Credit Card Approval Prediction

Group 9:
Stella Ahn, Matt Eichenberg, Tyler Ippolito,
David Shatsky, Spencer Warshaw

Intro

- Analyze credit card approval data to identify key factors influencing approval decisions
- Kaggle dataset…
- **Objectives:**
  1. Determine demographic trends affecting approval
  2. Improve performance in predicting approvals

Research Questions

1. What demographic factors most significantly influence credit card approval rates?

2. How can predictive modeling techniques be optimized to accurately forecast credit approval outcomes?

Data Description

- Application records
  - 438,557 entries

ID, gender, age, income, education, occupation, and family status

- Credit Record
  - 1,036,231 entries

Credit records of applicants

Merged on \`ID\`

77,715 records, 20 columns

Rikdifos. *Credit Card Approval Prediction*. Kaggle, 2021. https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction

Data Preprocessing

	FLAG_MOBIL	DAYS_BIRTH	AMT_INCOME_TOTAL	STATUS	OCCUPATION_TYPE
0	1	-12005	427500.0	C	NaN
1	1	-12005	427500.0	C	NaN
2	1	-12005	427500.0	C	NaN
3	1	-12005	427500.0	C	NaN
4	1	-12005	427500.0	C	NaN
...	...	...	...	...	
777710	1	-19398	202500.0	C	Drivers
777711	1	-19398	202500.0	C	Drivers
777712	1	-19398	202500.0	C	Drivers
777713	1	-19398	202500.0	C	Drivers
777714	1	-19398	202500.0	C	Drivers

777715 rows × 5 columns

	AGE	AMT_INCOME_TOTAL	STATUS_Approved	OCCUPATION_TYPE
0	32	0.258721	1	Unknown
1	32	0.258721	1	Unknown
2	32	0.258721	1	Unknown
3	32	0.258721	1	Unknown
4	32	0.258721	1	Unknown
...	...	...	...	...
777710	53	0.113372	1	Drivers
777711	53	0.113372	1	Drivers
777712	53	0.113372	1	Drivers
777713	53	0.113372	1	Drivers
777714	53	0.113372	1	Drivers

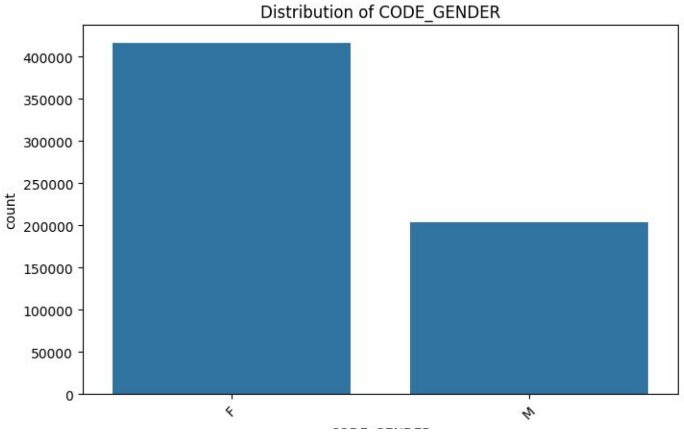
777715 rows × 4 columns

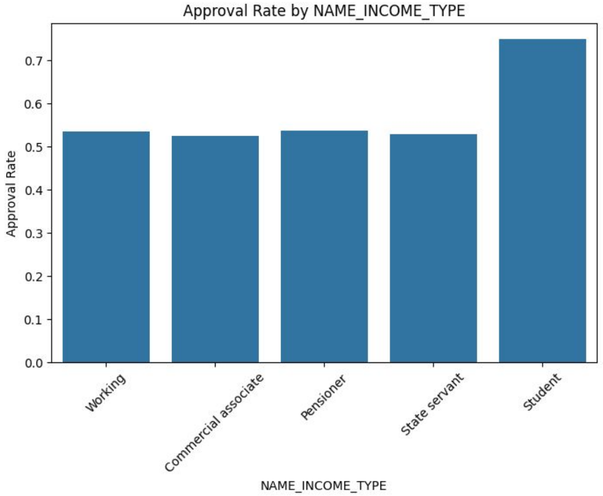
Feature Engineering

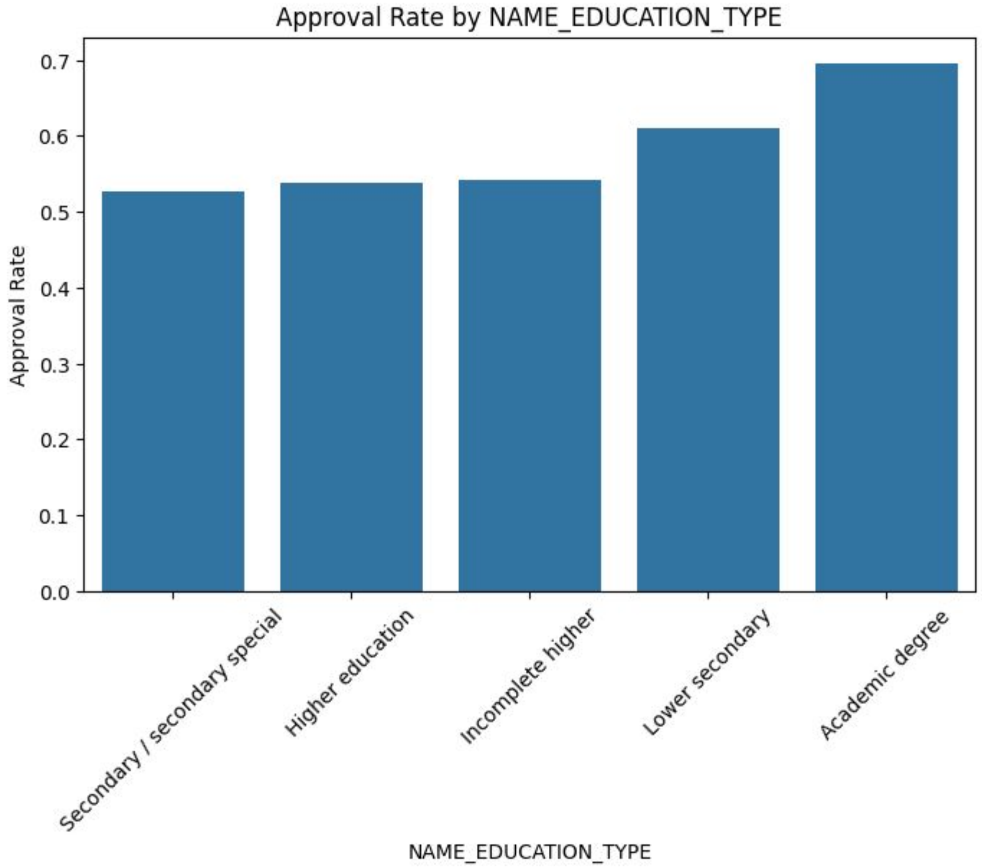
- Calculated age in years then Age Binning
- Categorical Encoding (target / one-hot encoding)
- Mapped the 'STATUS' values to create a binary target variable.
  - '0' to '5': Represent days past due
  - 'C': Indicates the credit is closed.
  - 'X': Indicates no loan for the month.
  - Assign 1 (approved) for 'C' and 'X'.
  - Assign 0 (not approved) for '0' to '5'.

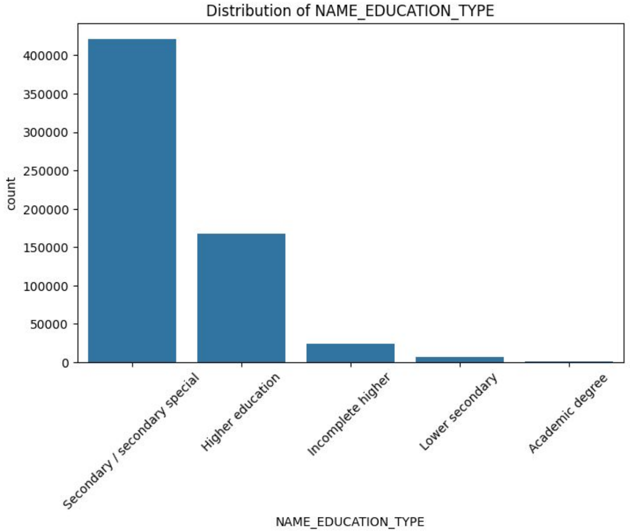
Exploratory Data Analysis

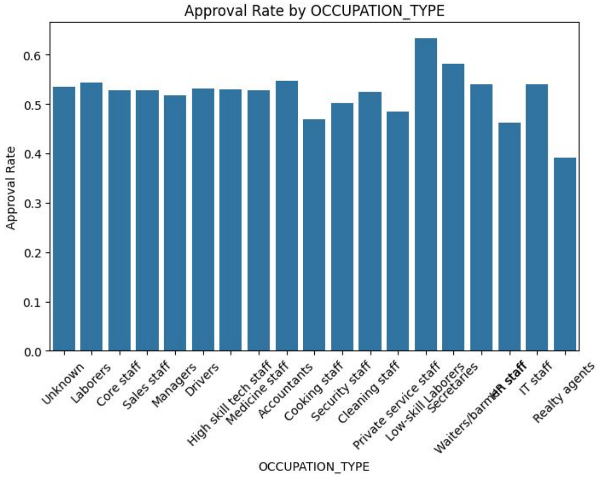
- Plotted distributions for categorical variables
- Evaluating approval rates by age, education and income type





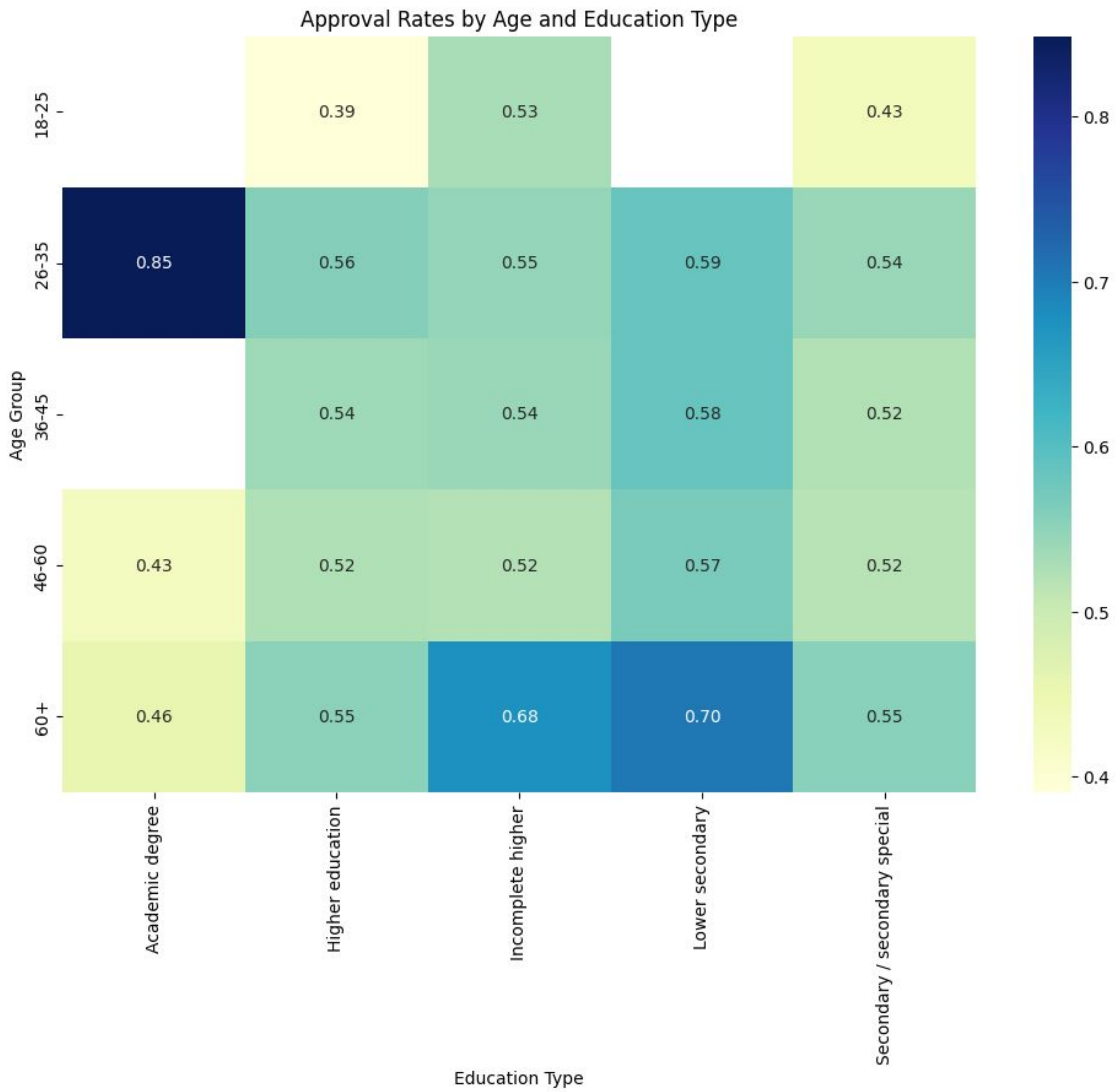




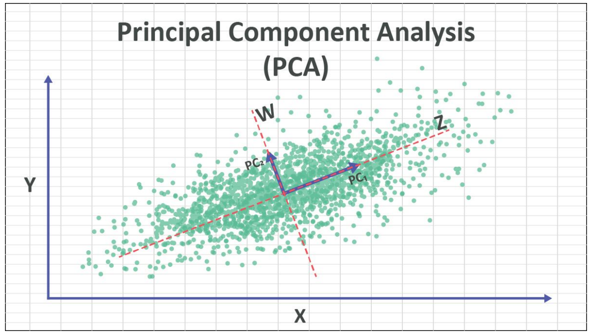


Exploratory Data Analysis

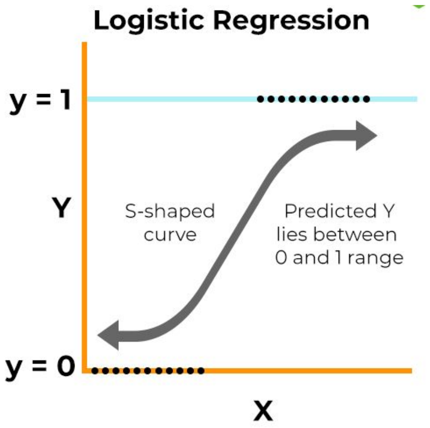
- 26-35 Age Group with an Academic Degree has the highest approval rate

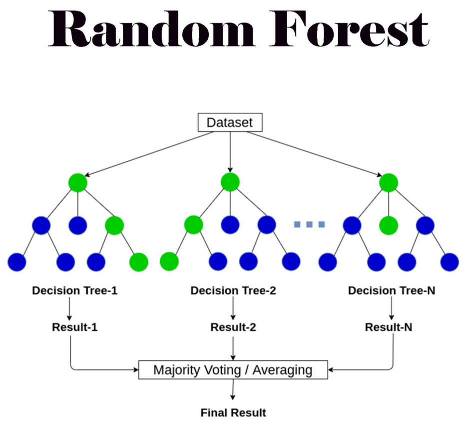


Machine Learning



1. Used stratified sampling to create a balanced sample of the data.
2. Scaled numerical features
3. PCA, regularization, and feature importance analysis
4. Train and test split
5. Models:
   - Logistic Regression → Area Under ROC Curve (AUC) score: **0.612177**
   - Random Forest Classifier → AUC: **0.712067**
6. Cross-Validation scores:
   - Logistic Regression → Area Under ROC Curve (AUC) score: **0.608112 ± 0.007**
   - Random Forest Classifier → AUC: **0.705041 ± 0.005**

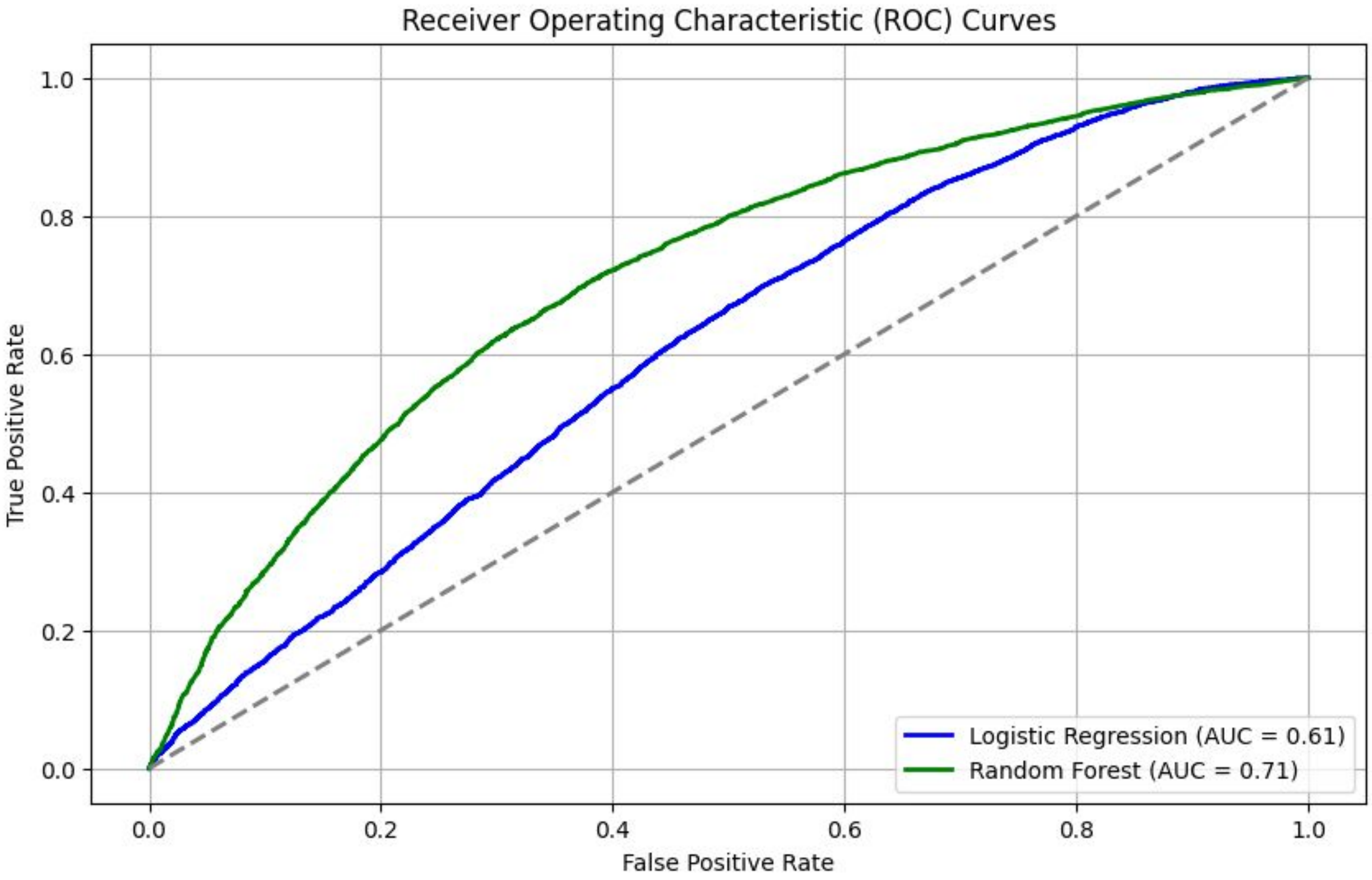




Selected Features

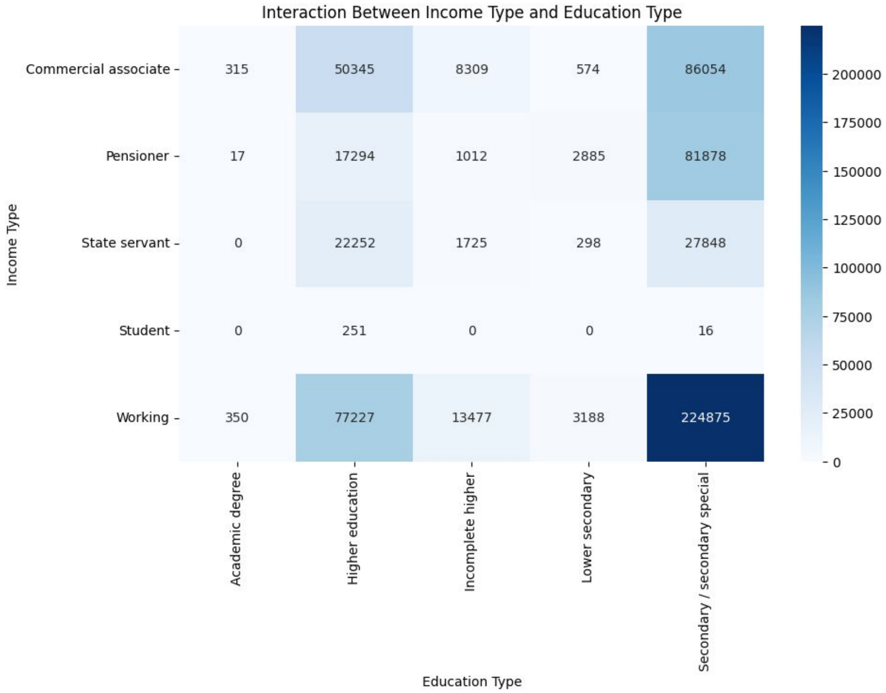
- 'CNT_FAM_MEMBERS',
- 'DAYS_EMPLOYED',
- 'AGE',
- 'MONTHS_BALANCE' - record month
- 'CNT_CHILDREN' - number of children
- 'OCCUPATION_TYPE'
- 'FLAG_PHONE' - if theres a phone

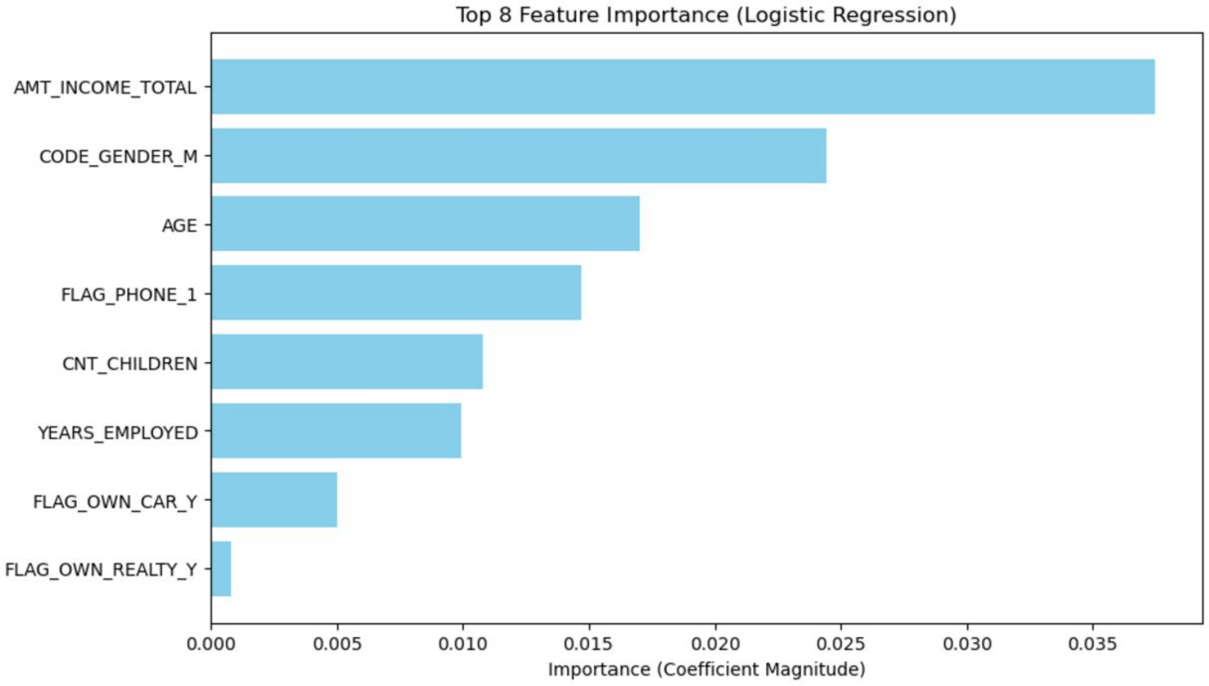
ROC Curves


Receiver Operating Characteristic (ROC) Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.61)
Random Forest (AUC = 0.71)

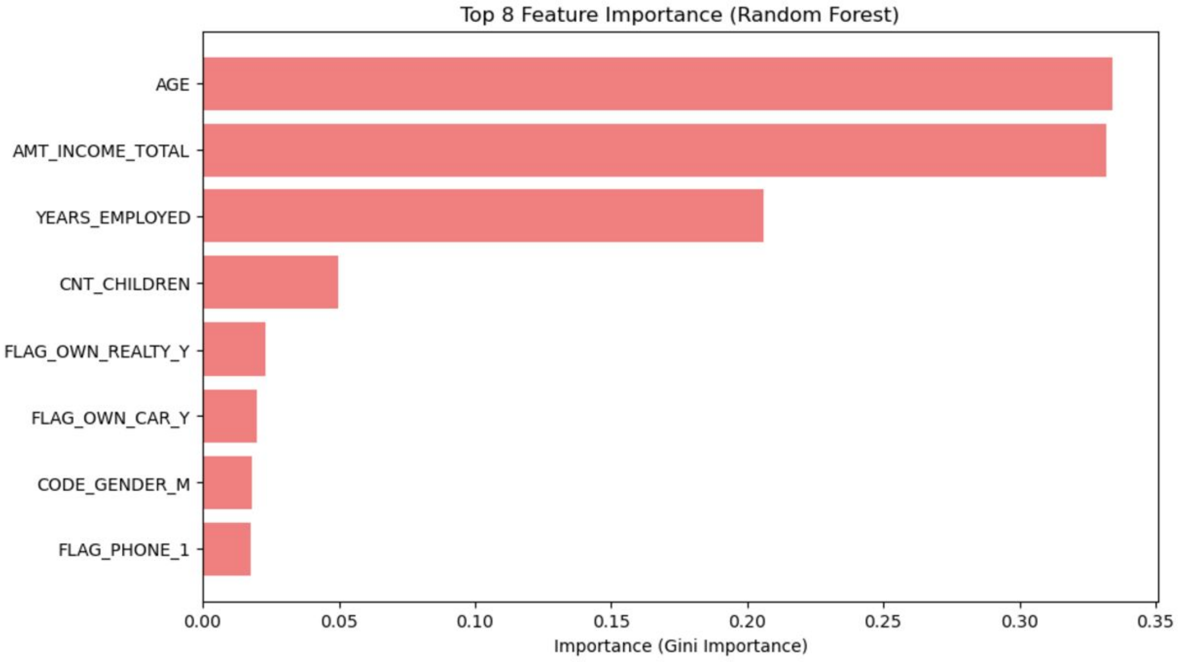
Challenges

- Handled missing data in occupation type
- Class imbalance (approved vs not approved)
  - stratified sampling
- Moderate predictive power
- Feature selection/engineering (interaction between variables)





Top 8 Feature Importance (Logistic Regression)
AMT_INCOME_TOTAL
CODE_GENDER_M
AGE
FLAG_PHONE_1
CNT_CHILDREN
YEARS_EMPLOYED
FLAG_OWN_CAR_Y
FLAG_OWN_REALTY_Y
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Importance (Coefficient Magnitude)



Top 8 Feature Importance (Random Forest)
AGE
AMT_INCOME_TOTAL
YEARS_EMPLOYED
CNT_CHILDREN
FLAG_OWN_REALTY_Y
FLAG_OWN_CAR_Y
CODE_GENDER_M
FLAG_PHONE_1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Importance (Gini Importance)

Conclusion

- Data Quality and Preprocessing
- Demographic Insights
- Approval Rate Analysis
- Modeling
- Bias and Fairness
- Considering more advanced methods

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON