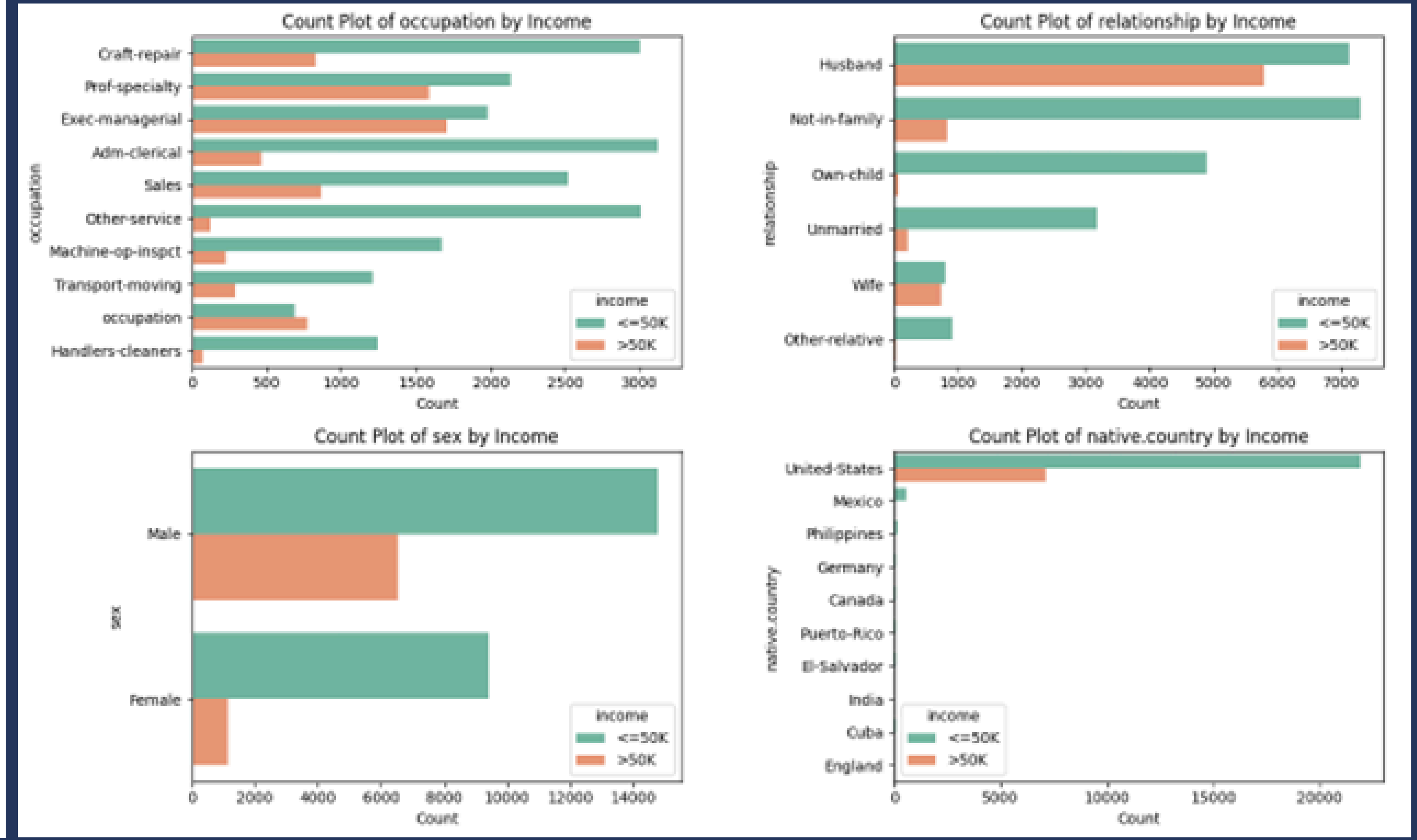
DROP

education → redundant with education.num
fnlwgt → outliers & low relevance

1994 Census Dataset Columns

Column Name	Description
age	Age of the individual.
workclass	Type of employment (e.g., Private, Self-emp-not-inc, Self-emp-inc, Federal-gov)
fnlwgt	Final weight; estimates the number of people each observation represents in the population
education	Highest level of education attained (e.g., Bachelors, HS-grad, 11th, Masters)
education,num	Numerical representation of education, often corresponding to years of education (e.g., 13 for Bachelors)
marital status	Marital status of the individual (e.g., Tech-support, Craft-repair, Other-sevice, Sales)
relationship	Relationship status (e.g., Wife, Own-child, Husband, Not-in-family, Unmarried)
race	Ethnicity or race (e.g., White, Black, Asian-Pac-Islander, Amer-indian-Estkimo)
sex	Gender of the individual (e.g., Male, Female)
native.country	Country of origin (e.g., United-States, Cambodia, England, Puerto-Rico)
income	Target variable; indicates if annual income exceeds $50K (>50K or <=50K)


Missing Values
Count
1750
1500
1250
1000
750
500
250
0
Workclass
Occupation
Native Country

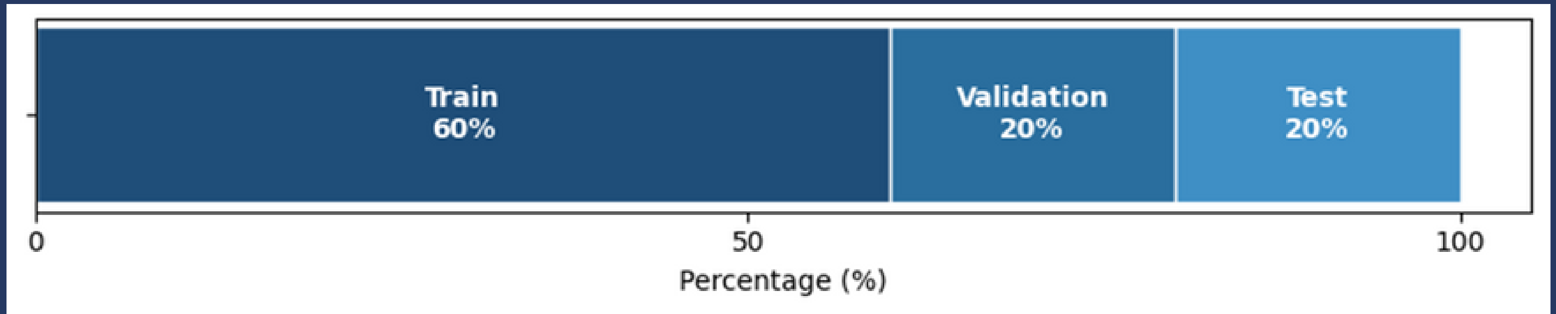

Count Plot of occupation by Income
Craft-repair
Prof-specialty
Exec-managerial
Adm-clerical
Sales
Other-service
Machine-op-inspct
Transport-moving
occupation
Handlers-cleaners
income
<=50K
>50K
Count
Count Plot of relationship by Income
Husband
Not-in-family
Own-child
Unmarried
Wife
Other-relative
relationship
Count
Count Plot of sex by Income
Male
Female
sex
Count
Count Plot of native.country by Income
United-States
Mexico
Philippines
Germany
Canada
Puerto-Rico
El-Salvador
India
Cuba
England
native.country
Count

Preprocessing

Data Preprocessing Techniques

Variable	Preprocessing Method
income	Binary (0: <=50K, 1: >50K)
age	MinMax Scaling
fnlwgt	MinMax Scaling
education.num	MinMax Scaling
workclass	One-Hot Encoding
education	One-Hot Encoding
marital.status	One-Hot Encoding
occupation	One-Hot Encoding
relationship	One-Hot Encoding
race	One-Hot Encoding
sex	One-Hot Encoding
native.country	One-Hot (all nationalities) or Binary (US: 1, else: 0)

Dataset Split



Model

Basic Models

	Train F1 Score	Validation F1 Score
Baseline Model	0.66	0.66
Logistic Regression	0.83	0.83
Decision Tree	0.95	0.79
Random Forest	0.95	0.81
Gradient Booster	0.84	0.84
SVM	0.82	0.83
BernoulliNB	0.77	0.76

Optimization: Used Grid Search and **Random Search**

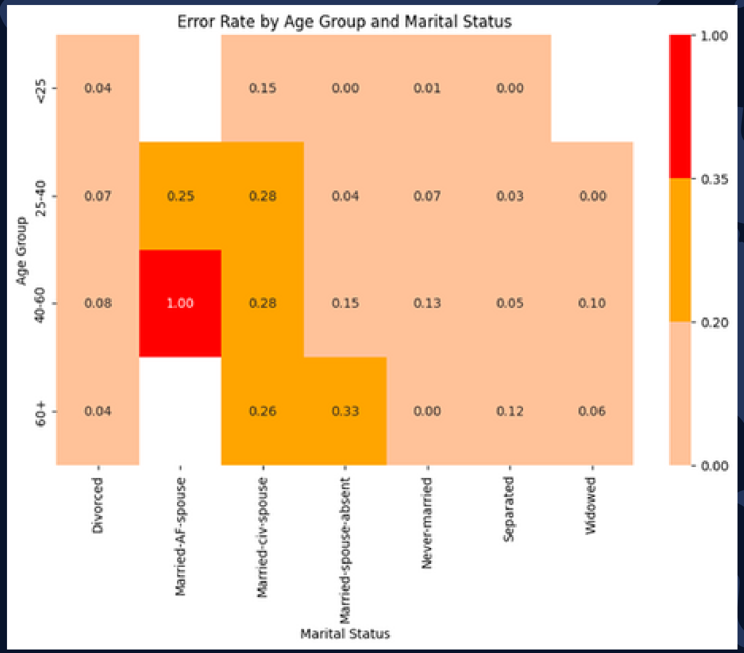
Final Decision: Gradient Boosting slightly outperformed Random Forest in Recall & Precision

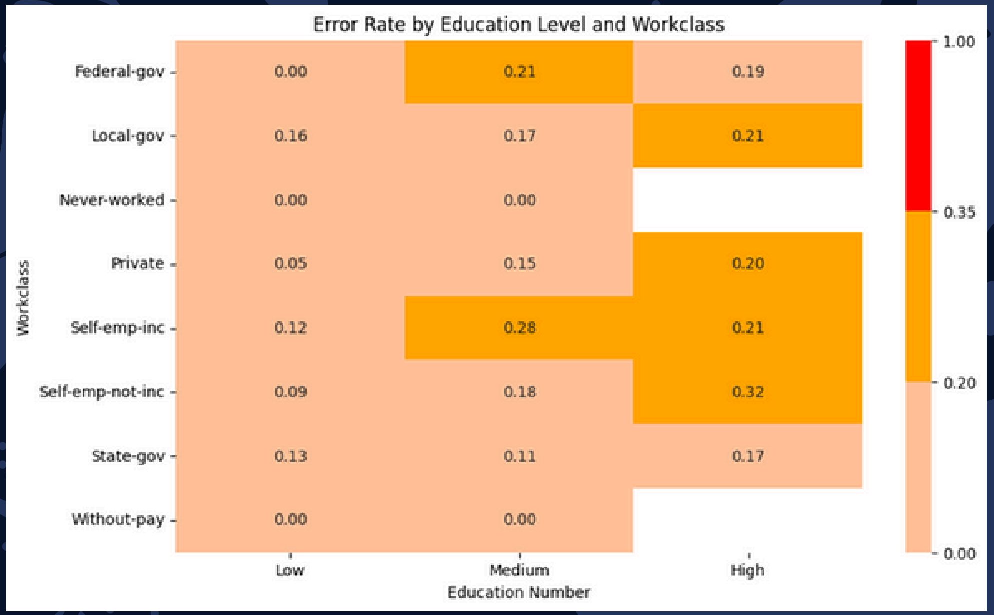
Optimized Models

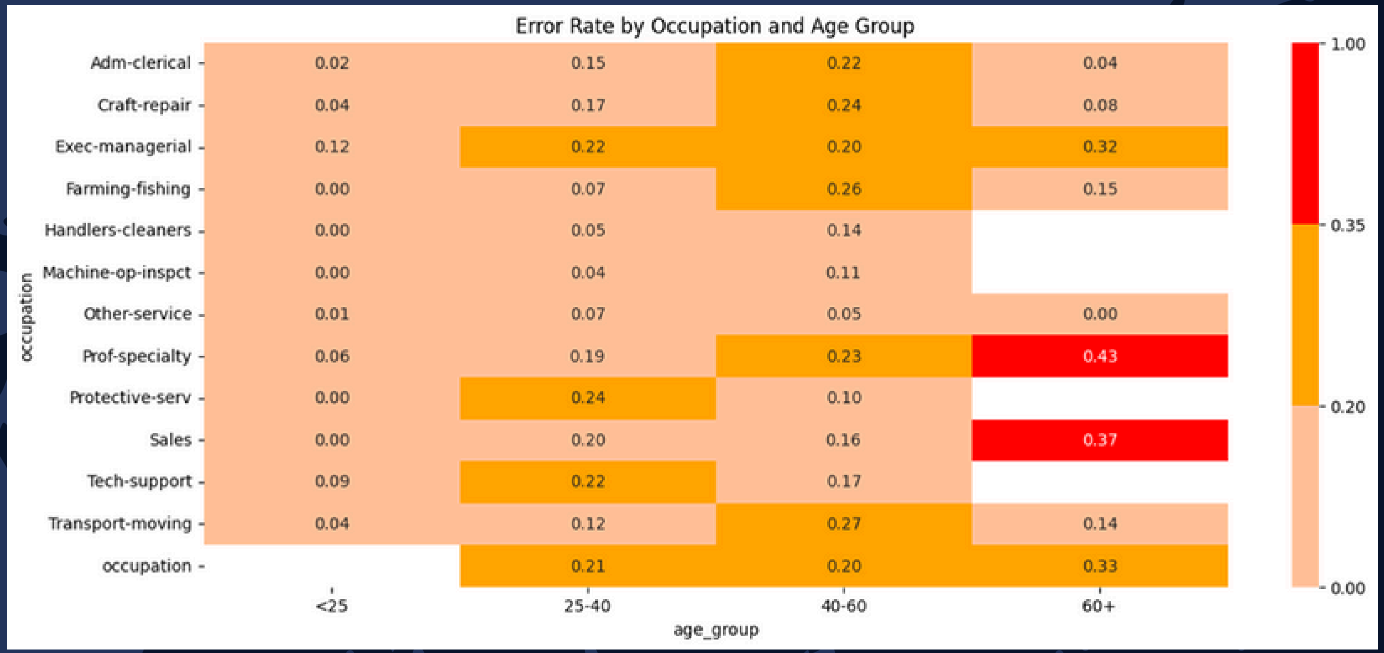
	Train Score	Validation Score
Random Forest	0.84	0.84
Gradient Boosting	0.84	0.84
SVC	0.83	0.83
Logistic Regression	0.83	0.83

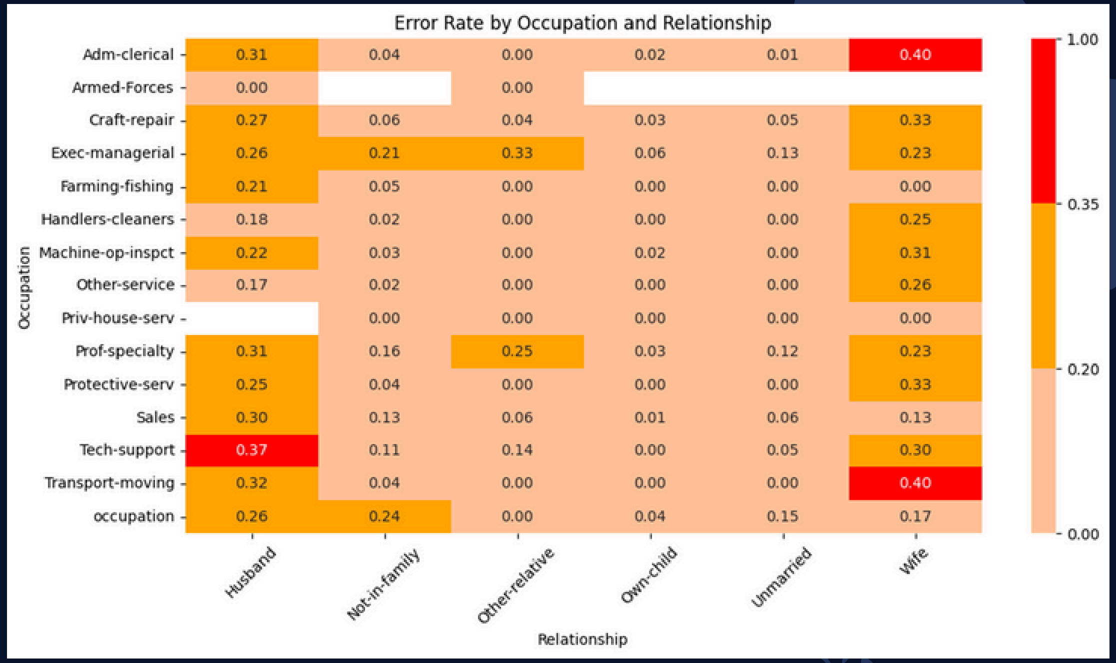
Analysis of the best model

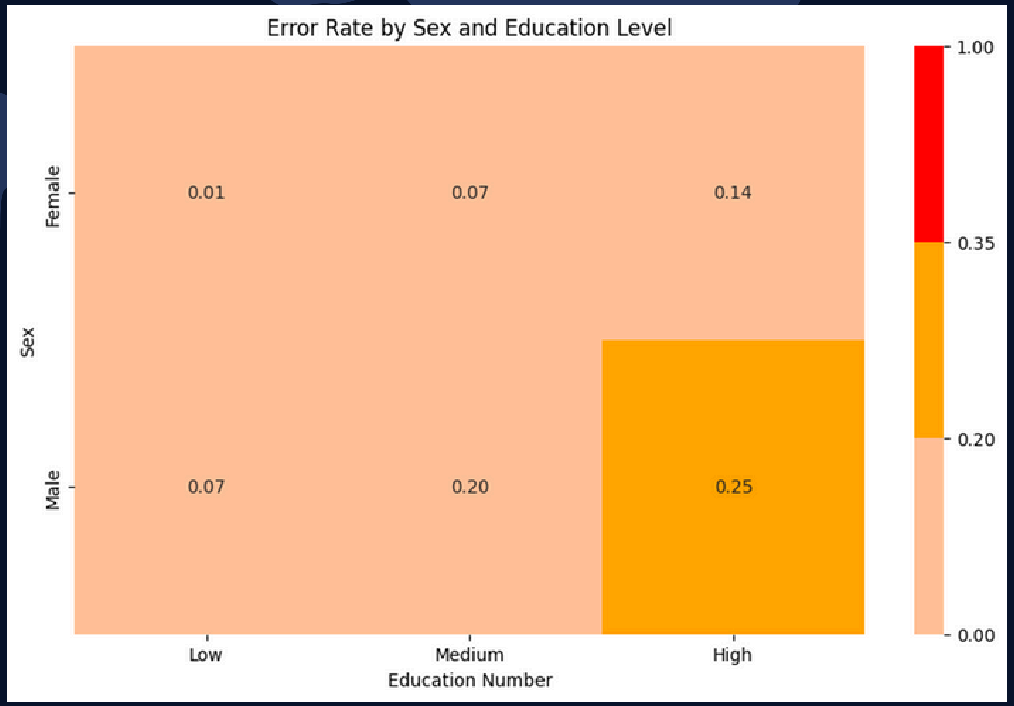
Challenges and error analysis of the best model

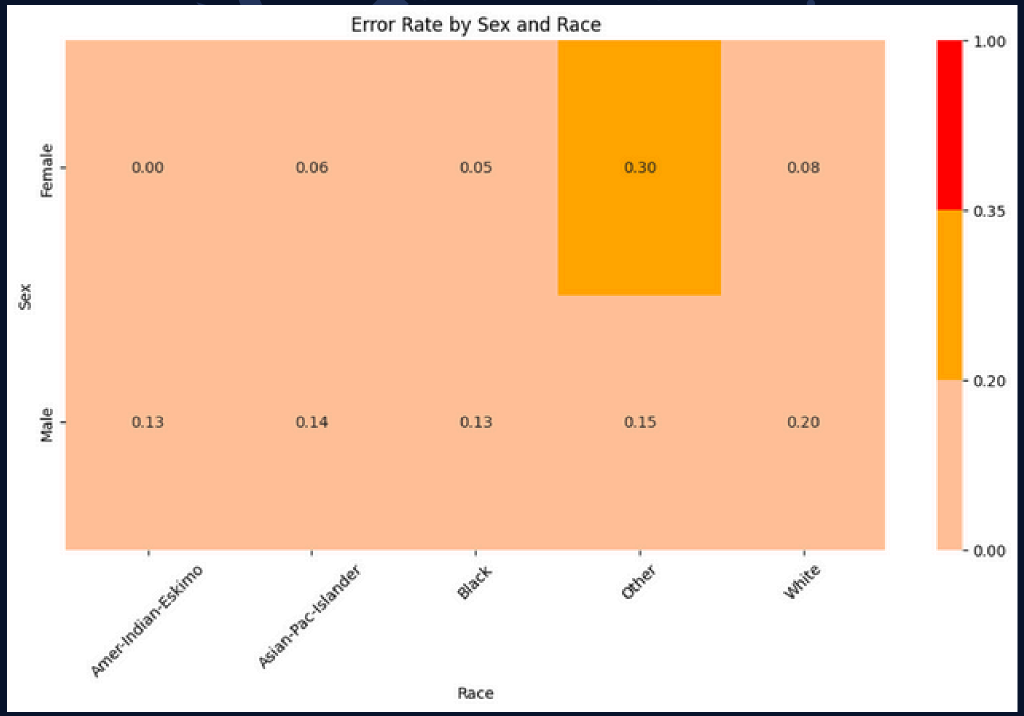








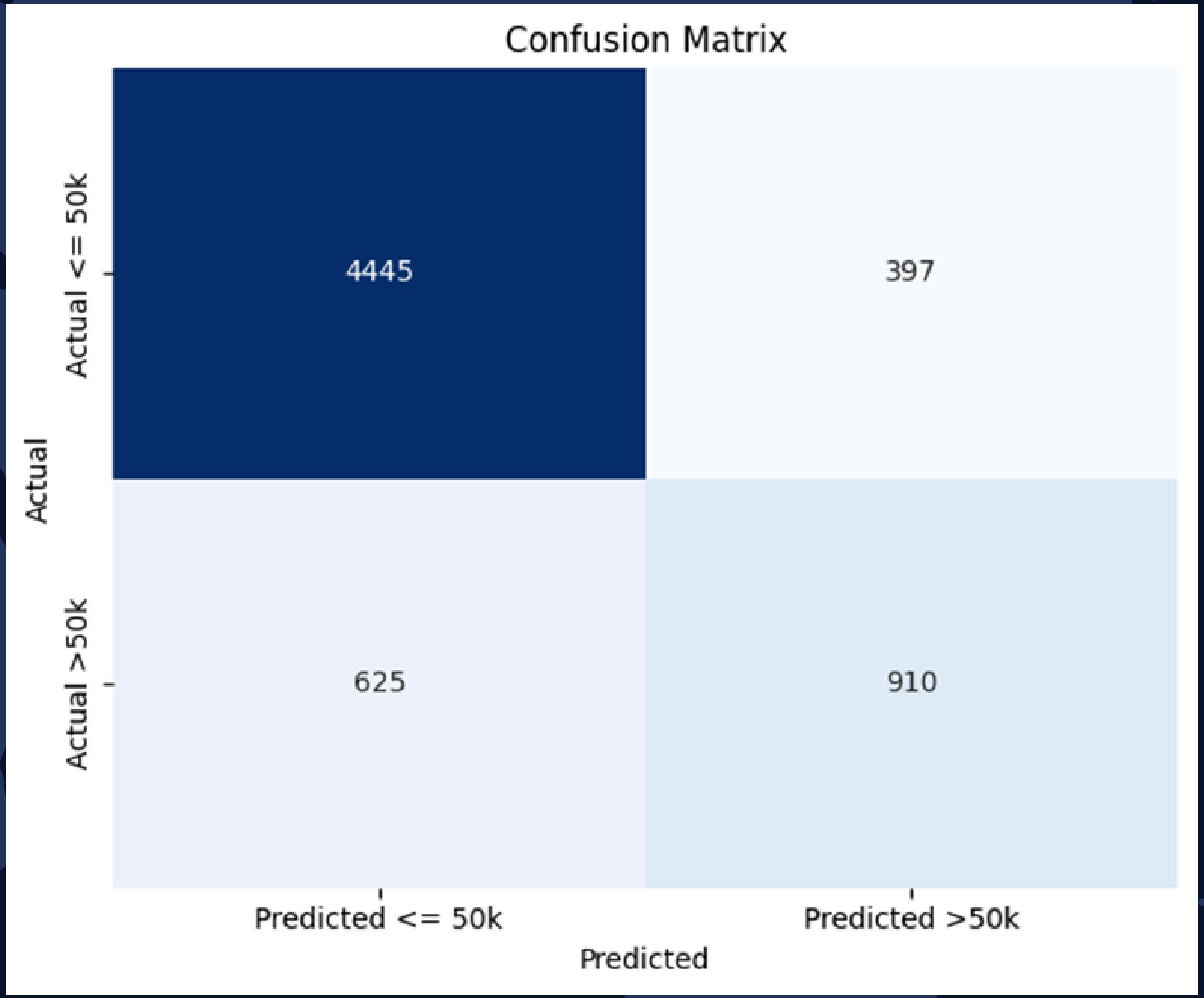




Results on the test data

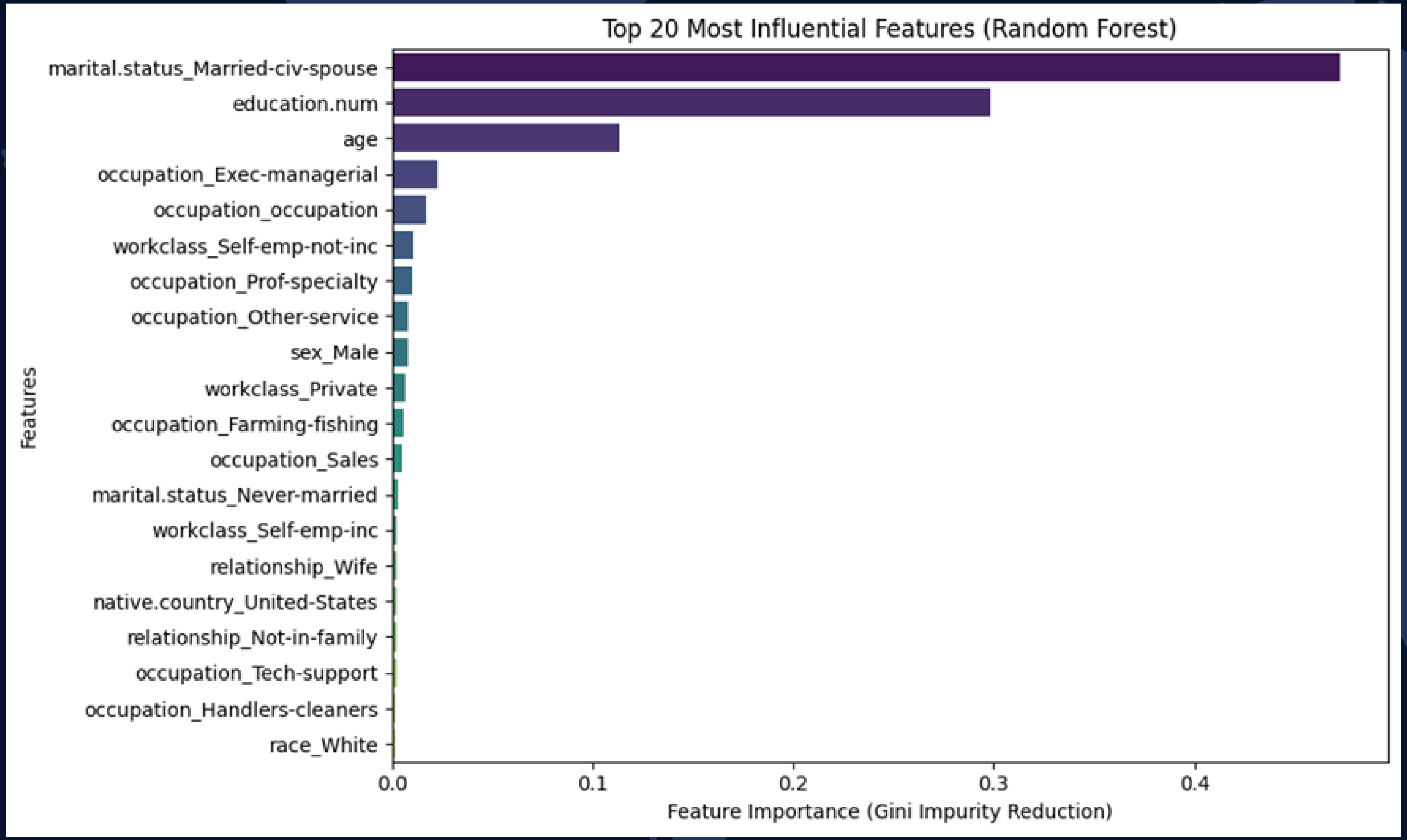
	Precision	Recall	F1-Score
0	0.88	0.92	0.9
1	0.7	0.59	0.64
accuracy			0.84
macro avg	0.79	0.76	0.77
weighted avg	0.83	0.84	0.84

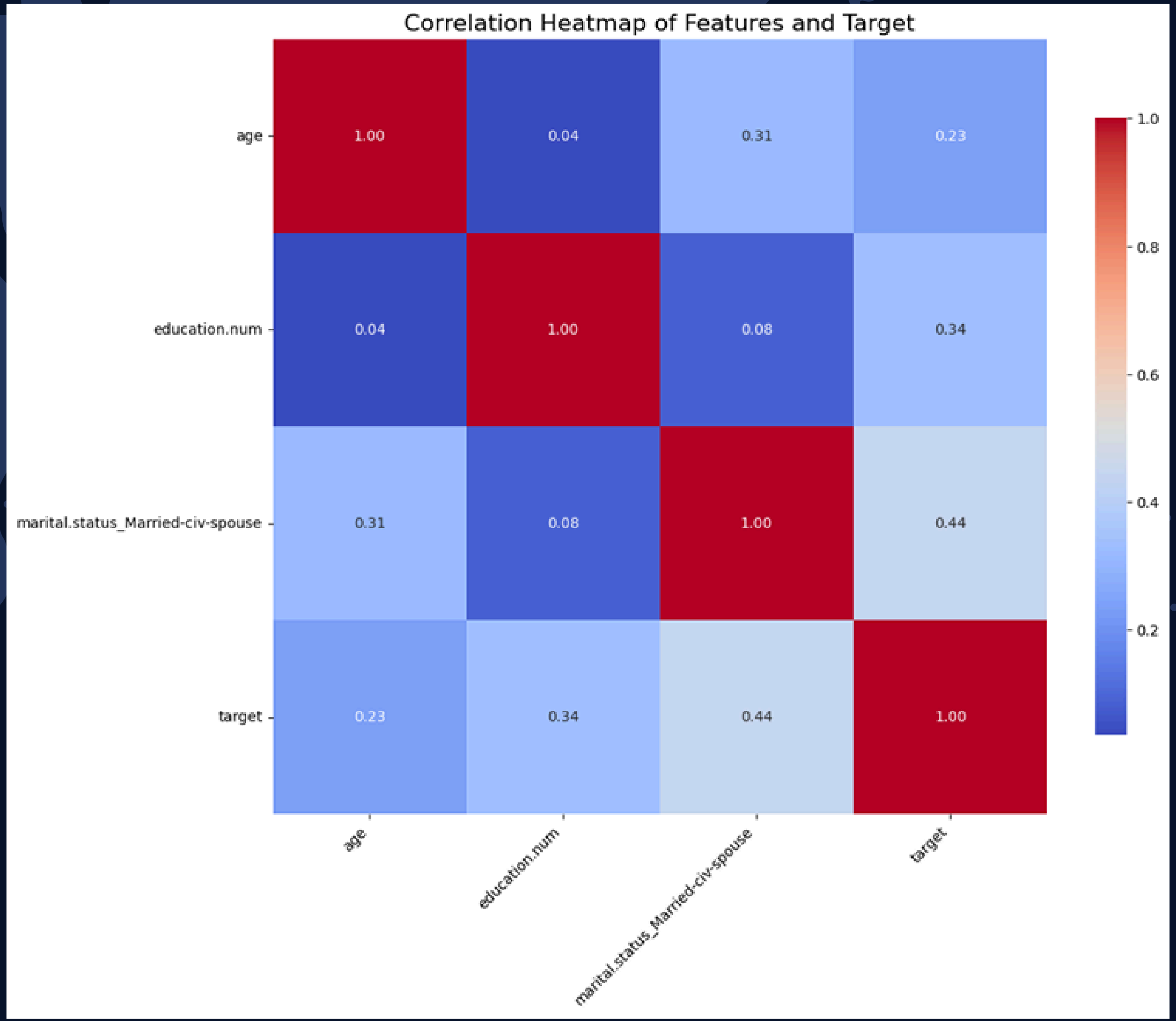
✅ **No overfitting** observed
🎯 High accuracy on <=*50K* class (11× more correct than wrong)
🥇 Outperformed 88% of Kaggle models



Importance of features to the result

📊 Model Used: Random Forest
🏆 Top 3 features: **Marital Status**, **Education Num**, **Age**
📈 All show **positive correlation** with income > $50K





UNIVERSITÄT MANNHEIM

THANKS FOR YOUR ATTENTION

Mannheim University
Miguel Mendes,2179726
João Ferreira, 2179738
Maria Beili Mena, 2177377
Paola Tomorri, 2033630
Klea Hoxha, 1961755
Sueda Sogutlu, 1978962