Bank Telemarketing Success
Classification problem

Rui Yuan

Clients and Problem

- Bank wants to implement a telemarketing campaign.
- And they want to know about the campaign performance: success rate; what their target clients are (that are likely to subscribe bank term deposit or other financial products).

Data

- The data is download from UCI ML Repository, related with direct marketing campaigns of a Portuguese banking institution dated from May 2008 to November 2010.
- The marketing campaigns were based on phone calls.
- 45,211 rows and 16 columns.
- Target: Whether client subscribe bank term deposit (1 = yes, 0 = no)

15 Features

bank clients' data:

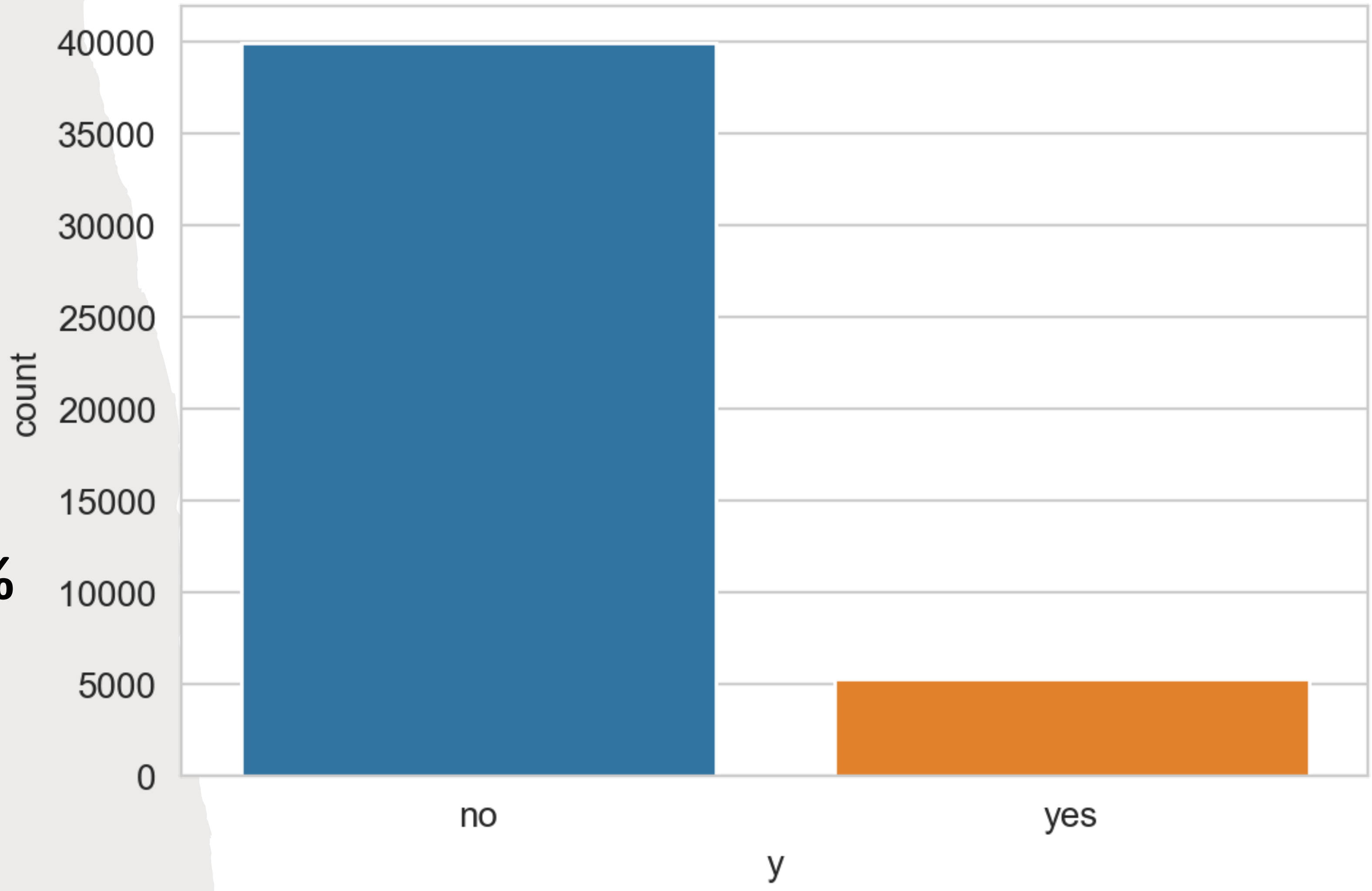
- 1 - **age**: numeric, 18 to 95.
- 2 - **job** : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
- 3 - **marital status** (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
- 4 - **education** (categorical: "unknown","secondary","primary","tertiary")
- 5 - **default**: has credit in default? (binary: "yes","no")
- 6 - **balance**: average yearly balance, in euros (numeric)
- 7 - **housing**: has housing loan? (binary: "yes","no")
- 8 - **loan**: has personal loan? (binary: "yes","no")

campaign data and other attributes:

- 9 - **contact**: contact communication type (categorical: "cellular","telephone", "unknown")
- 10 - **day**: last contact day of the month
- 11 - **month**: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
- 12 - **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 13 - **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- 14 - **previous**: number of contacts performed before this campaign and for this client (numeric)
- 15 - **poutcome**: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")

Class Distribution

Success rate= yes/all = 12%



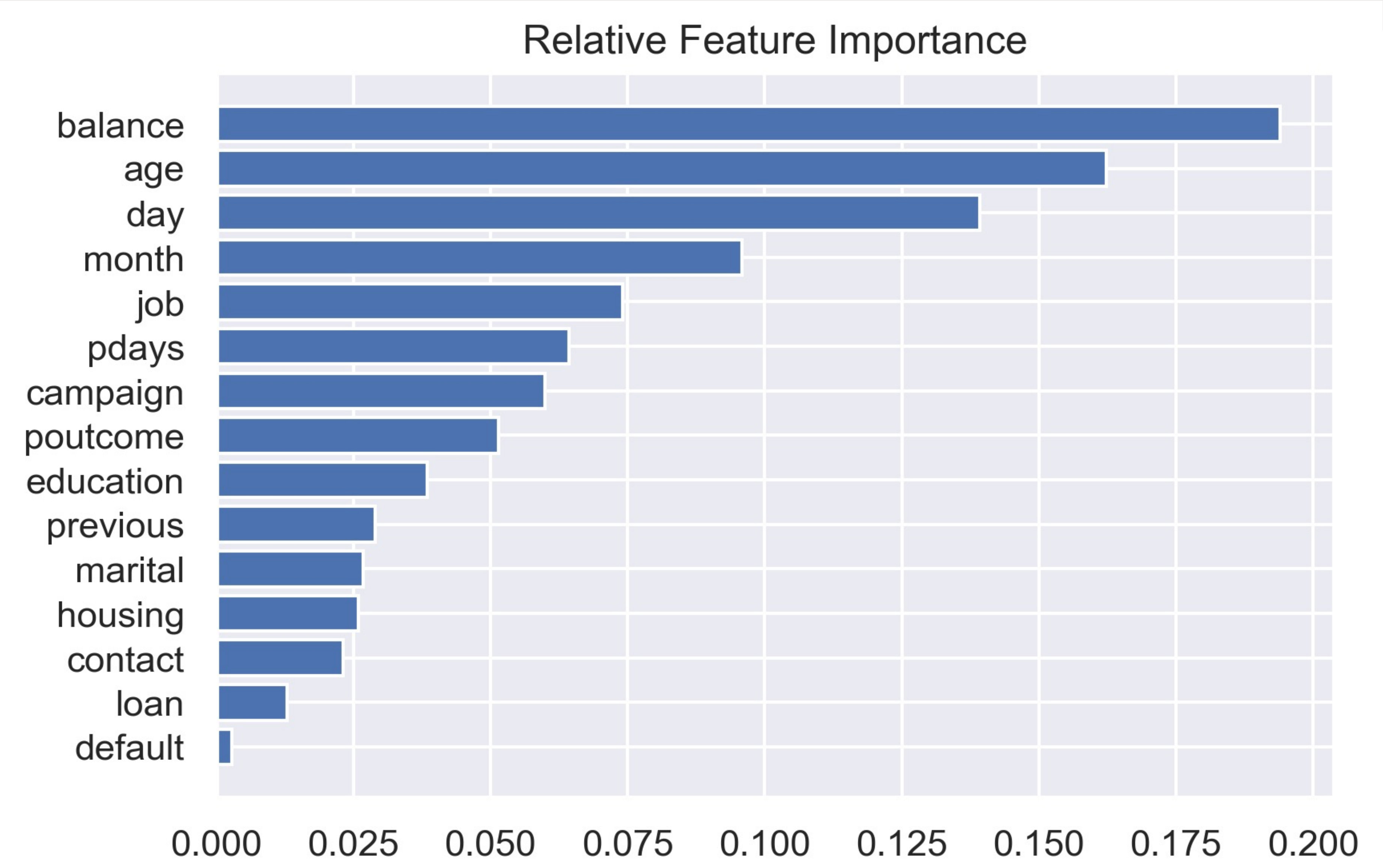
Classification Modeling Goal

- Goal: ↑ precision score on positive class & number of true positives (subscribe)
- In Business Sense: Model being able to target clients upon changing needs.
  - Business capability can vary depending on the amount of sources, such as number of employees, phone plan fees, etc.
  - So how many clients bank can reach to or how many phone calls bank can do in a given period (week/month/year) may vary, and we want the model being able to target potential clients upon business changing capability.

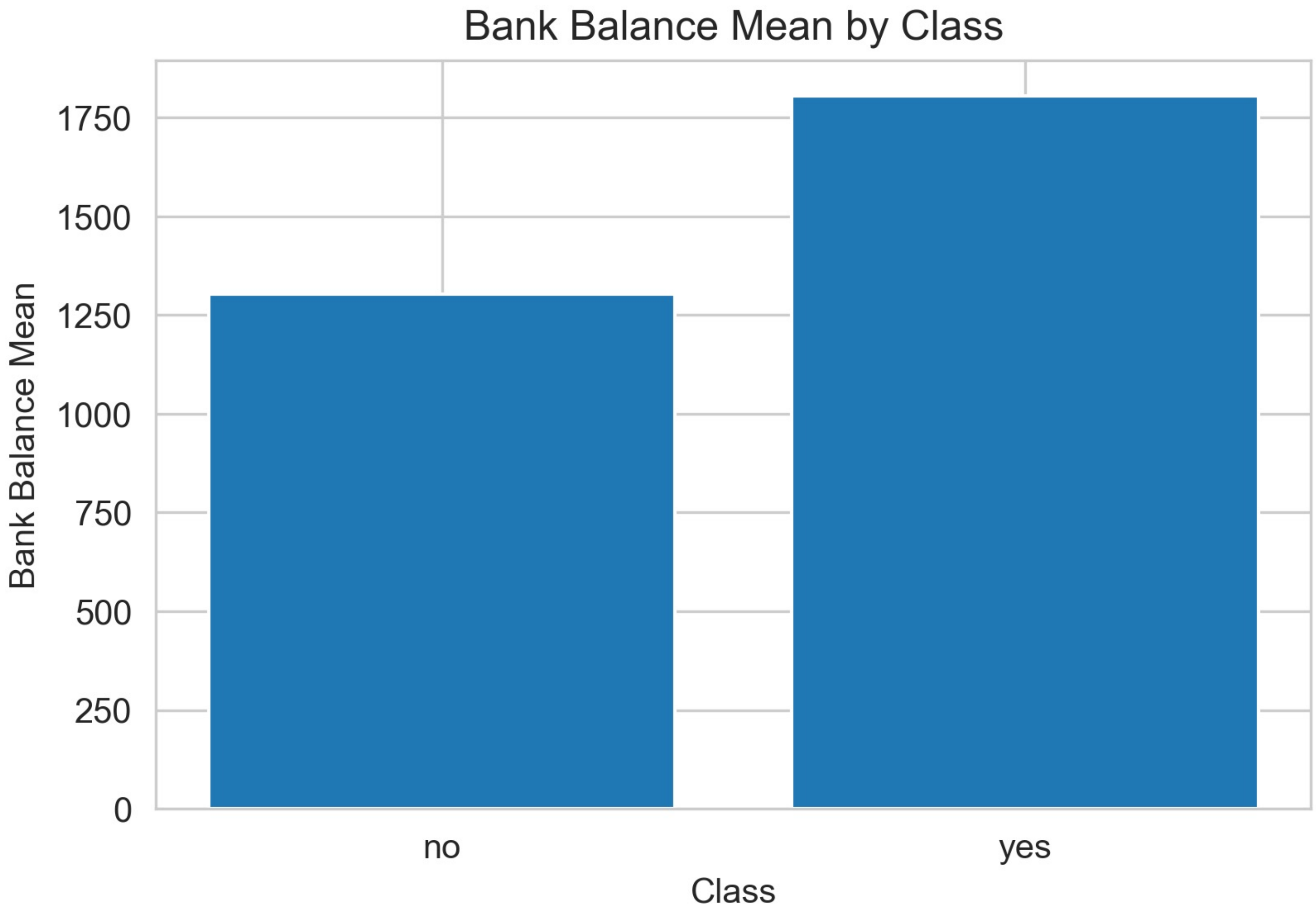
Modeling & Performance Metric Report

Phase 1: Model Testing				
Model	**Precision**	**Recall**	**F-1**	**Accuracy**
1. K-Nearest Neighbor Baseline	0.43	0.12	0.18	0.88
2. K-Nearest Neighbor Optimized with Grid Search	0.48	0.04	0.07	0.88
3. Logisic Regression Baseline	0.50	0.00	0.00	0.88
4. Logisic Regression Regularized	0.48	0.04	0.07	0.88
5. Random Forest Baseline	0.69	0.21	0.32	0.89
6. Random Forest Optimized with Random Search	**0.71**	0.19	0.30	0.89
Phase 2: Handle Class Imbalance				
Model	Precision	Recall	F-1	Accuracy
7. Random Forest with Sampling method	0.55	0.29	0.38	0.89
8. Random Forest with Adjusted Class Weight	0.69	0.19	0.30	0.89
9. Random Forest with Probability Threshold Adjustment	Adjustable	Adjustable	Adjustable	Adjustable

Feature Importance from RF model

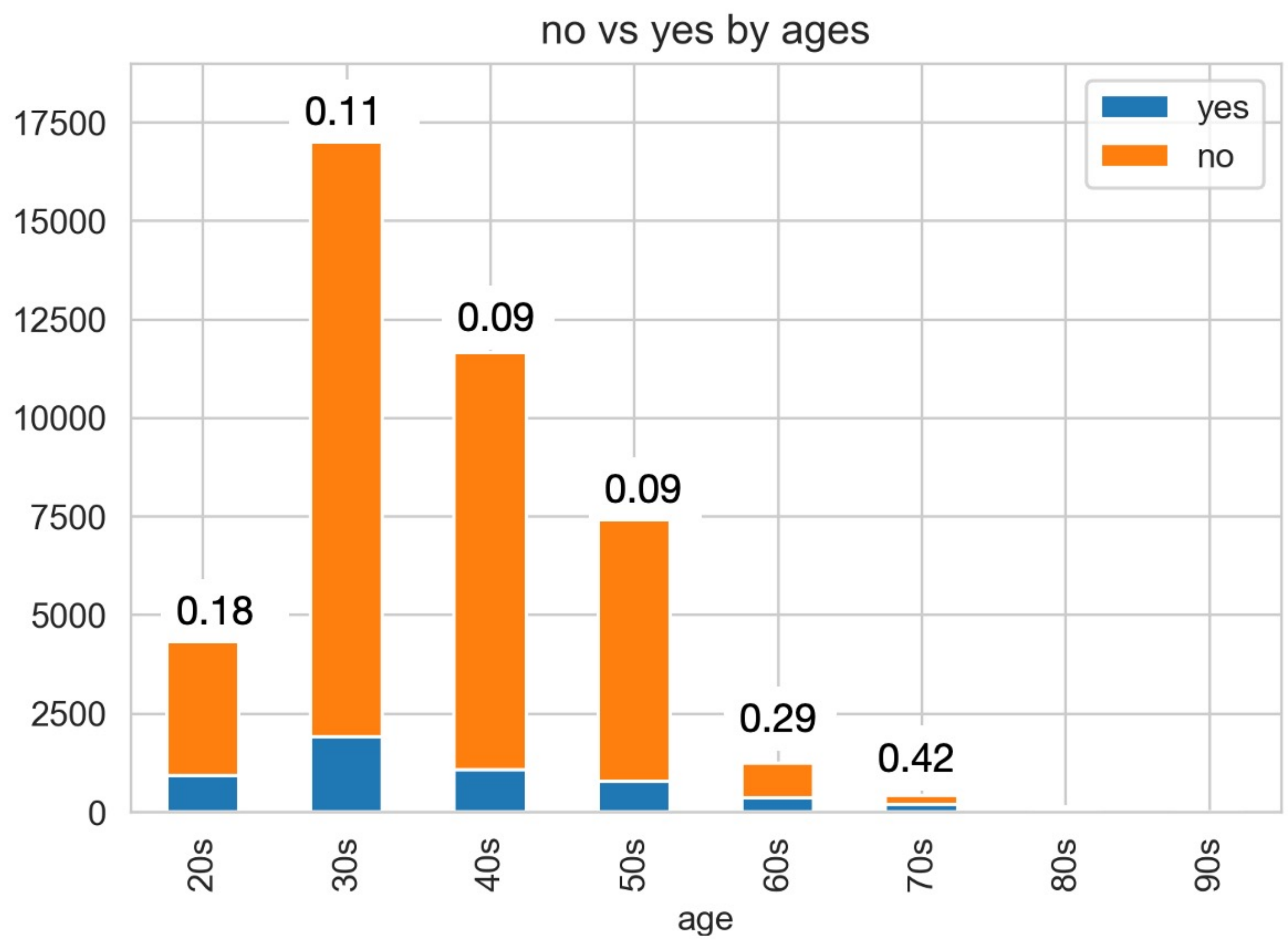


Closer look at important features: balance

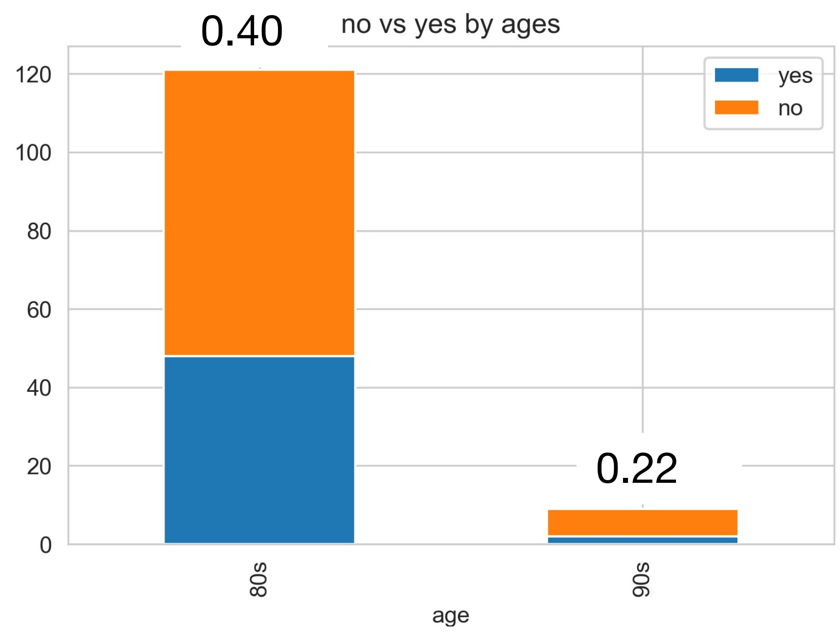


Positive class has higher average bank balance

Closer look at important features: age

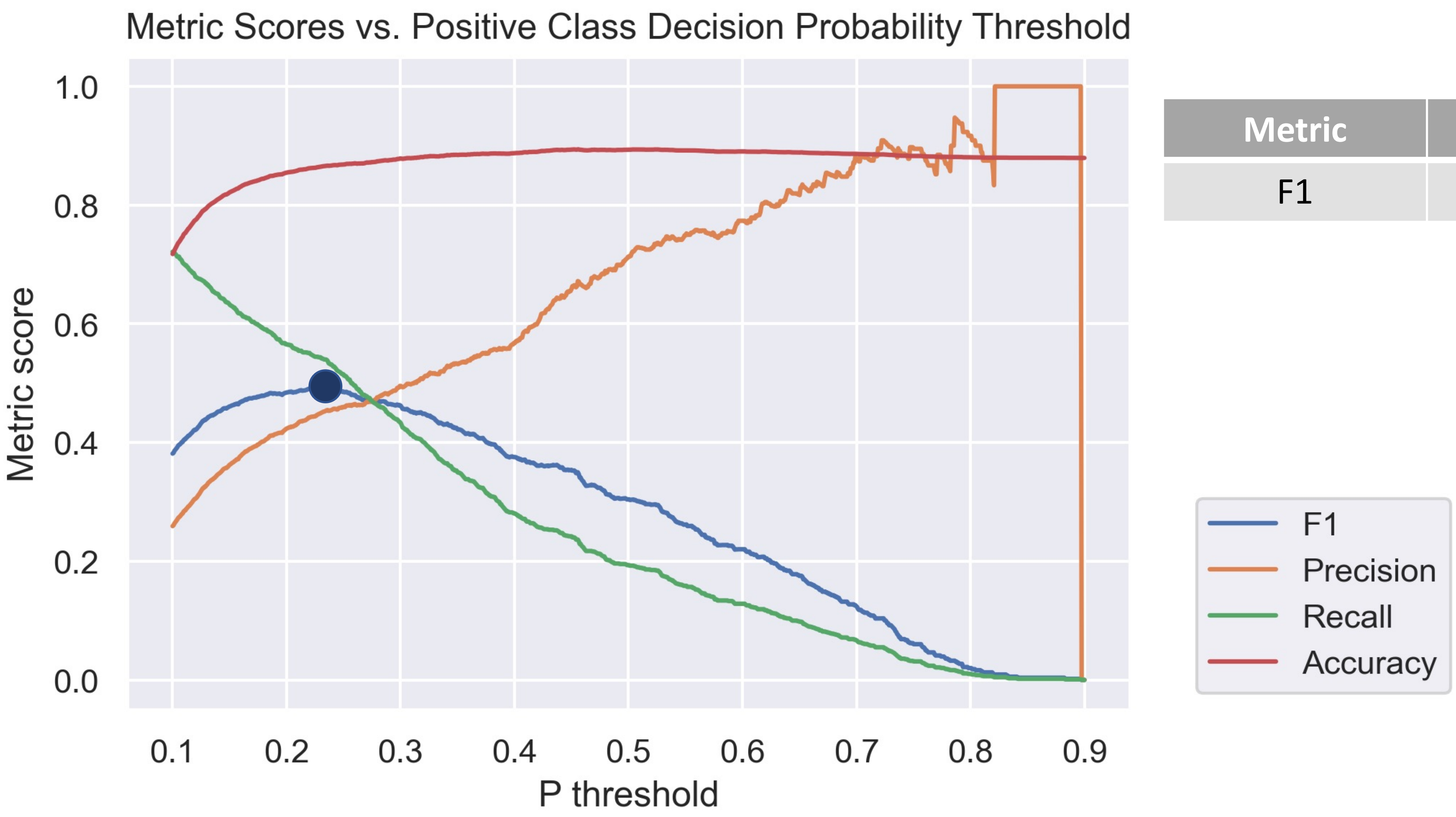


Zoom in 80s and 90s



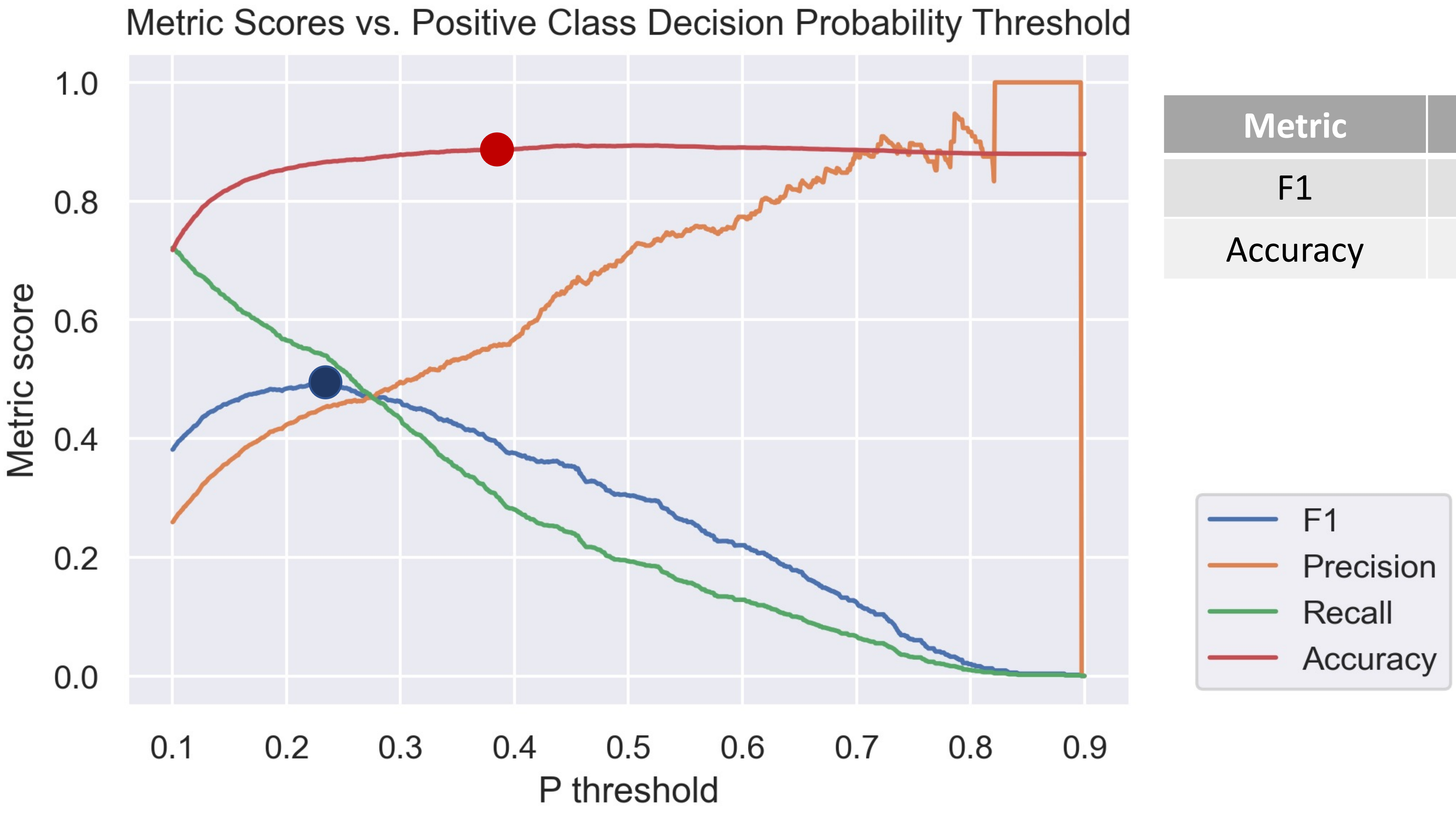
Elderly groups have higher success rate but lower number of success

Solution: RF model w/ Probability Threshold Controlling



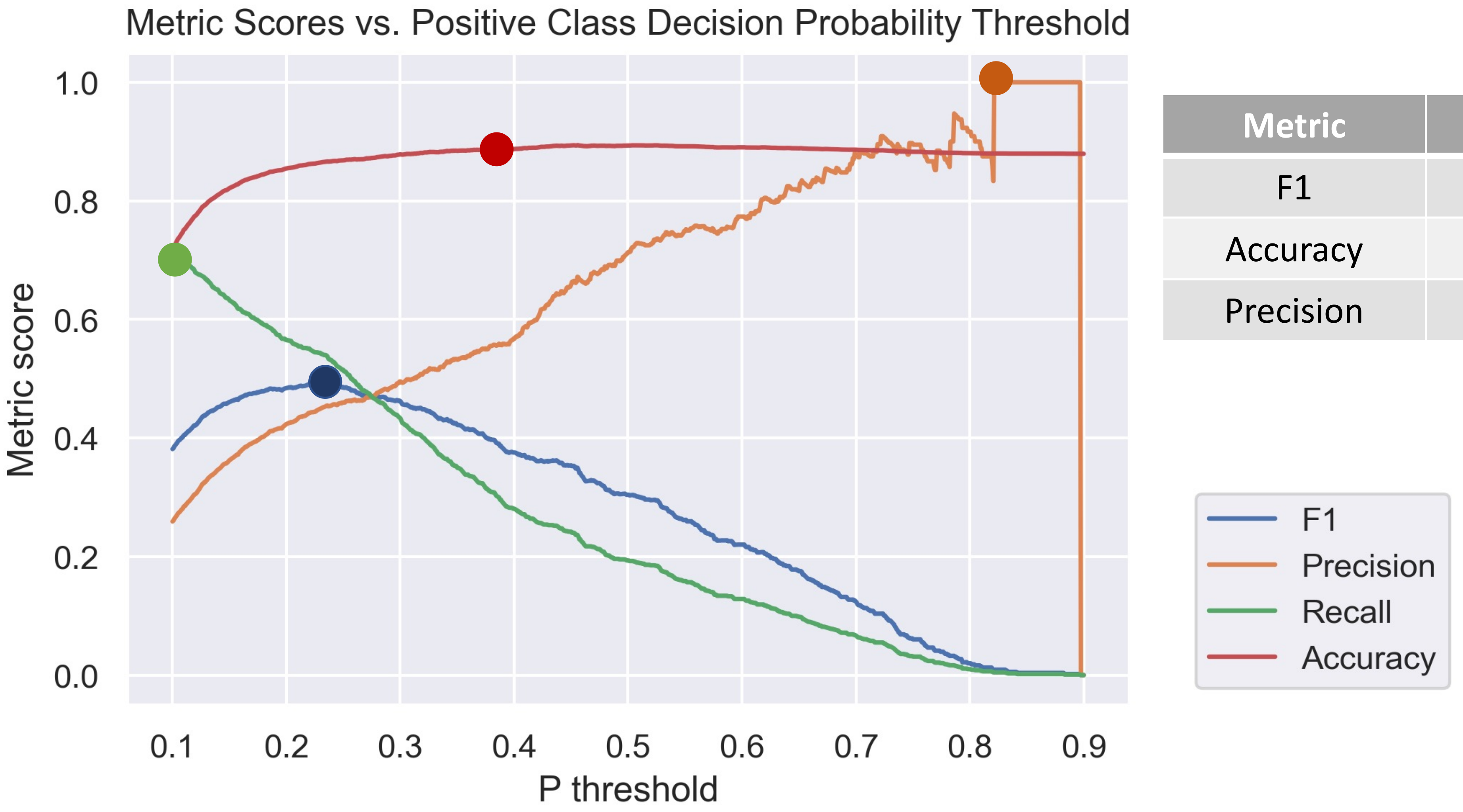
Metric	Best Score	Probablity
F1	0.47	0.24

Solution: RF model w/ Probability Threshold Controlling



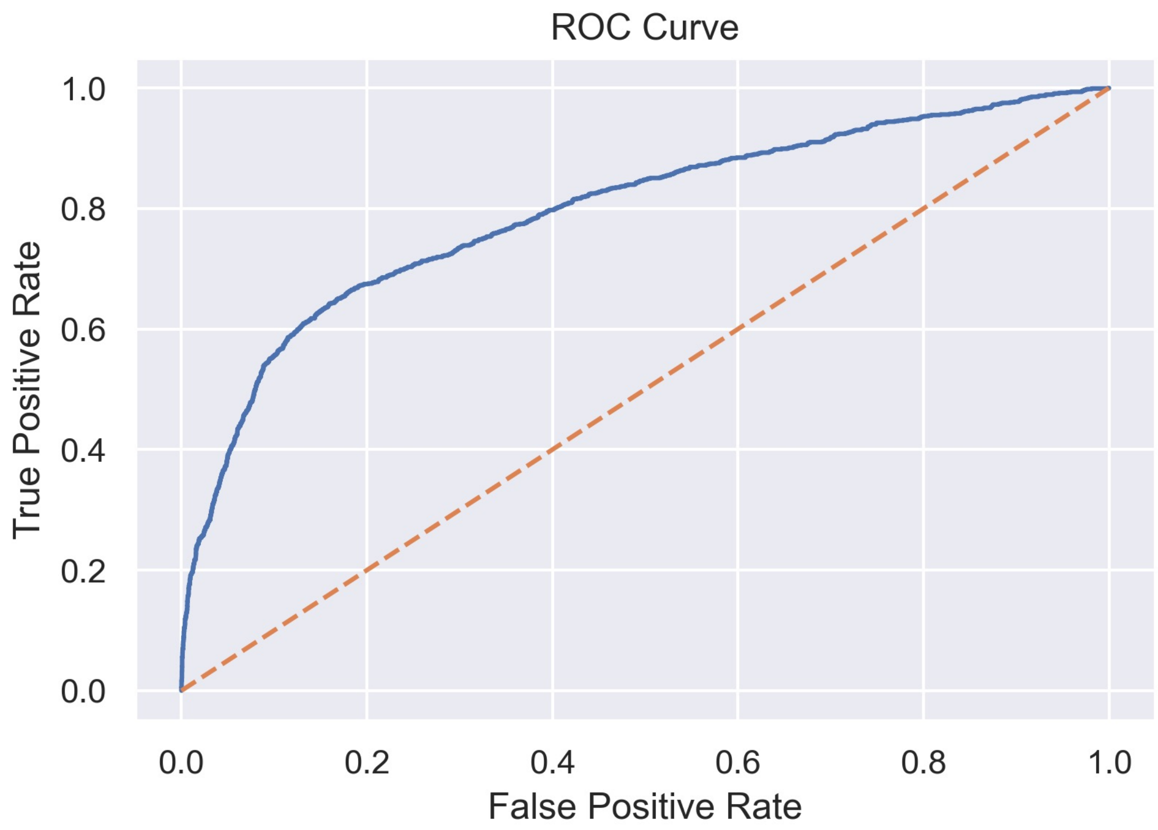
Metric	Best Score	Probablity
F1	0.47	0.24
Accuracy	0.89	0.46

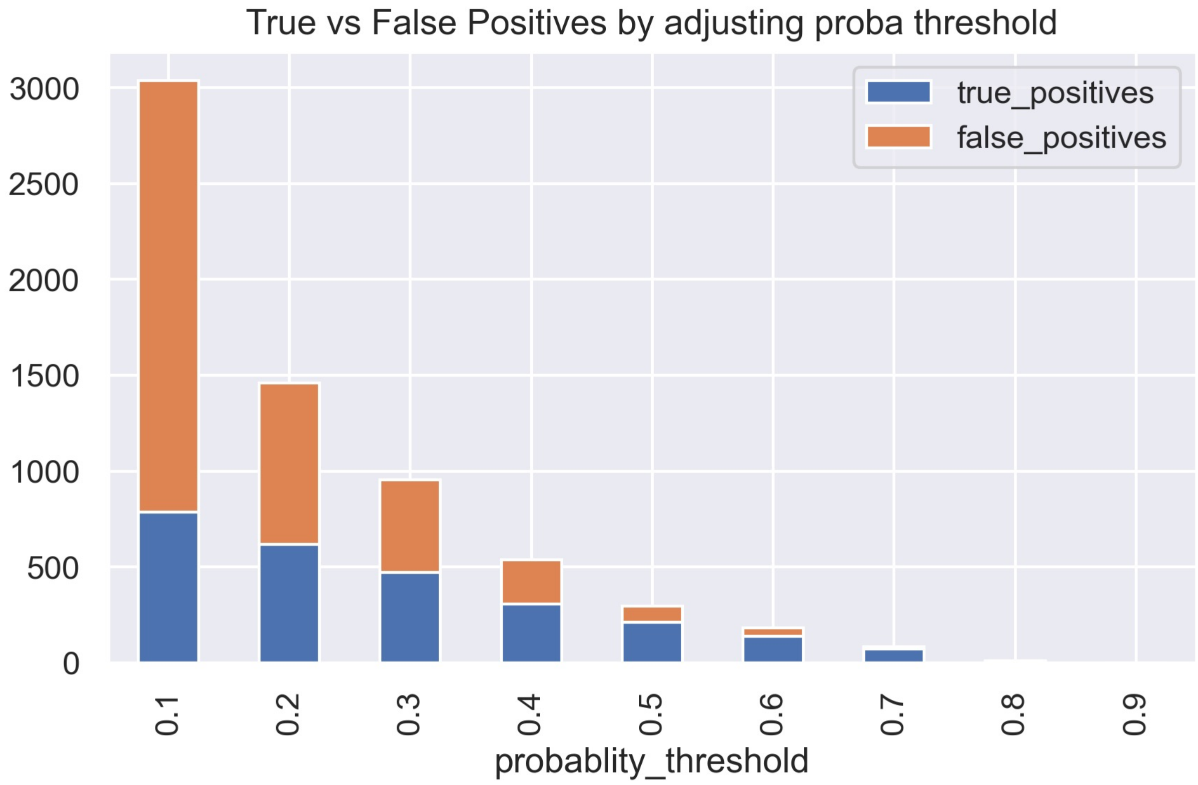
Solution: RF model w/ Probability Threshold Controlling



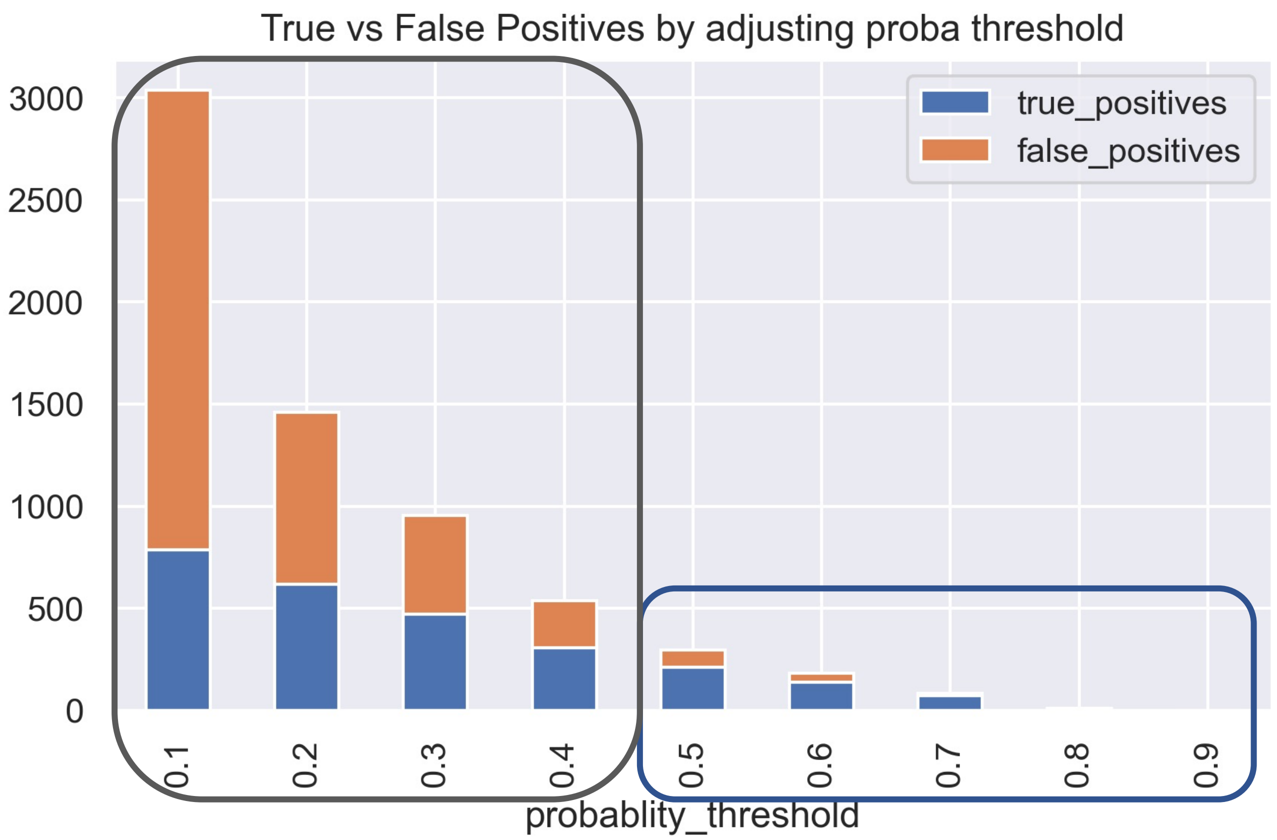
Metric	Best Score	Probablity
F1	0.47	0.24
Accuracy	0.89	0.46
Precision	1.00	0.82

Solution: RF model w/ Probability Threshold Controlling





Solution: RF model w/ Probability Threshold Controlling

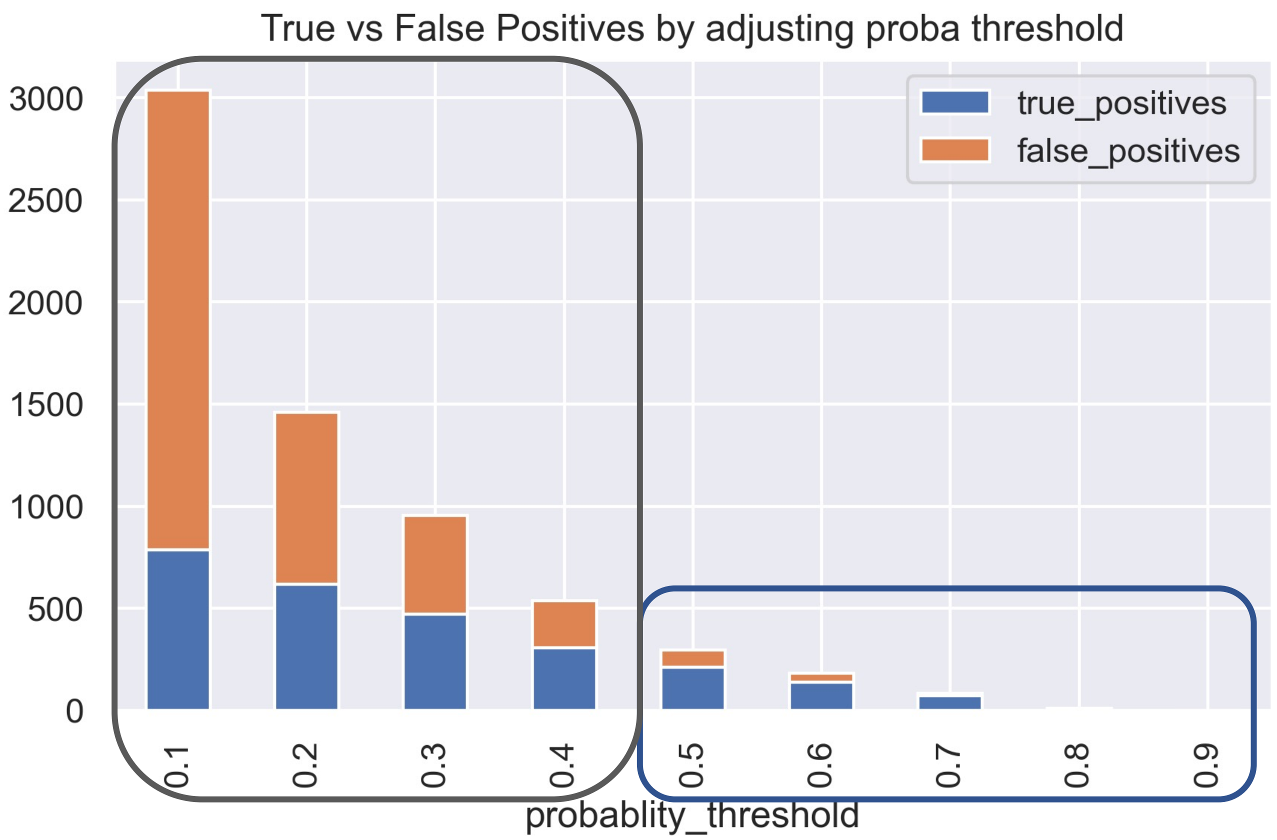


Proba threshold	True positives	False Positives	Predicted Positives	Precision Score
0.1	787	2250	3037	0.26
0.2	617	841	1458	0.46
0.3	472	481	953	0.49
0.4	306	233	539	0.56
0.5	211	85	296	0.71
0.6	140	41	181	0.77
0.7	72	10	82	0.88
0.8	11	1	12	0.92
0.9	0	0	0	

Prob threshold 0.5 to 0.9, precision score ranges from around 0.7 to 0.9.

Prob threshold 0.1 to 0.4, precision score ranges from 0.26 to 0.56.

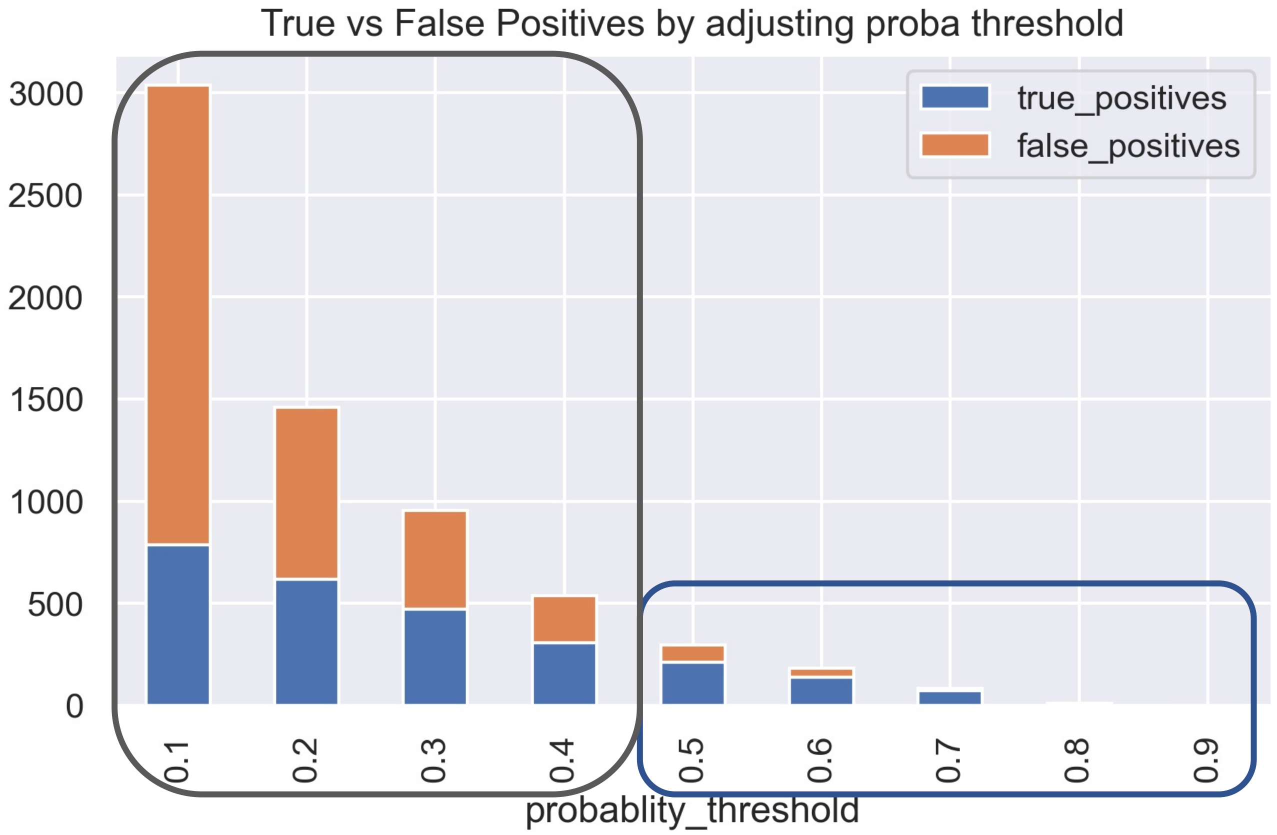
Solution: RF model w/ Probability Threshold Controlling



Proba threshold	True positives	False Positives	Predicted Positives	Precision Score
0.1	787	2250	3037	0.26
0.2	617	841	1458	0.46
0.3	472	481	953	0.49
0.4	306	233	539	0.56
0.5	211	85	296	0.71
0.6	140	41	181	0.77
0.7	72	10	82	0.88
0.8	11	1	12	0.92
0.9	0	0	0	

Prob 0.1 to 0.4 have greater true positives than prob 0.5 to 0.9

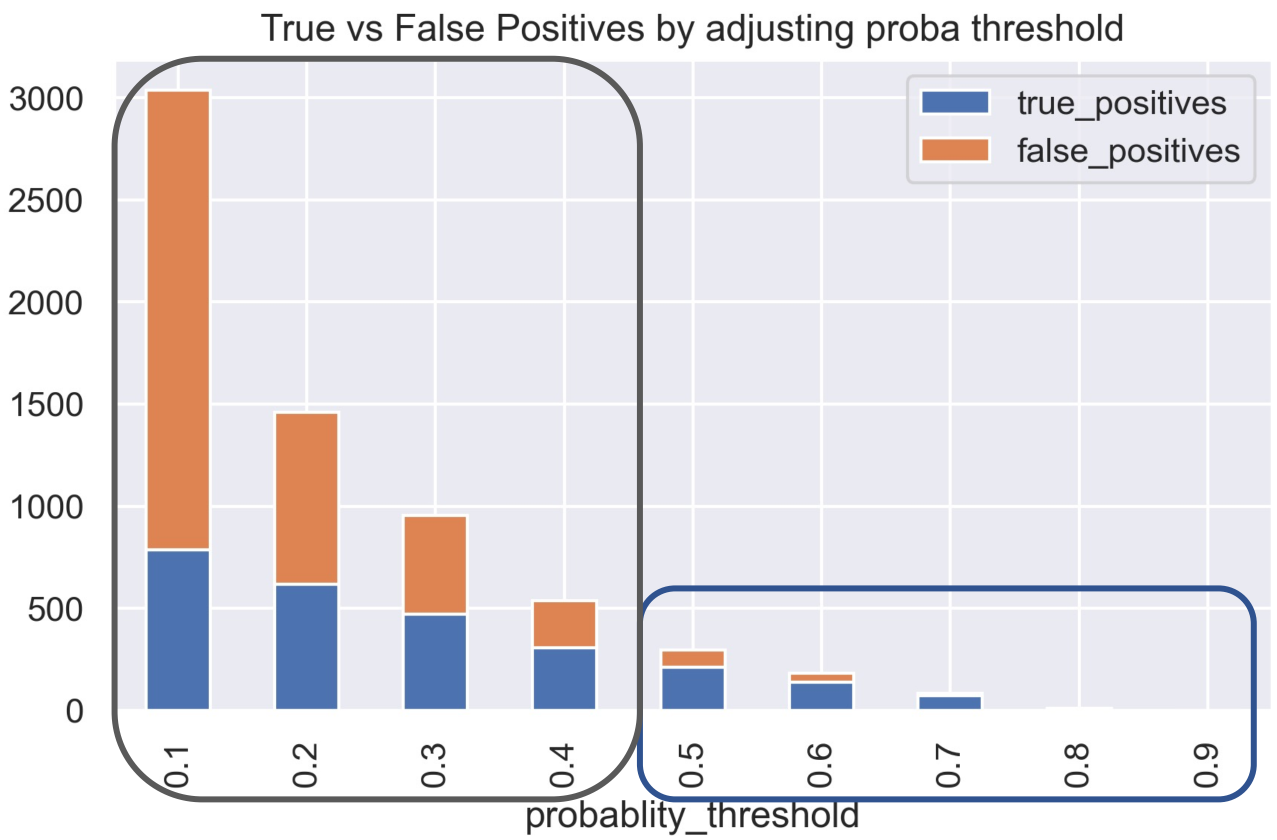
Solution: RF model w/ Probability Threshold Controlling



Proba threshold	True positives	False Positives	Predicted Positives	Precision Score
0.1	787	2250	3037	0.26
0.2	617	841	1458	0.46
0.3	472	481	953	0.49
0.4	306	233	539	0.56
0.5	211	85	296	0.71
0.6	140	41	181	0.77
0.7	72	10	82	0.88
0.8	11	1	12	0.92
0.9	0	0	0	

At prob threshold 0.8, precision 0.92 with only 11 positives.
At prob threshold 0.9, no postives being captured.
Best Precision ≠ Best Result

Solution: RF model w/ Probability Threshold Controlling



Proba threshold	True positives	False Positives	Predicted Positives	Precision Score
0.1	787	2250	3037	0.26
0.2	617	841	1458	0.46
0.3	472	481	953	0.49
0.4	306	233	539	0.56
0.5	211	85	296	0.71
0.6	140	41	181	0.77
0.7	72	10	82	0.88
0.8	11	1	12	0.92
0.9	0	0	0	

Compare Proba threshold 0.1 and 0.2:

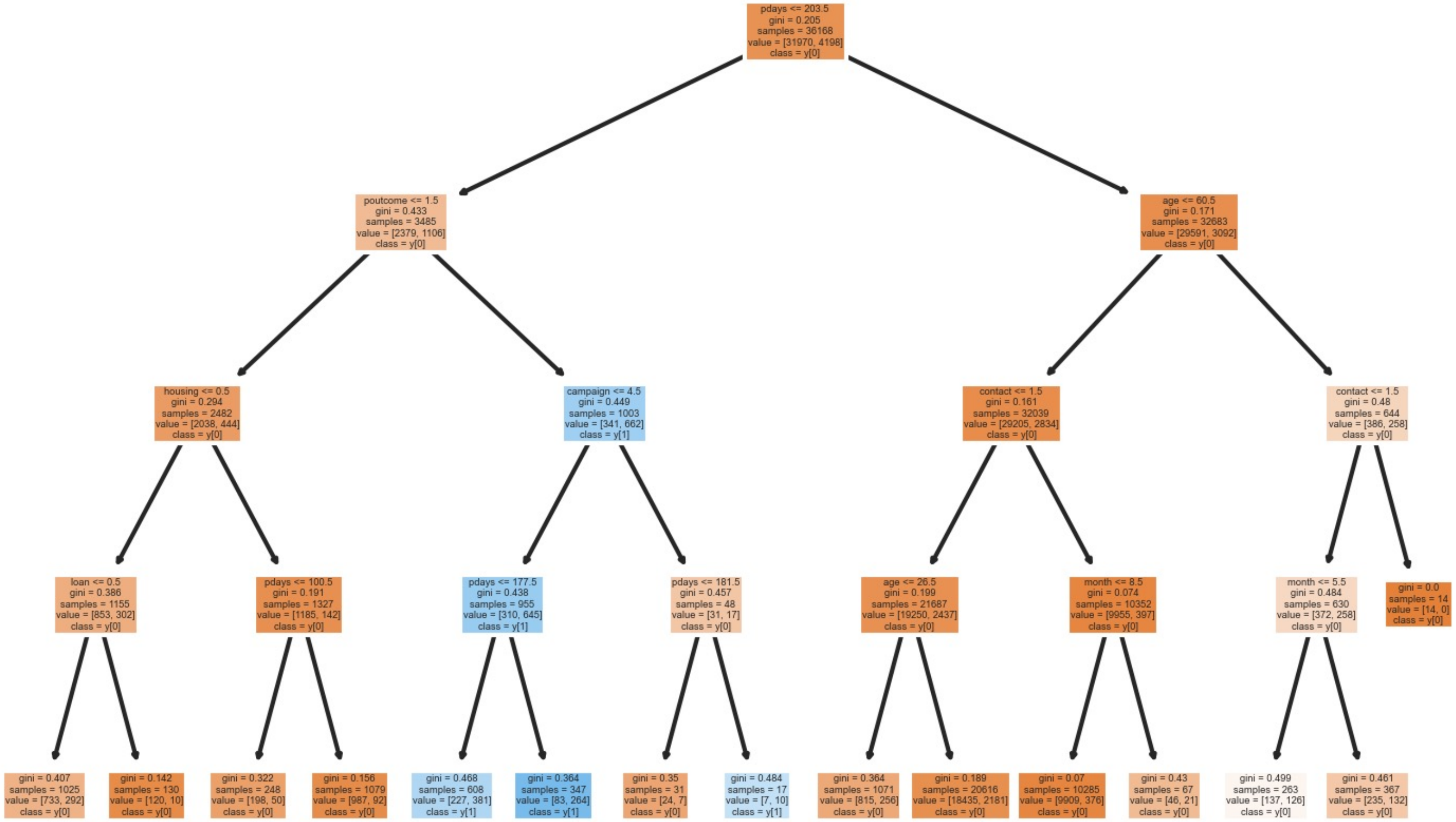
With slightly greater number of true positives, Prob 0.1 has more than twice predicted positives as much as of prob 0.2

Thank you!

Futher Work

- Boosting with Ada Boost, XG boost
- More Models: SVM, Naive Bayes, etc.

Appendix (1): Visualize the Tree (max_depth=4)



Appendix (2): Visualize the Tree (Greedy Apporach)