Predicting response to a Bank telemarketing campaign

Jillian Etheredge

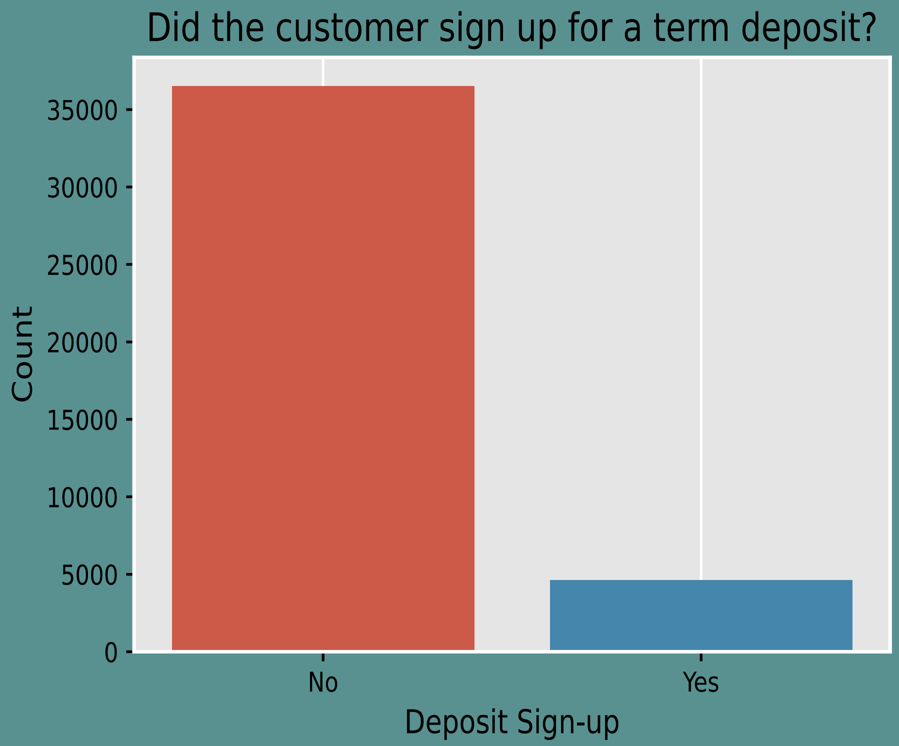
Goal of the model:

Predicting which potential customers will sign up for a term deposit account

- Maximize time efficiency
- Reach more people likely to sign up
- Increase campaign success rates

About the Data:

- UCI's Bank Marketing Data
  - Portugese Bank
  - May 2008- November 2010
- Types of Information in the Data:
  - Client Information
  - Marketing Campaign Data
    - Previous
    - Current
  - Social/Economic Context Data

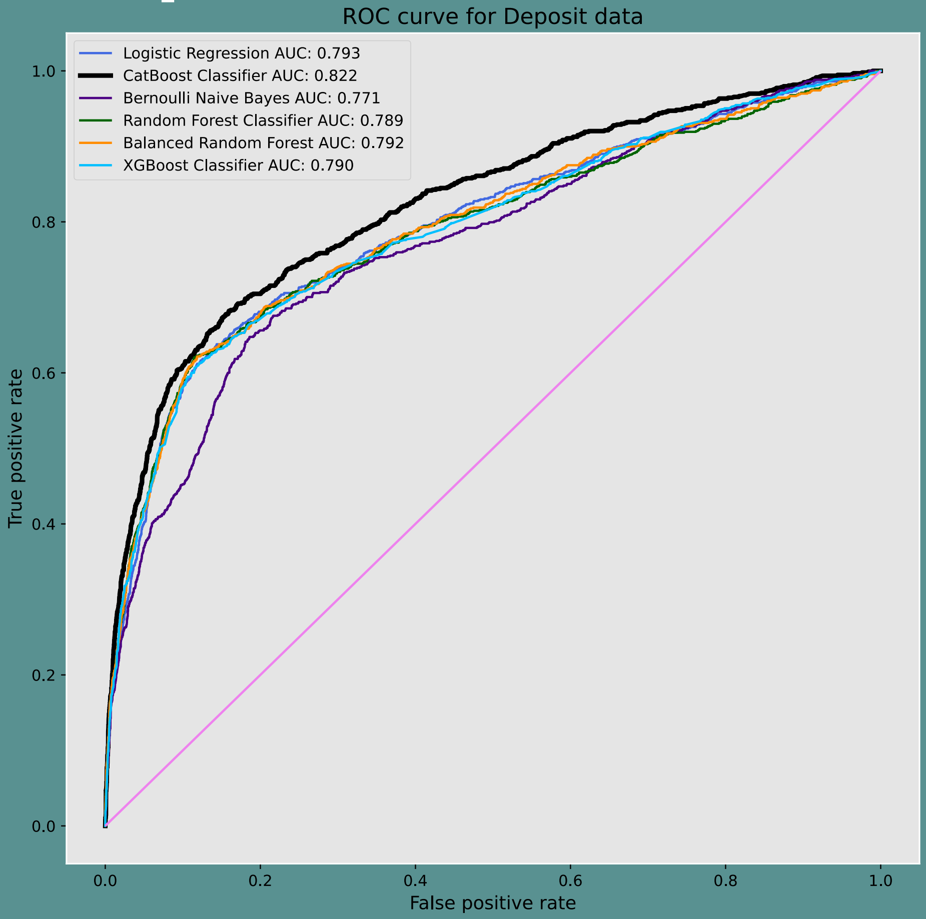


Interesting Demographic Information

Noticeable trends with signing up:

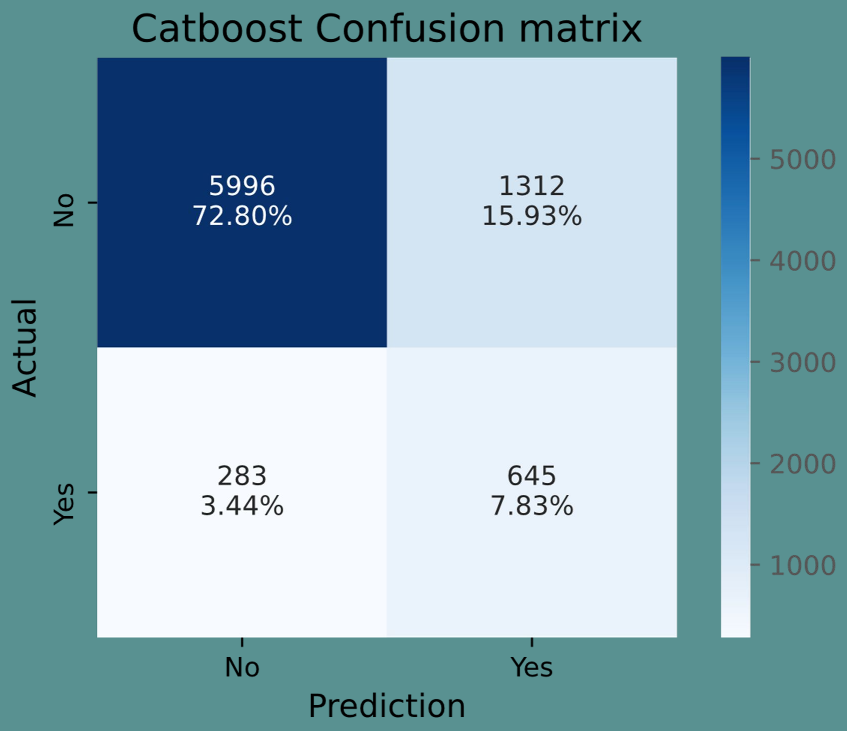
- **31%** of students and **25%** of retired customers
- **16%** of young adults and **40%** of seniors
- **15%** of cell phone users vs **5%** of landline users
- None with >23 contacts

Comparison of Model Performance



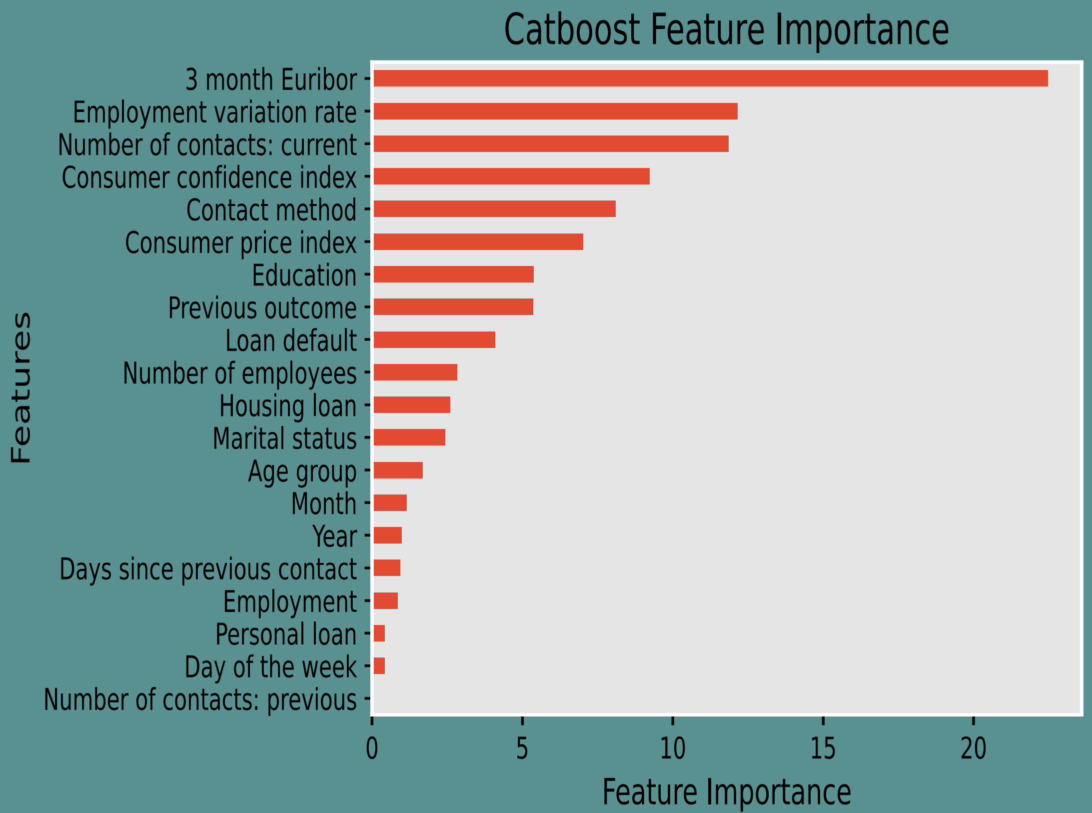
Final Model

- Catboost Classifier
- ROC AUC: **0.822**
- F1 Macro: **0.66**
- Accuracy: **0.81**



Insights and Recommendations

- Features of highest importance are mostly social/economic context features
- Campaign contacts and education were important features for multiple models



Future Improvements

- Add more economic data
- More tuning with Catboost
- Ensemble models
- Create flask app

Questions?

Appendix

Features

Client Data

- Age
- Job
- Marital Status
- Education
- Loan Default
- Housing Loan
- Personal Loan
- Employment *
- Age Group *

Campaign Data

- Contact Method
- Last contact Month
- Last contact Weekday
- Duration of call
- Number of contacts
- Days since last contact from previous campaign
- Number of contacts from previous campaign
- Outcome of previous campaign
- Contact year *
- Contact season *

Social/Economic Data

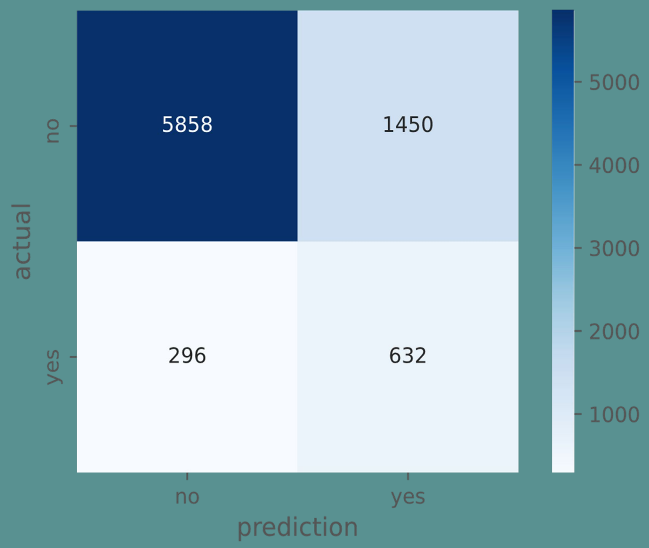
- Employment variation rate
- Consumer price index
- Consumer confidence index
- 3 month EURO InterBank Offer Rate
- Number of employees

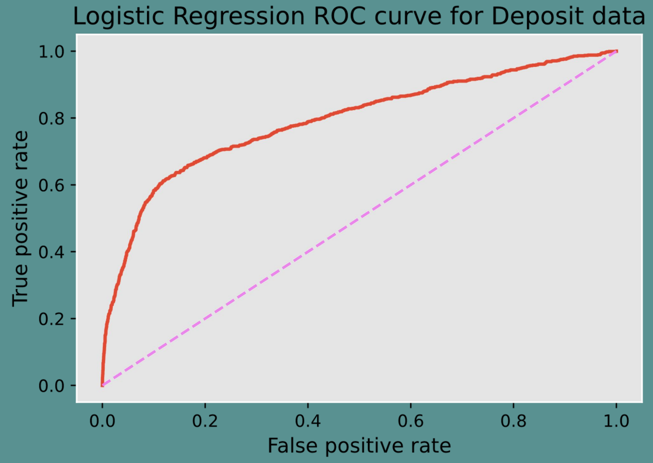
* Engineered Feature

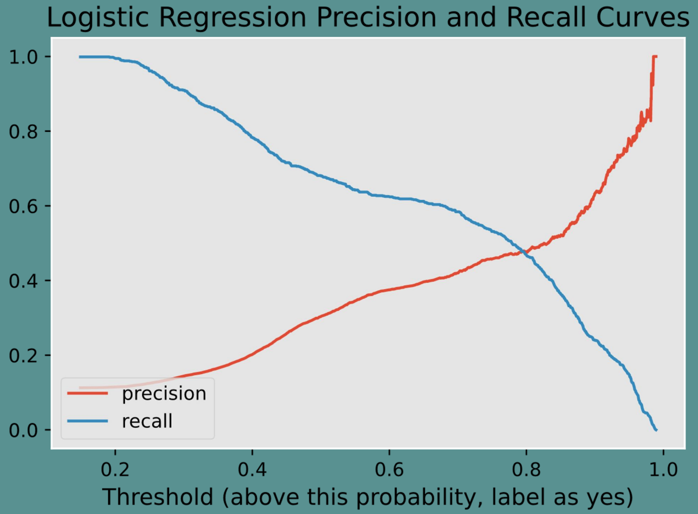
Model	ROC AUC	F1 Macro Threshold=0.5	Accuracy
Dummy (Baseline)	0.500	0.47	0.89
Logistic Regression	0.793	0.65	0.79
Catboost	0.822	0.66	0.81
Bernoulli Naive Bayes	0.771	0.63	0.79
Random Forest	0.789	0.68	0.84
Balanced Random Forest	0.792	0.66	0.81
Xgboost	0.790	0.67	0.83

Logistic Regression

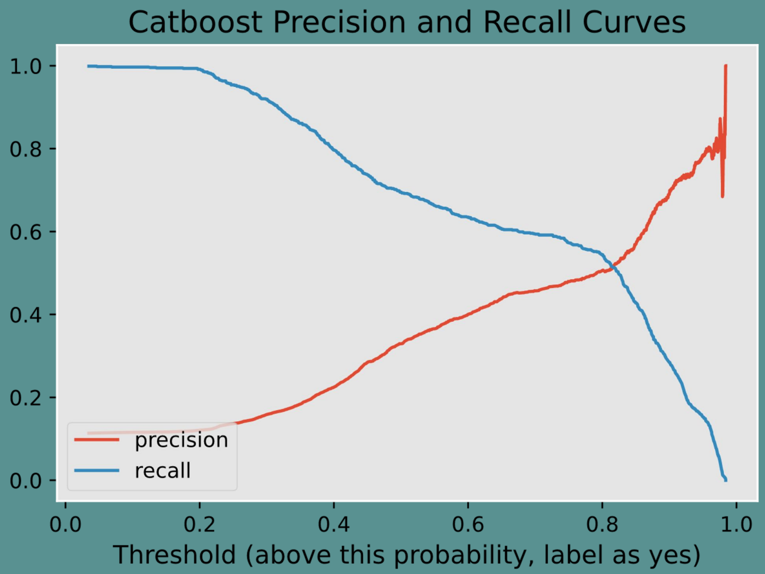
Logistic Regression Confusion Matrix

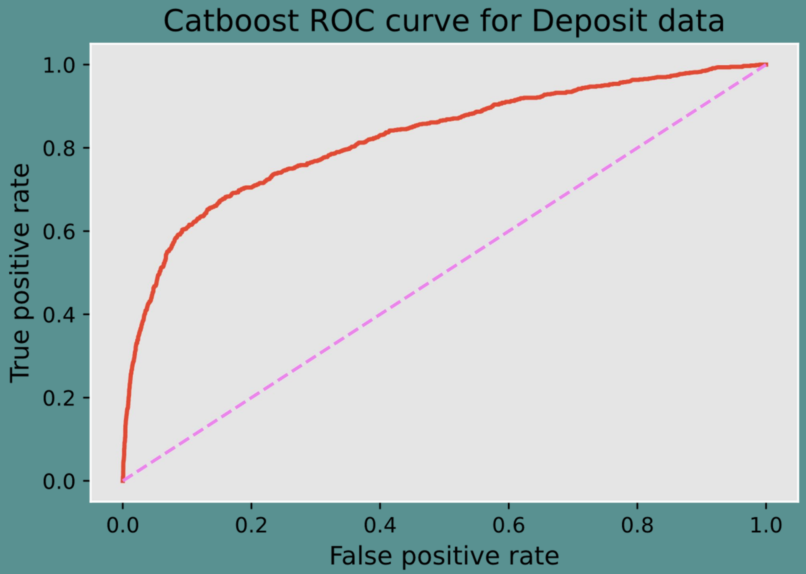




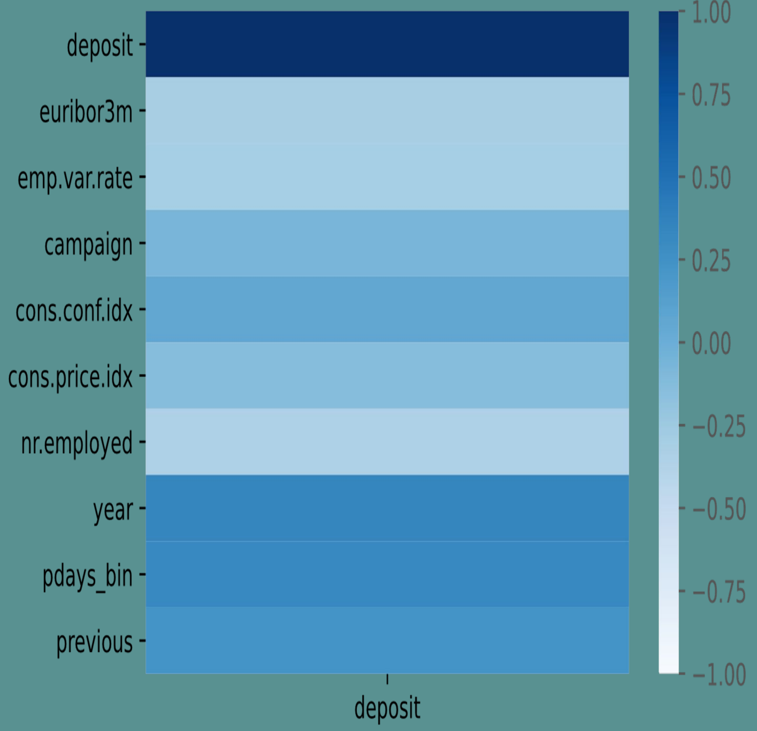


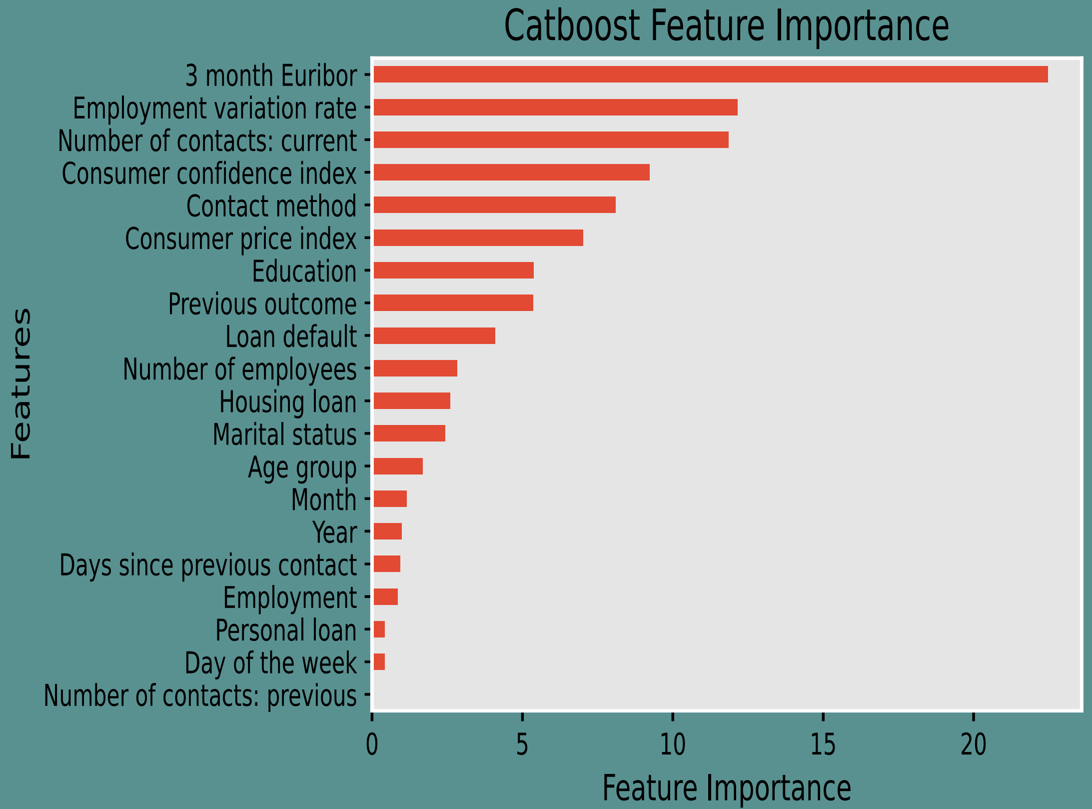
Catboost Classifier



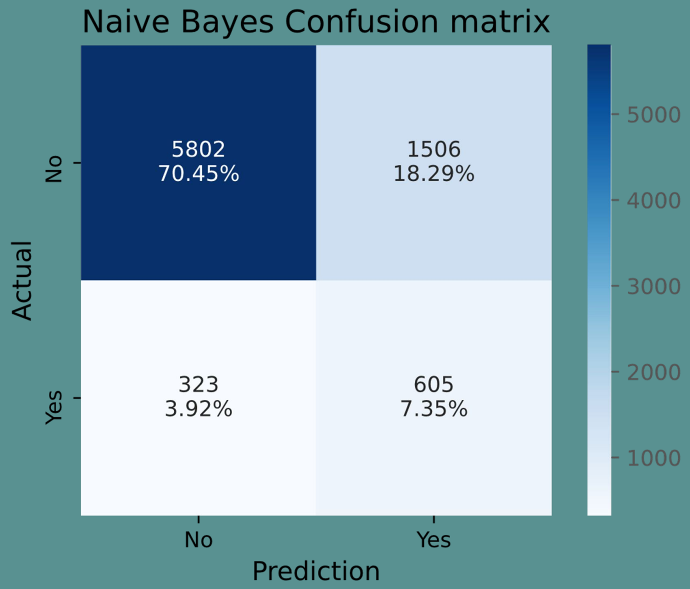


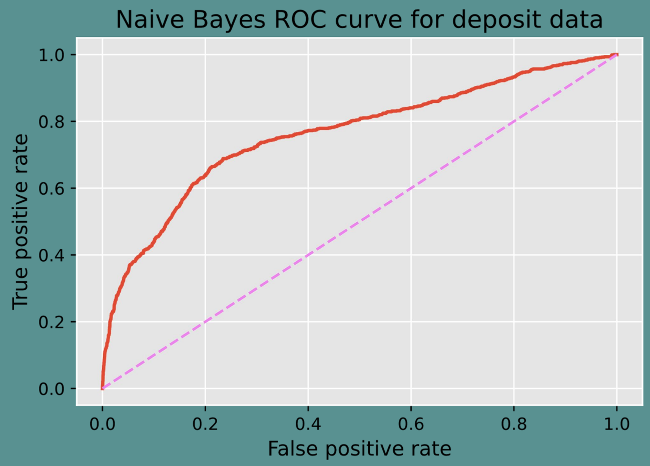
Catboost Classifier

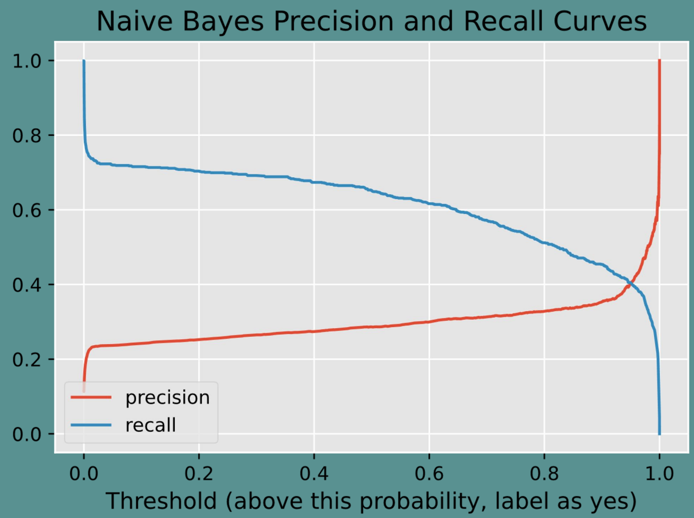




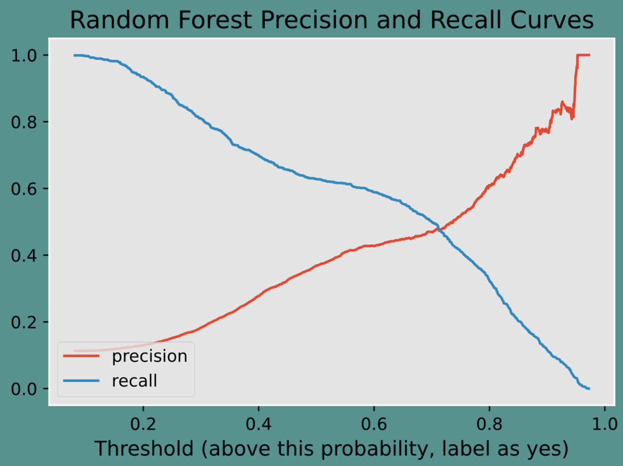
Bernoulli Naïve Bayes

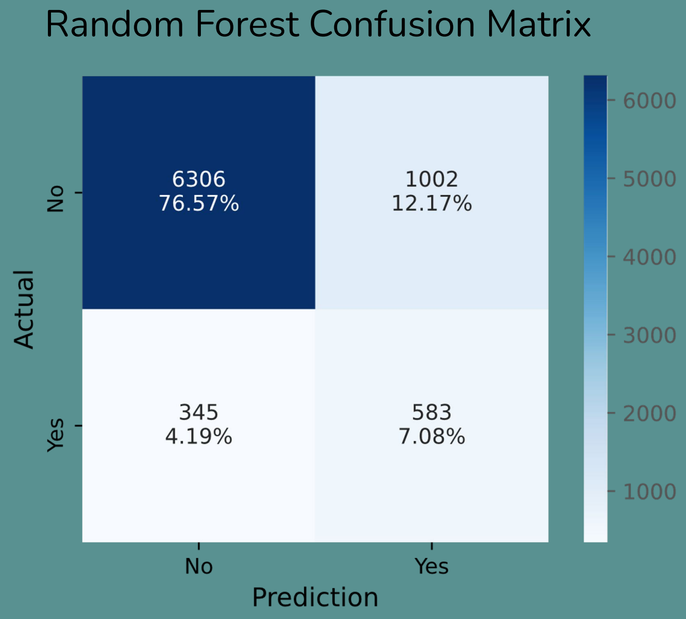


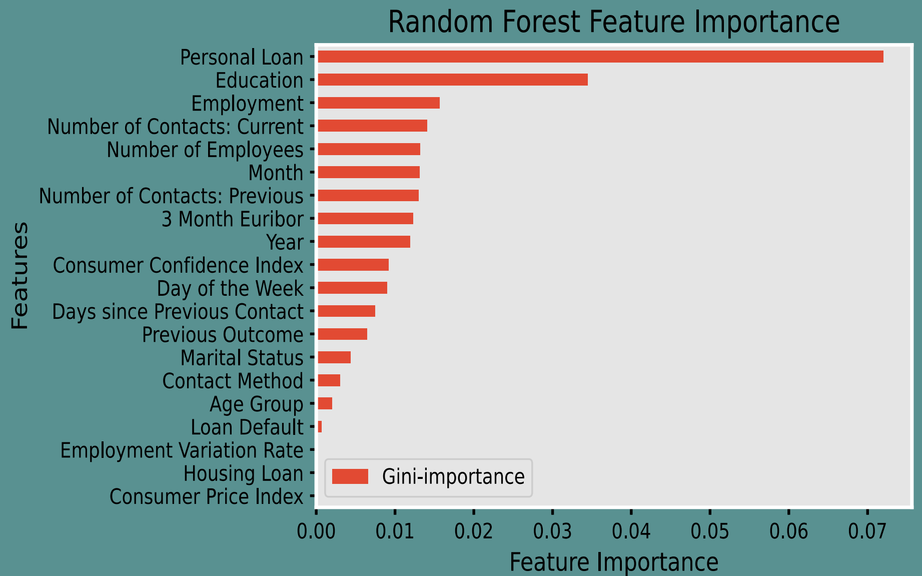




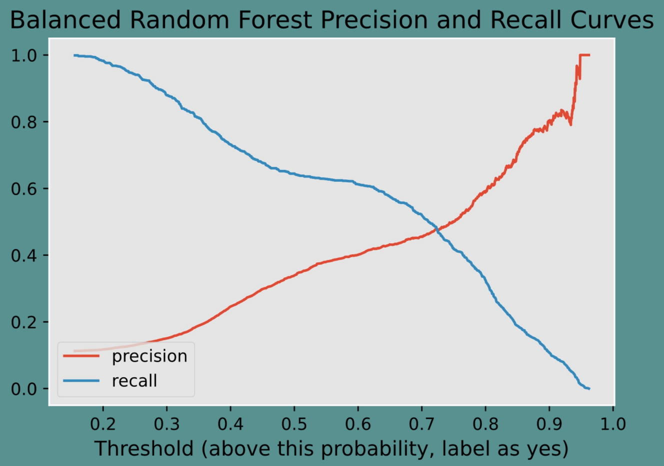
Random Forest Classifier

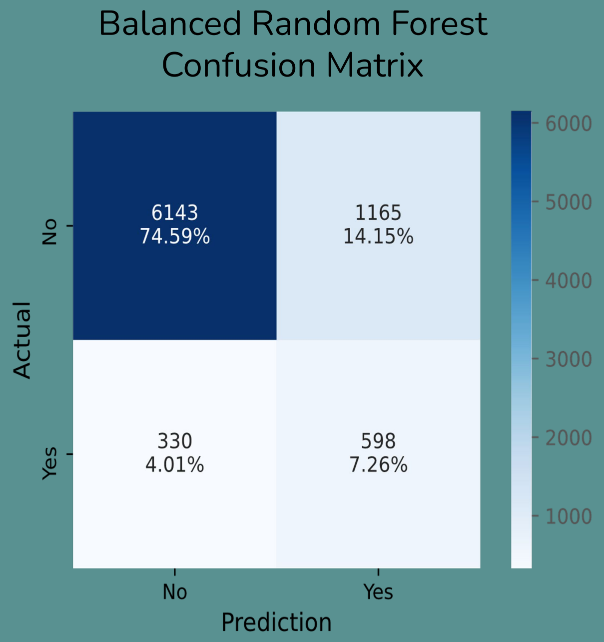


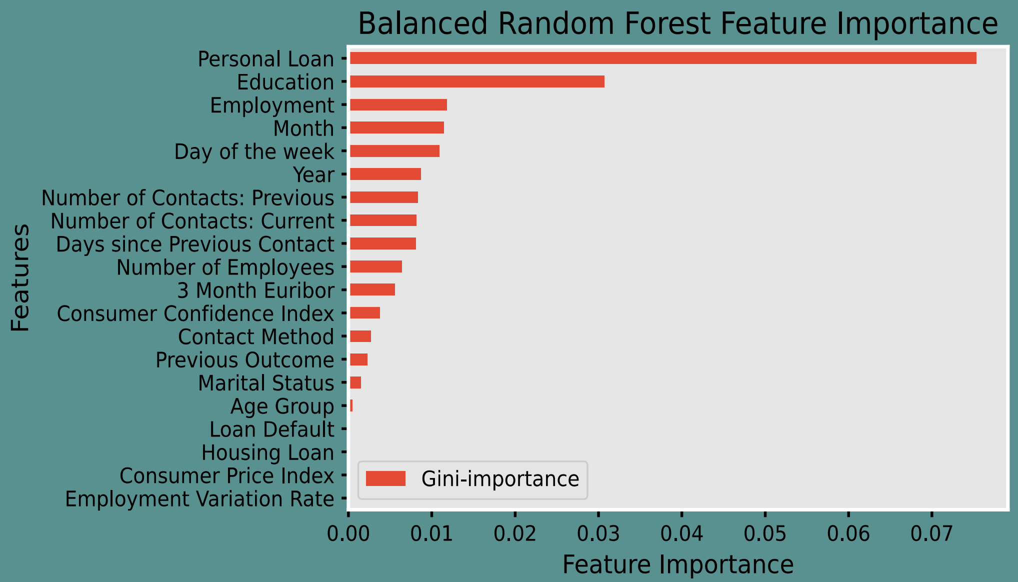




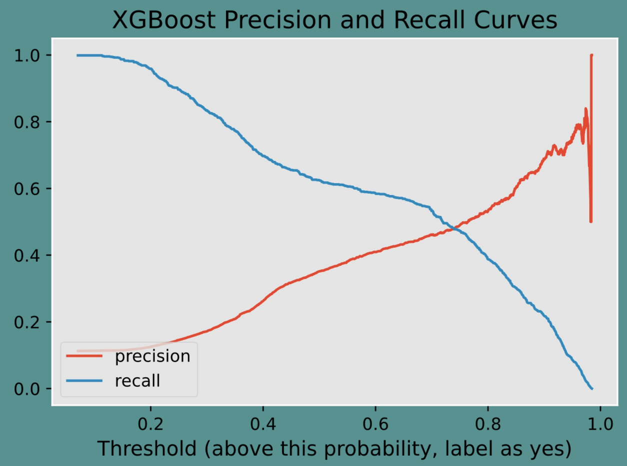
Balanced Random Forest



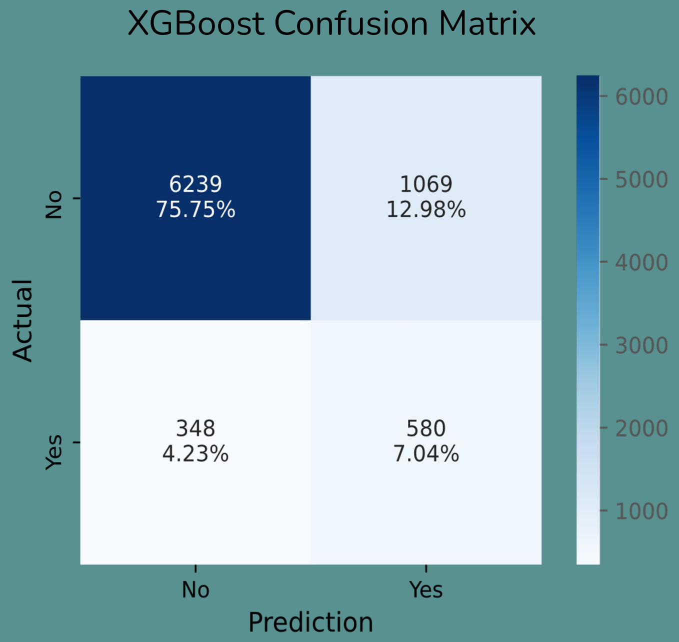




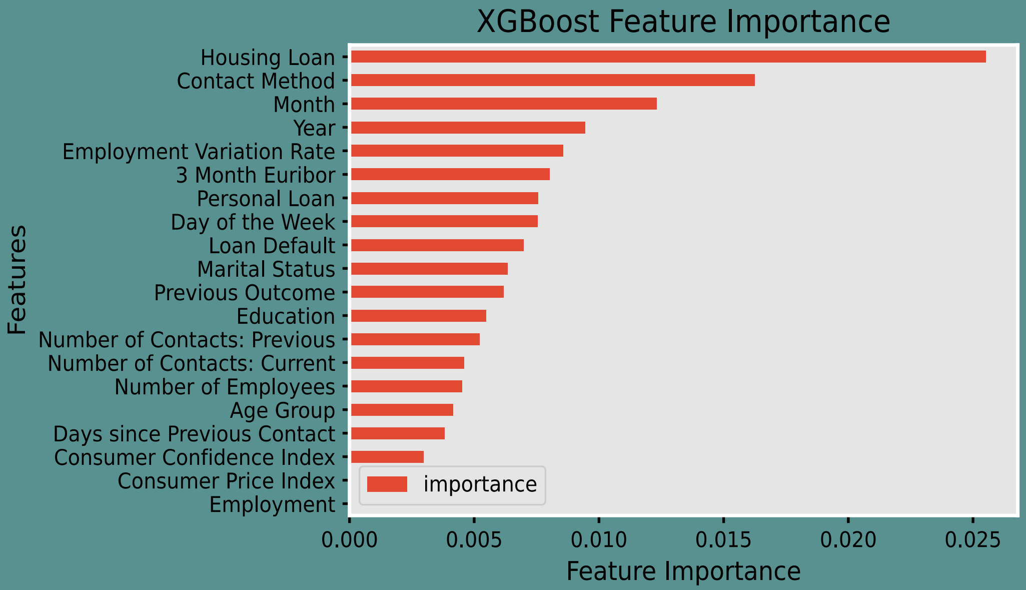
XGBoost



XGBoost Precision and Recall Curves
1.0
0.8
0.6
0.4
0.2
0.0
precision
recall
0.2
0.4
0.6
0.8
1.0
Threshold (above this probability, label as yes)



XGBoost Confusion Matrix
6239
75.75%
1069
12.98%
348
4.23%
580
7.04%
No
Yes
Actual
No
Yes
Prediction
6000
5000
4000
3000
2000
1000



XGBoost Feature Importance
Housing Loan
Contact Method
Month
Year
Employment Variation Rate
3 Month Euribor
Personal Loan
Day of the Week
Loan Default
Marital Status
Previous Outcome
Education
Number of Contacts: Previous
Number of Contacts: Current
Number of Employees
Age Group
Days since Previous Contact
Consumer Confidence Index
Consumer Price Index
Employment
Features
importance
0.000
0.005
0.010
0.015
0.020
0.025
Feature Importance