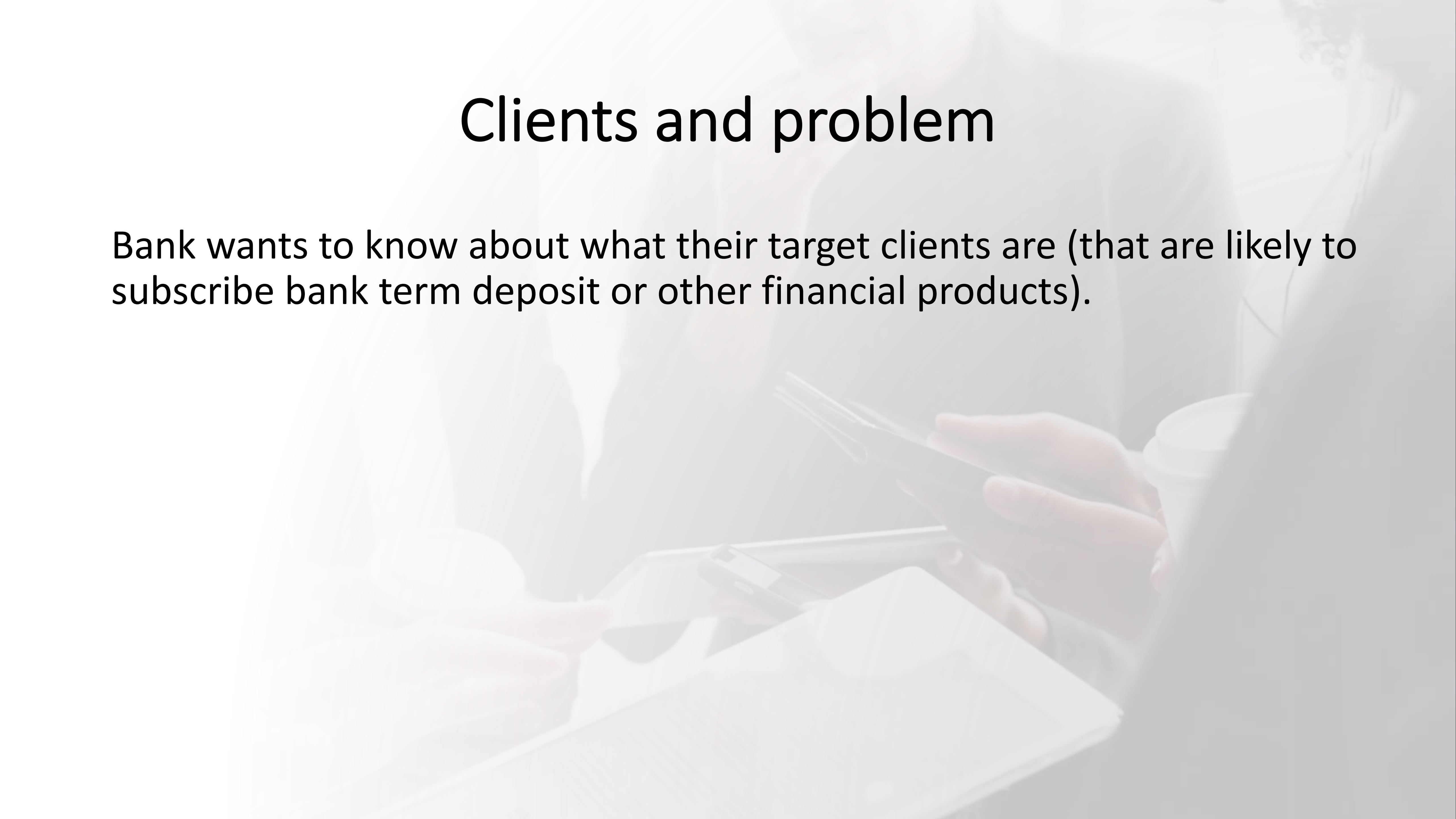
Bank Marketing Success Classification problem

Rui Yuan

Clients and problem

Bank wants to know about what their target clients are (that are likely to subscribe bank term deposit or other financial products).



Data

- The data is 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.
- 45211 rows and 16 columns.
- Target: Whether client subscribe bank term deposit (1 = yes, 0 = no)

15 Features

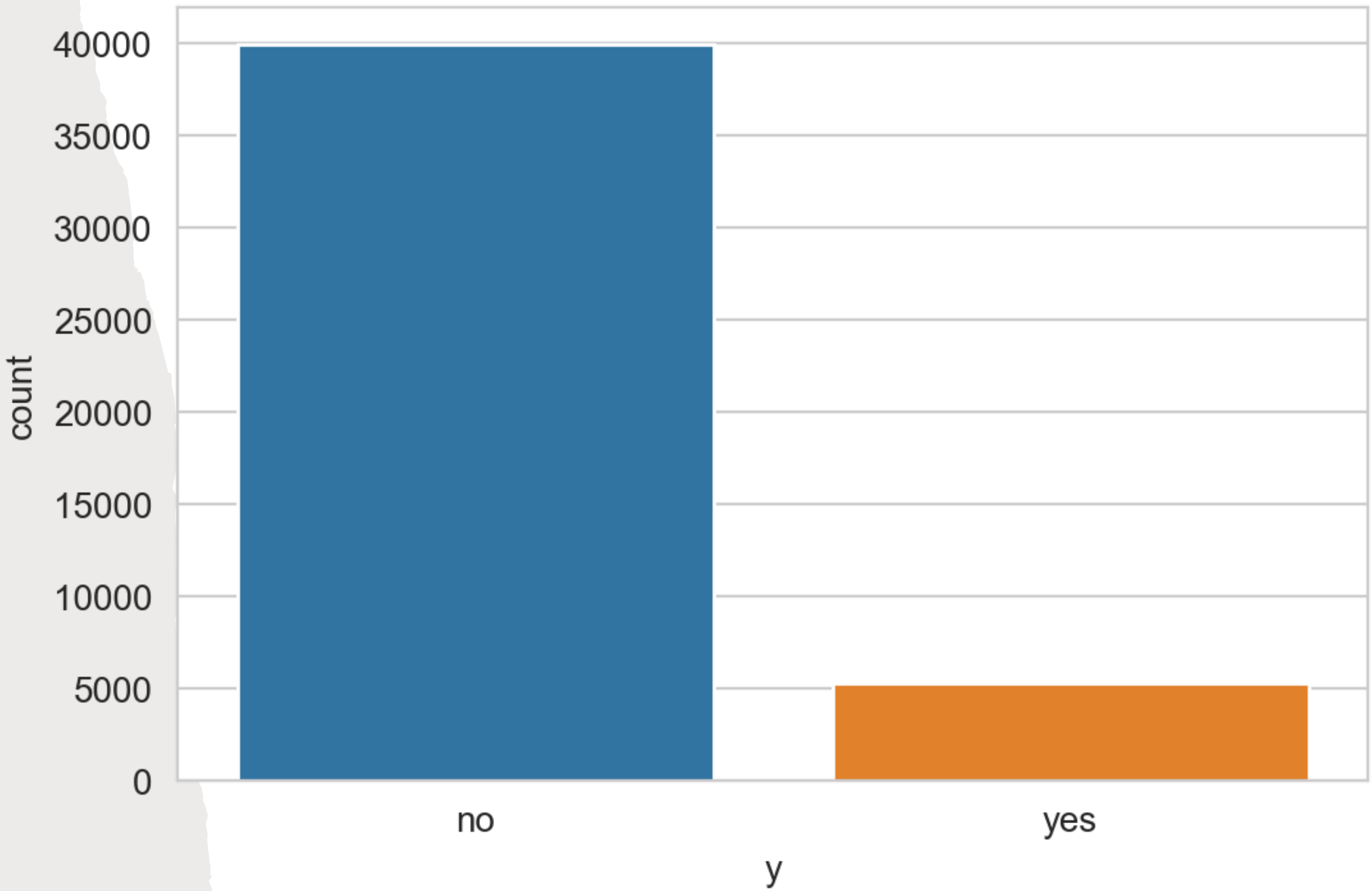
bank clients' data:

- 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



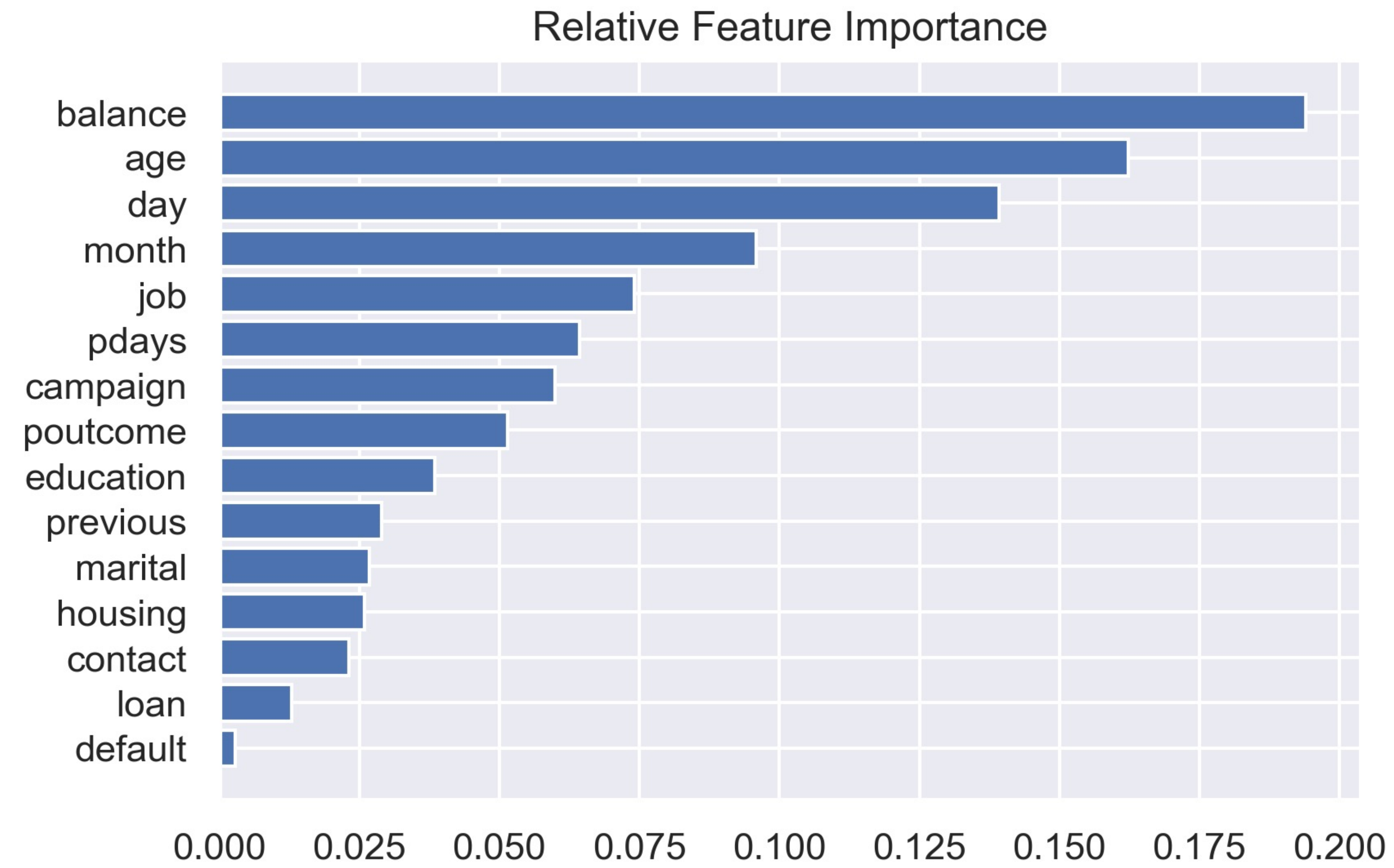
Classification Modeling Goal

- Goal: ↑ precision score on positive class (subscribe) & number of positives
- In Business Sense: Adjustable model depending on the business capability. I.e., how many clients bank can reach to or how many phone calls bank can do in a given period (week/month/year).

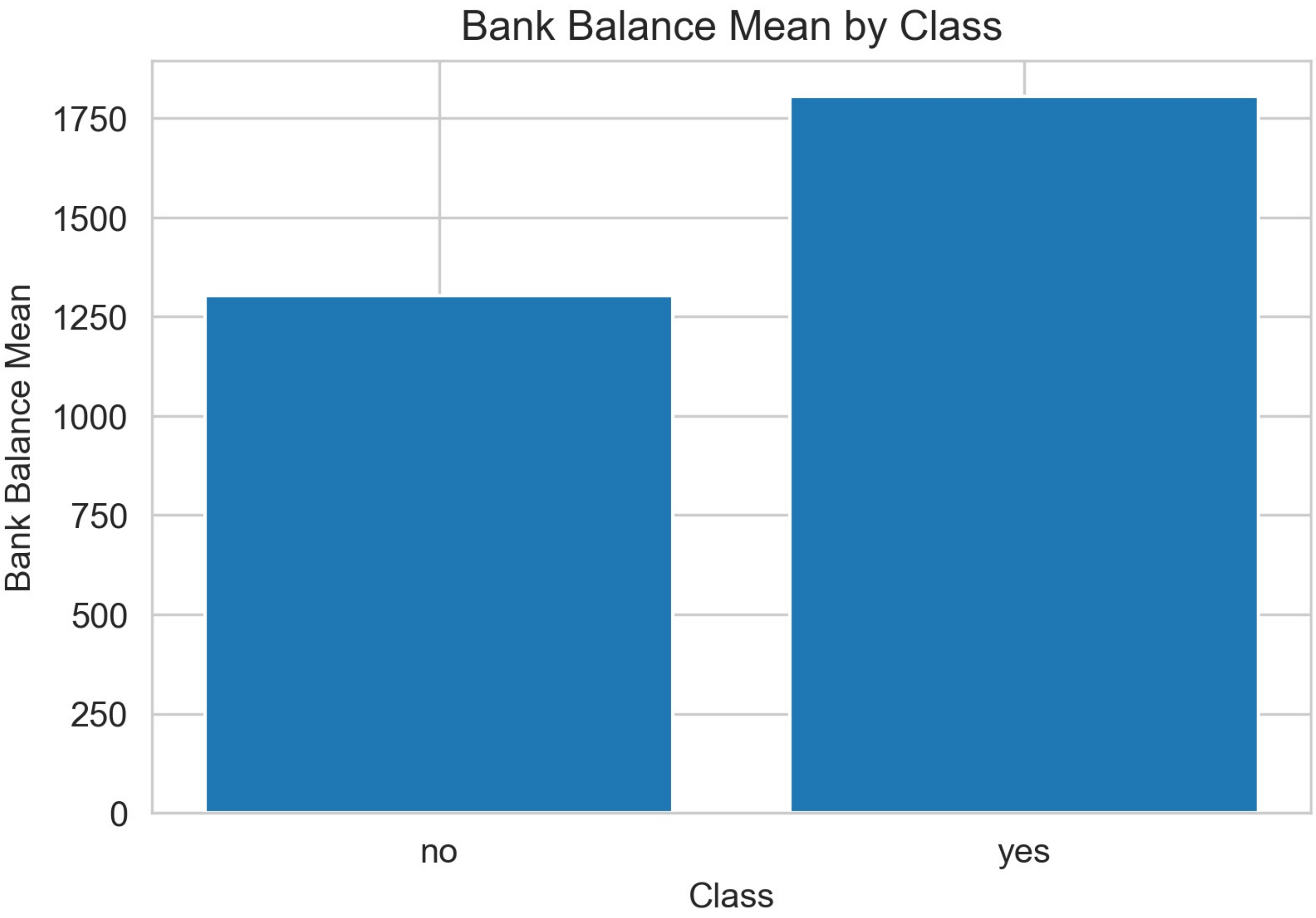
Modeling & Performance Metric Report

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

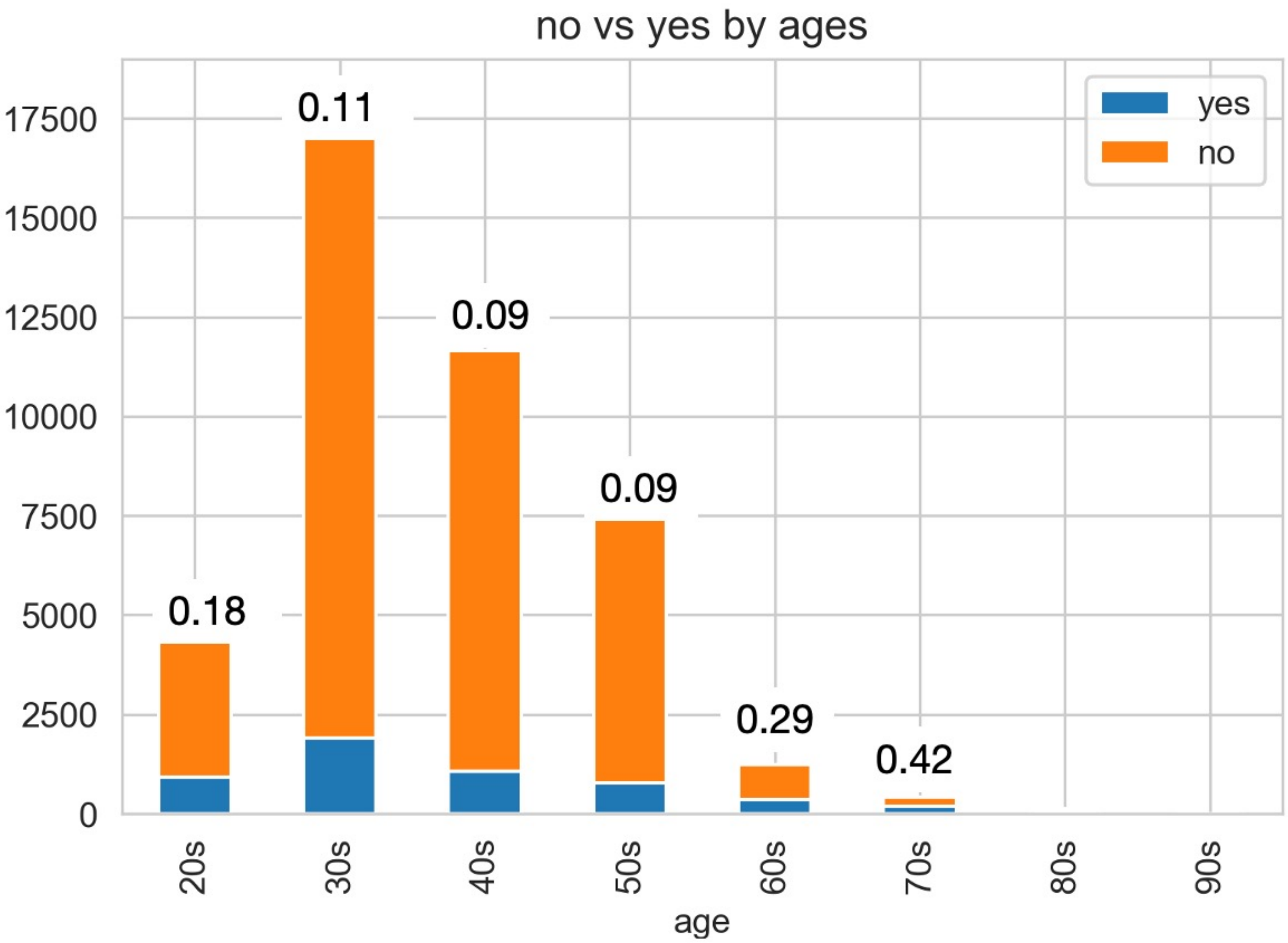
Feature Importance from RF model



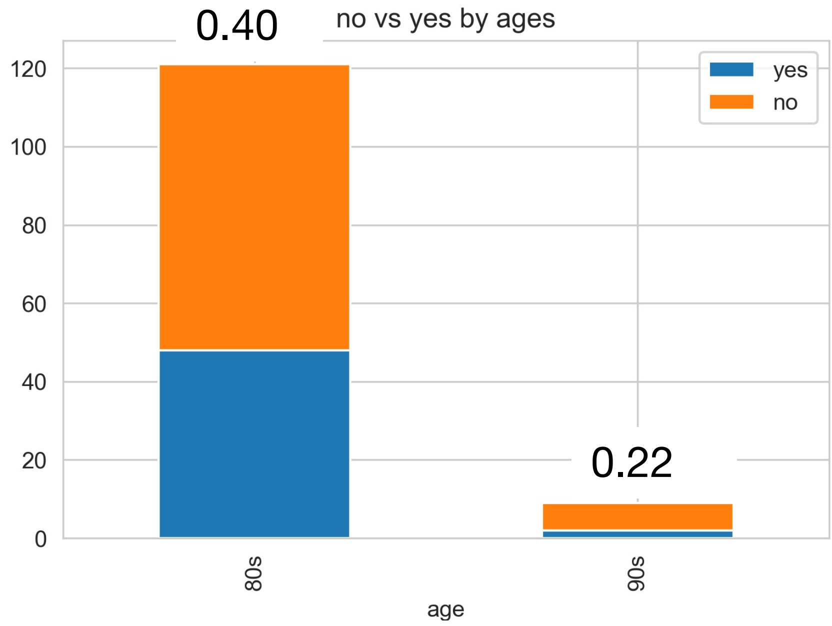
Closer look at important features: balance



Closer look at important features: age

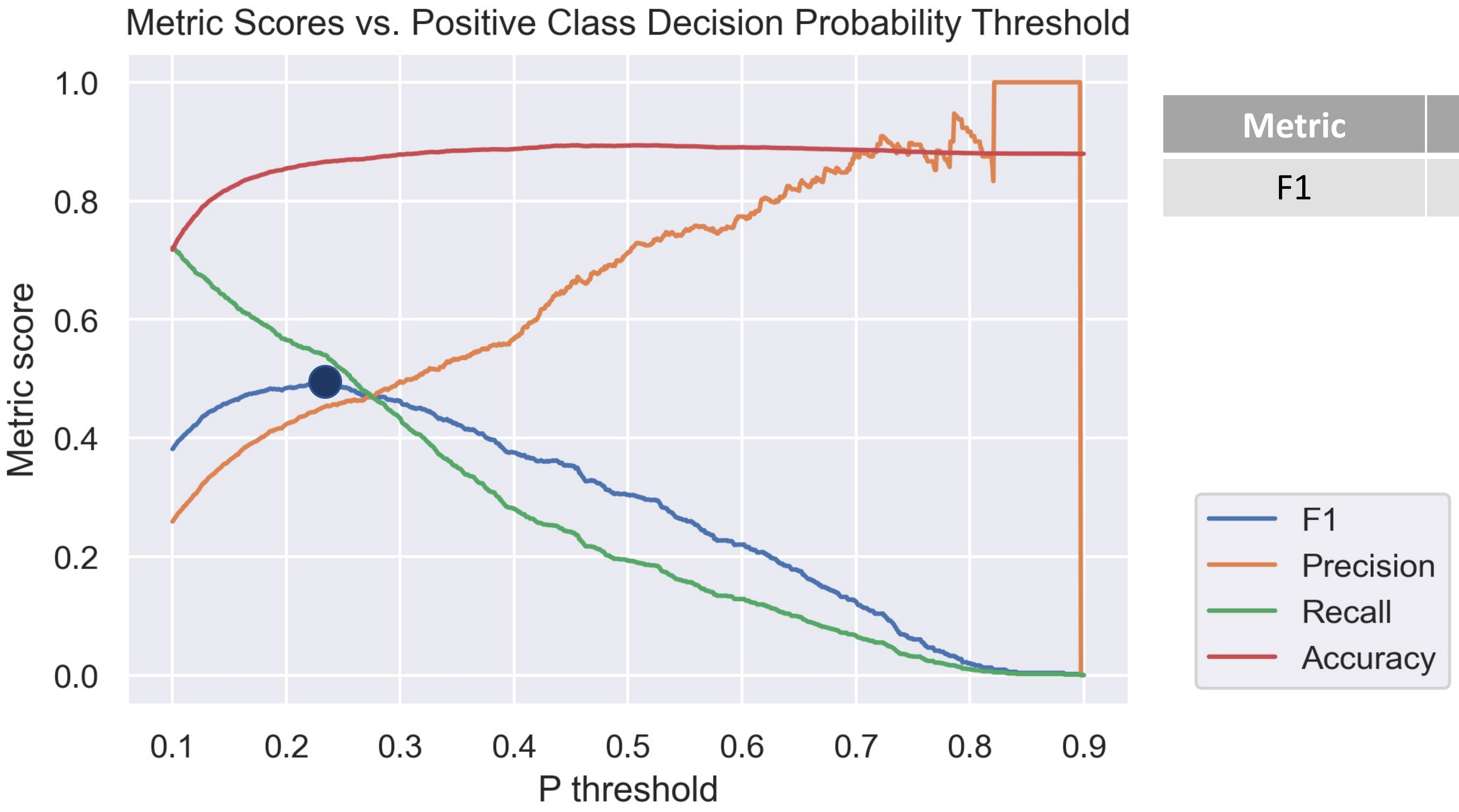


Zoom in 80s and 90s



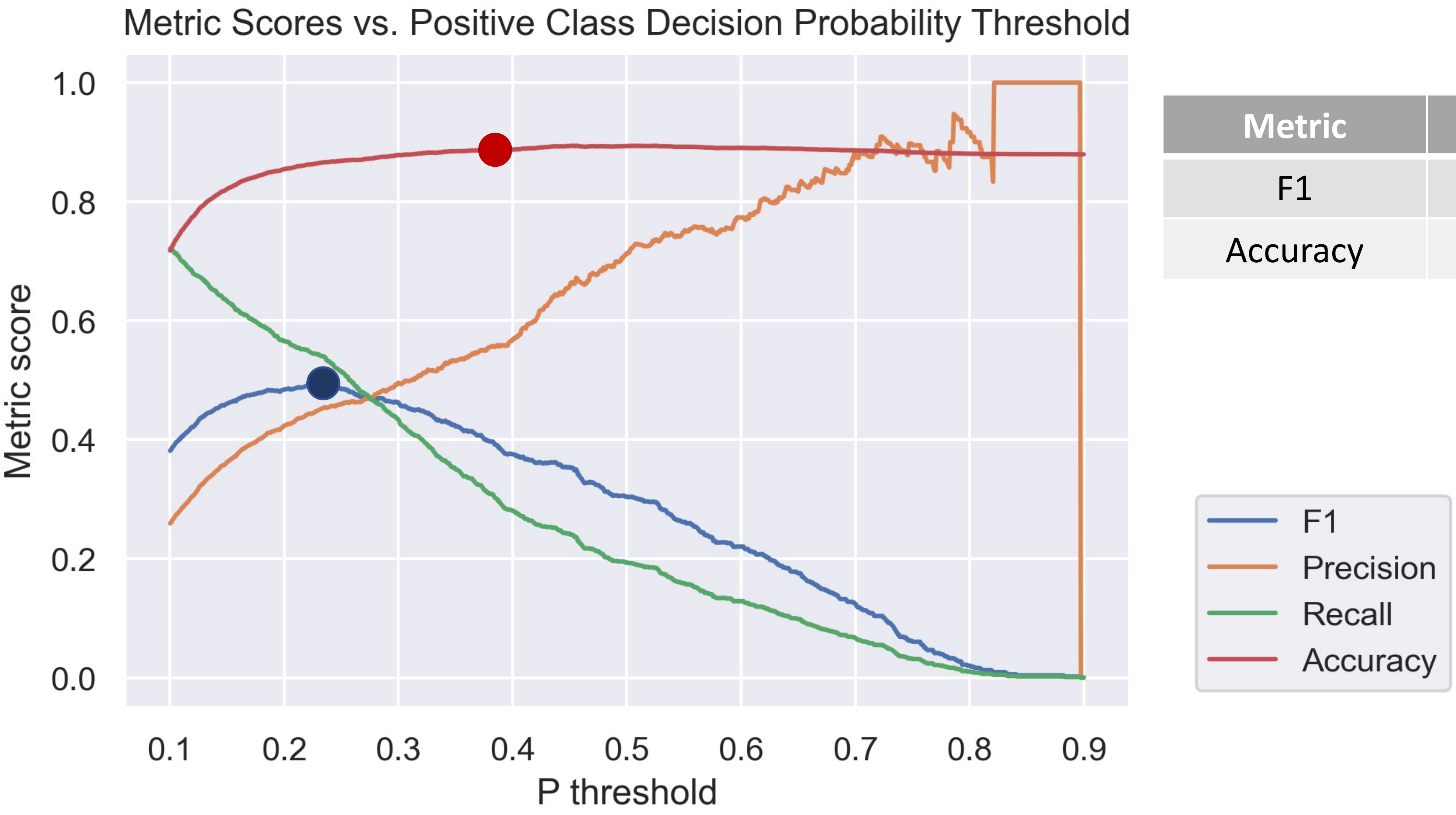
Elderly groups have higher success rate but lower number of success

Solution: RF model w/ Probability Threshold Controlling



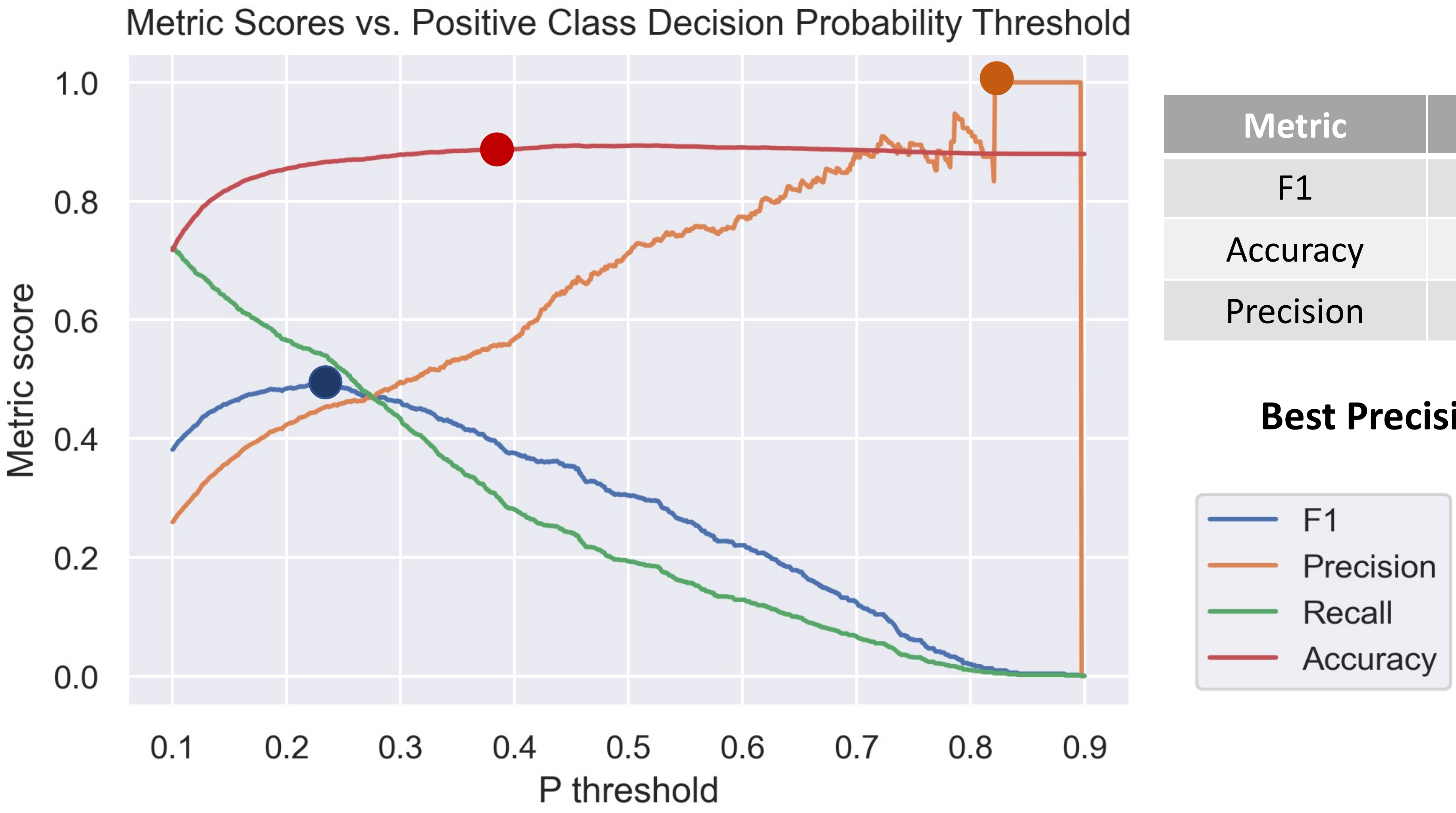
Metric	Best Score	Probablity
F1	0.469	0.240

Solution: RF model w/ Probability Threshold Controlling



Metric	Best Score	Probablity
F1	0.469	0.240
Accuracy	0.894	0.458

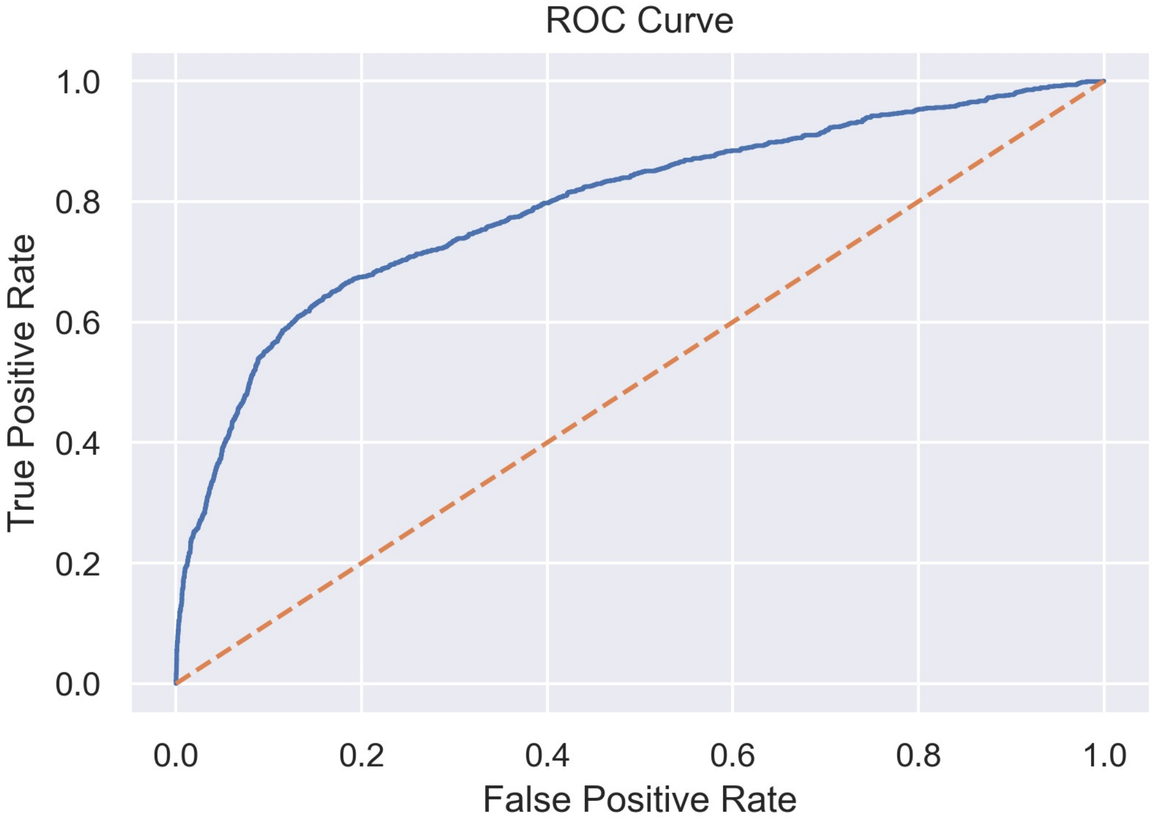
Solution: RF model w/ Probability Threshold Controlling

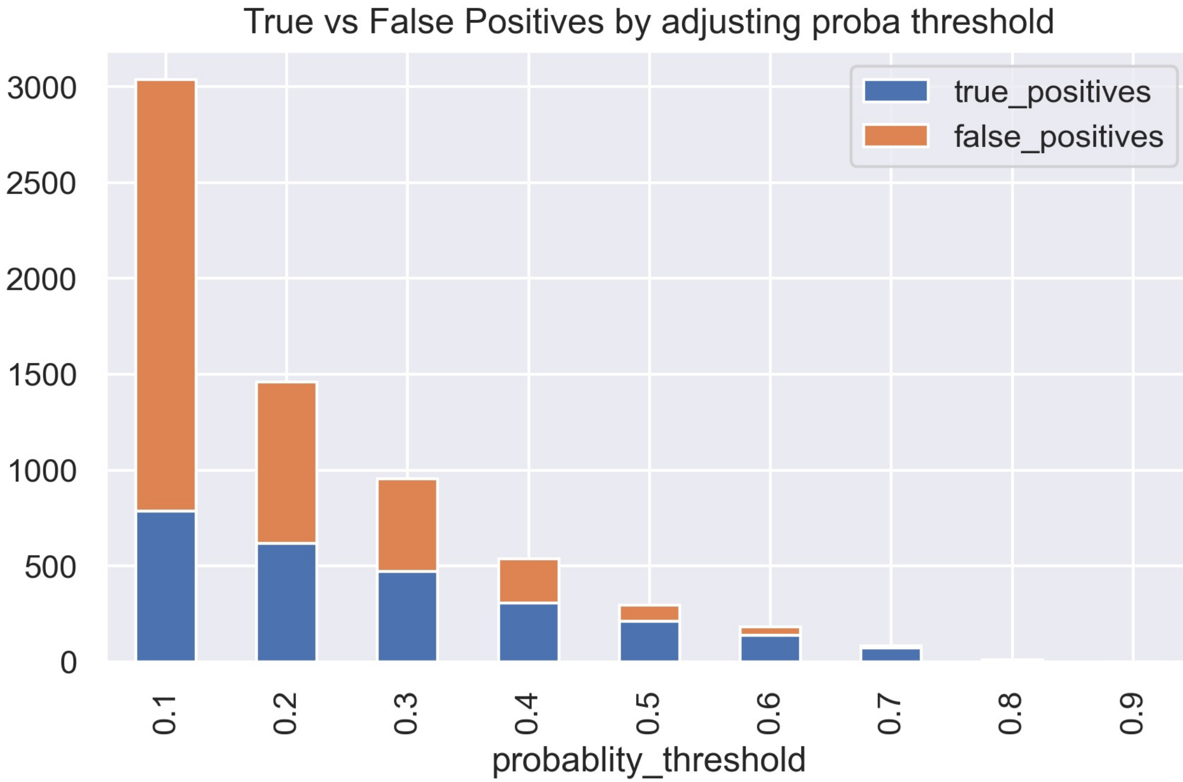


Metric	Best Score	Probablity
F1	0.469	0.240
Accuracy	0.894	0.458
Precision	1.00	0.822

Best Precision ≠ Best Result

Solution: RF model w/ Probability Threshold Controlling





Visualize the Tree

Futher Work

- Ensembling with Ada Boost, XG boost
- More Models