Lead Conversion Analysis and Prediction

FOR TERM DEPOSIT ACCOUNT OPENINGS

FEMI ONAFALUJO

SPRINGBOARD DATA SCIENCE

SEPTEMBER 26, 2021

"The more you know about the past, the better prepared you are for the future."

- THEODORE ROOSEVELT (1858-1919)

Project Objectives

Problem

- Help banks determine promising sales leads

Solution

- Develop a predictive model

It is a Journey: Solutions Areas & Scopes

Dataset

- Data collection

Exploratory Data Analysis

- Relationships between positive response and features

Machine Learning

- Model development
- Model selection
- Model application

Dataset

COLLECTING DATA

Primary Dataset

Source: UCMachine Learning

Information: Marketing records for selling term deposit accounts

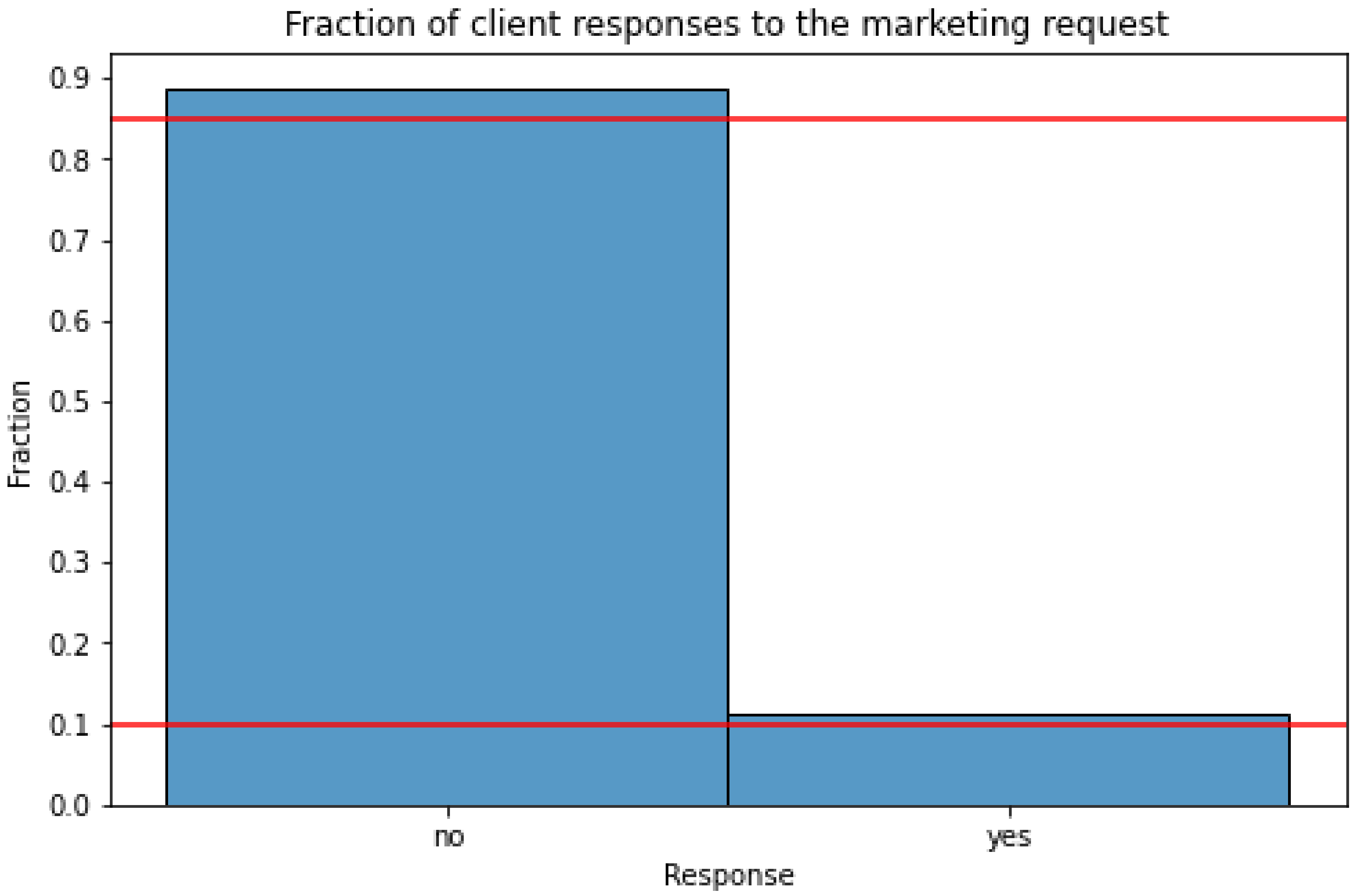
Size: 45,211 records, 21 features

Examples of features

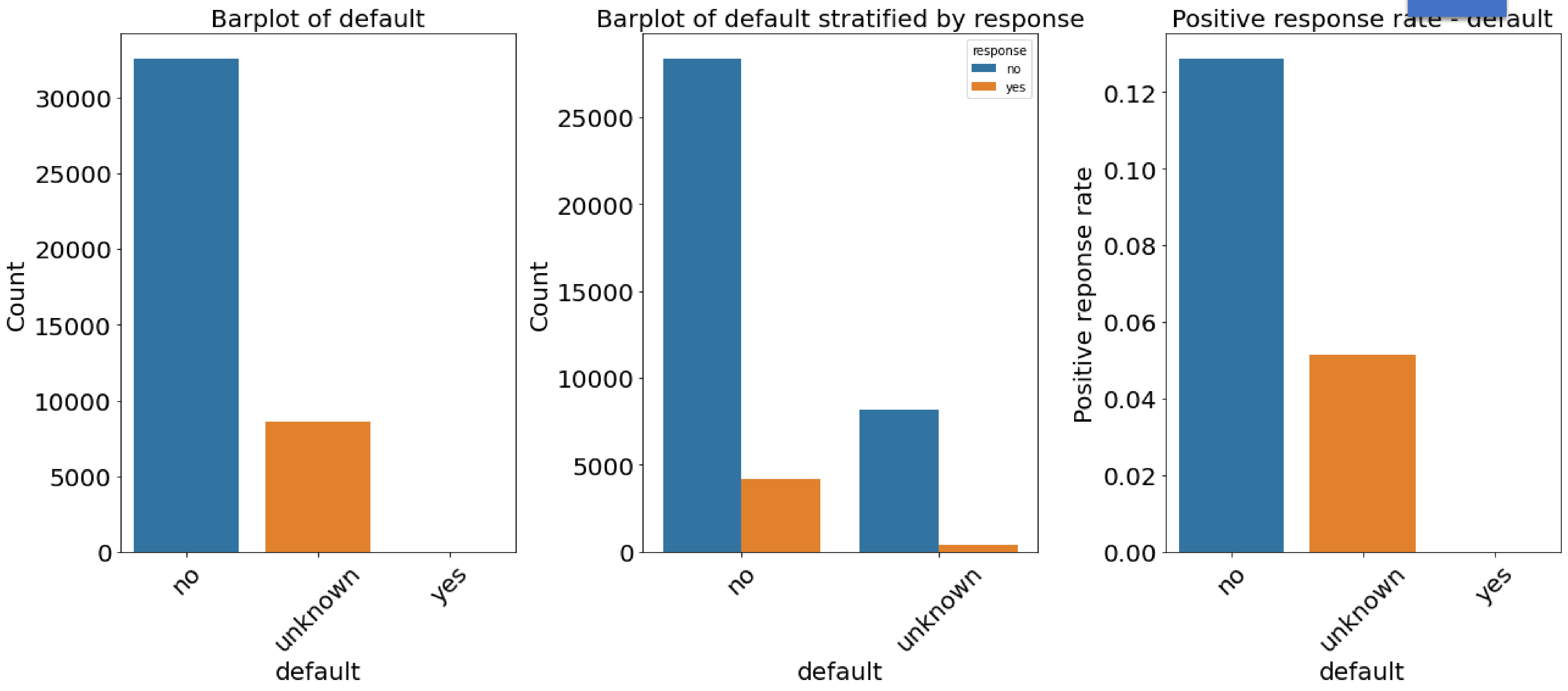
- Age
- Job
- Duration of sales phone call
- Type of contact – e.g. cell phone versus landline
- Marital status

Exploratory Data Analysis

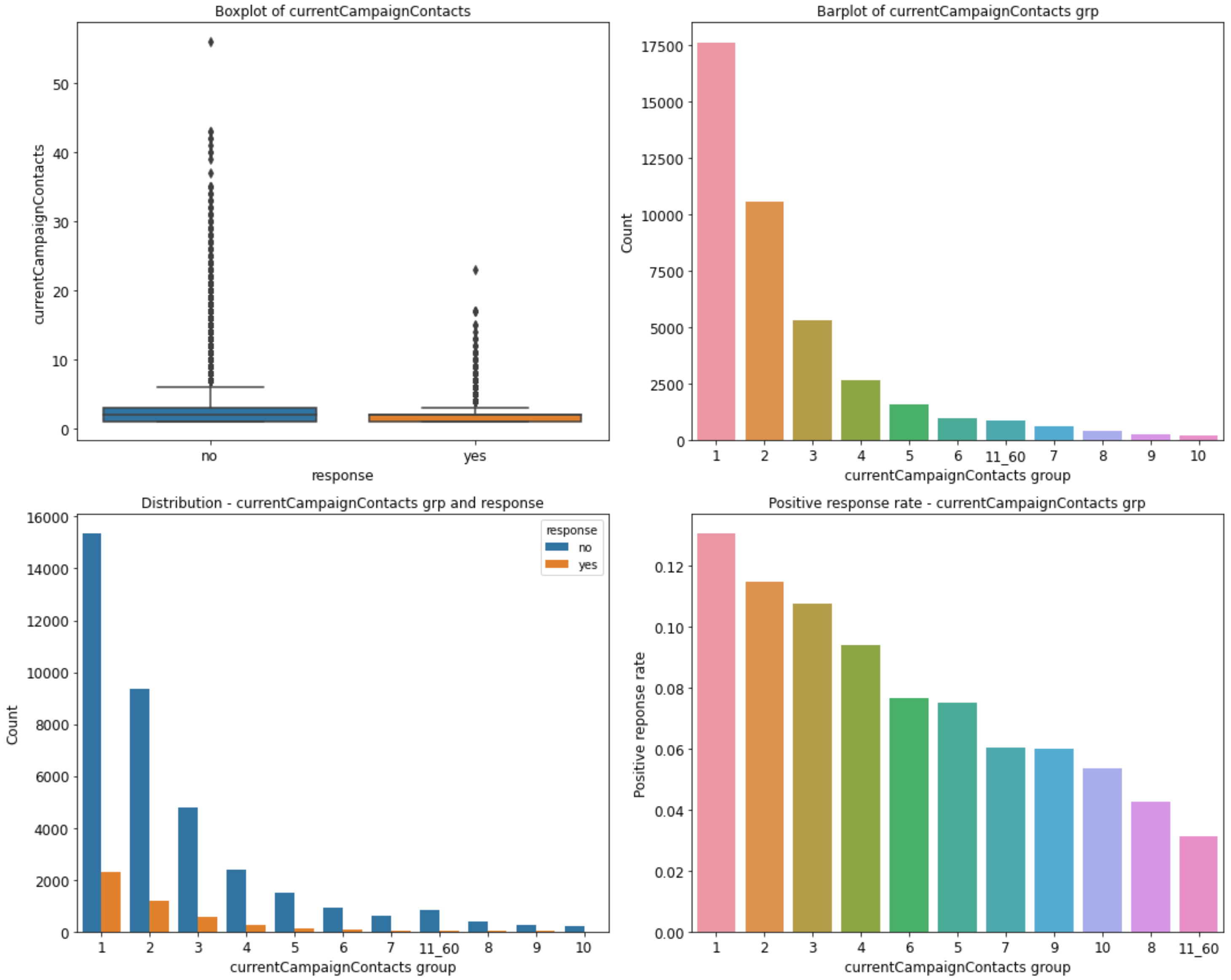
RENT VS. NUMERICAL AND CATEGORICAL FEATURES



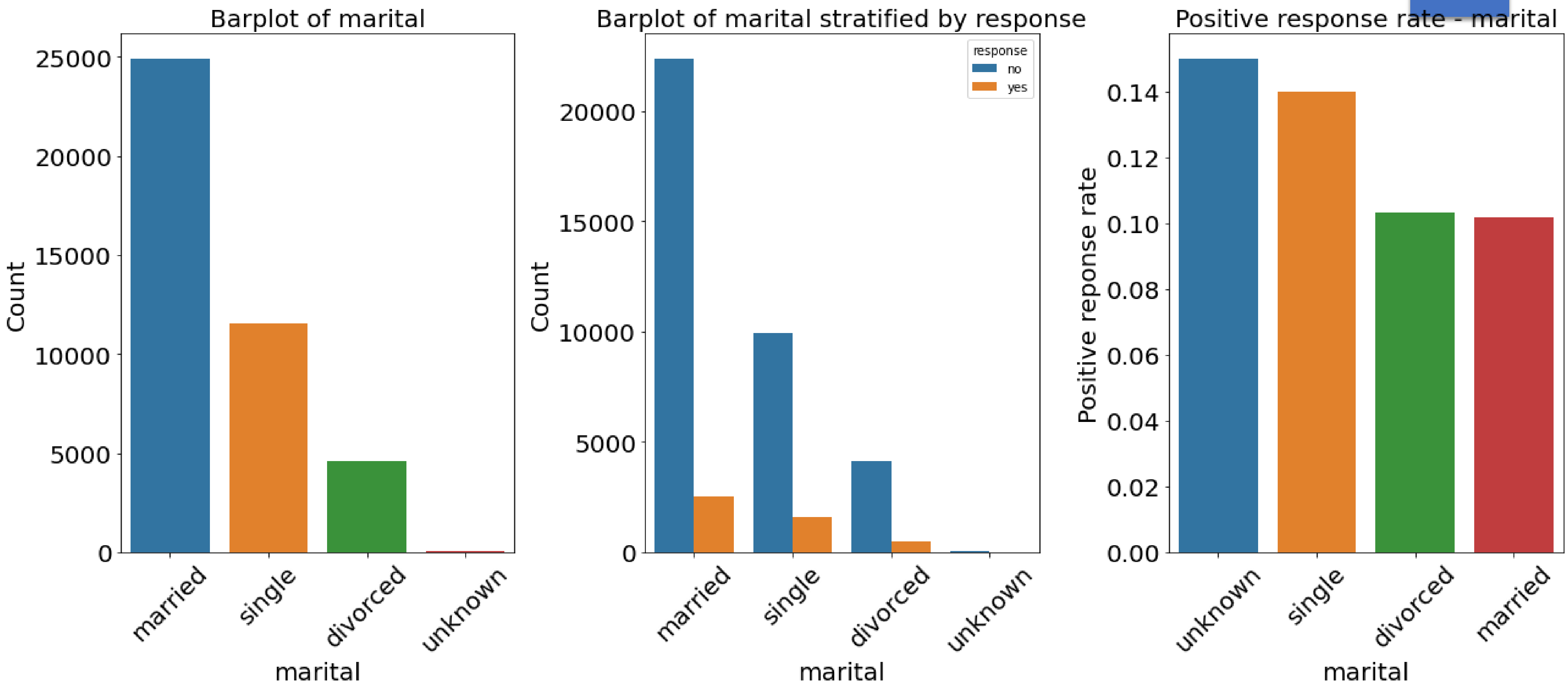
Distribution of response showing imbalanced nature of dataset



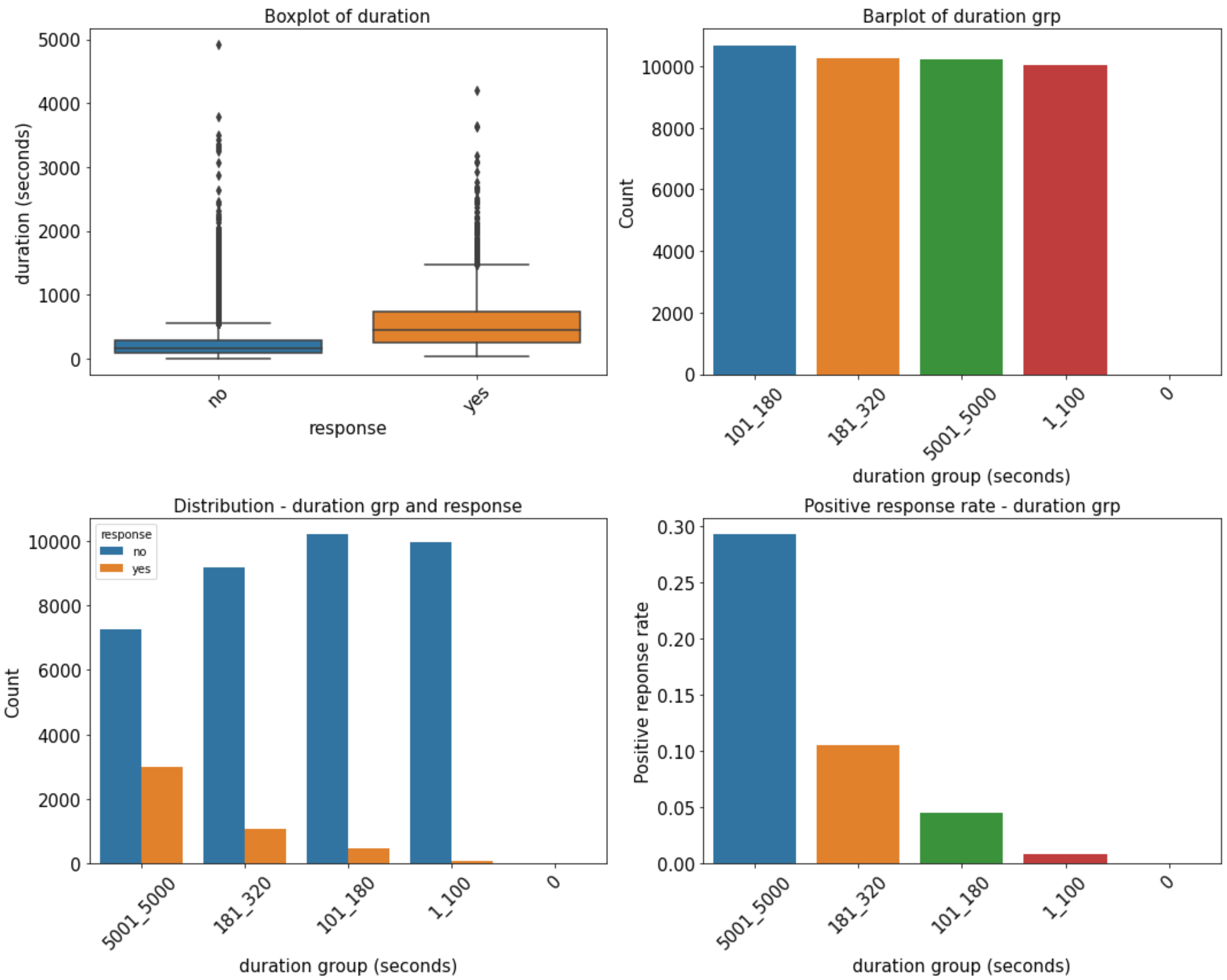
Relationship between a loan default and response



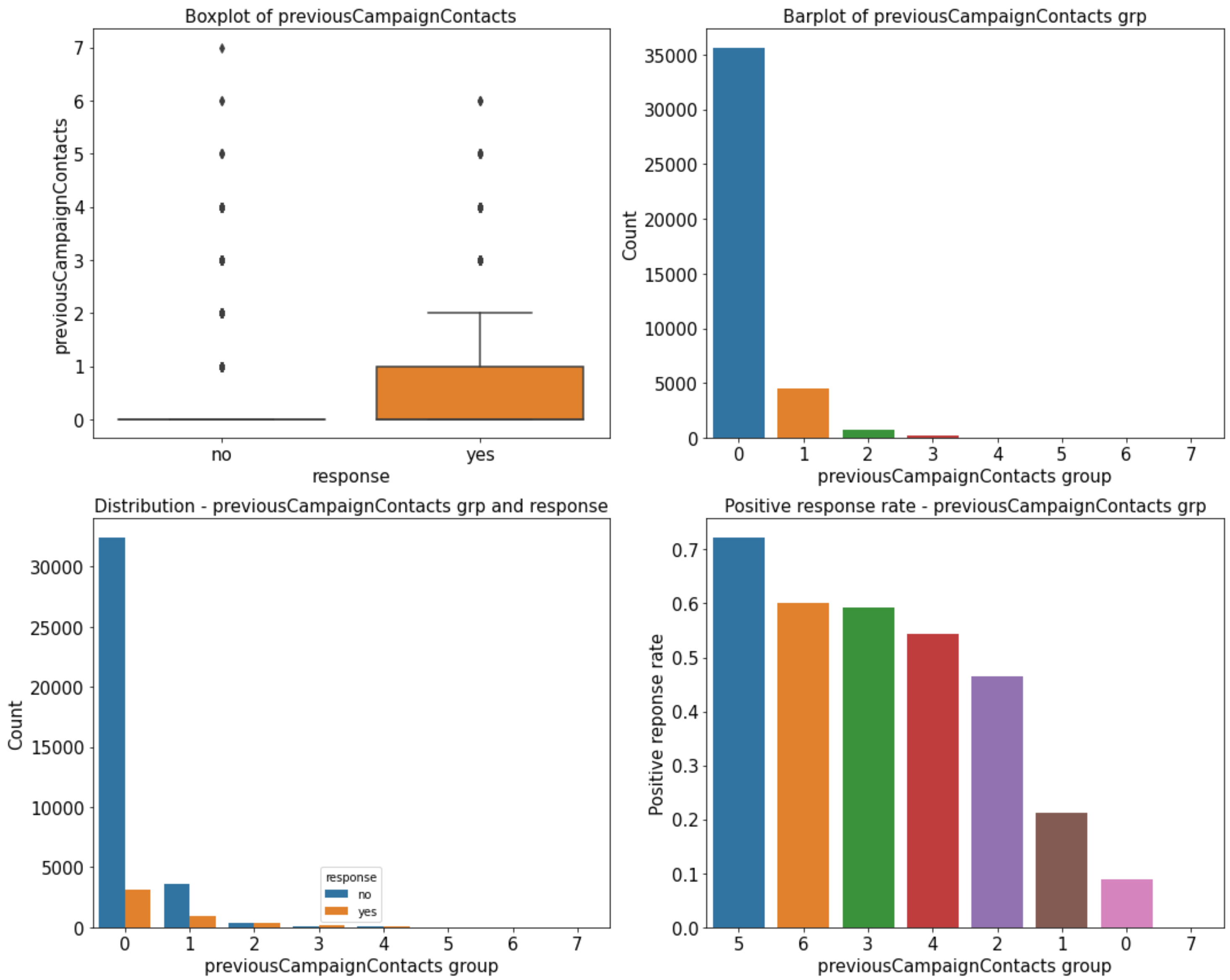
Relationship between current campaign contacts and response



Relationship between marital status and response



Relationship between duration and response



Relationship between previous campaign contacts and response

Machine Learning

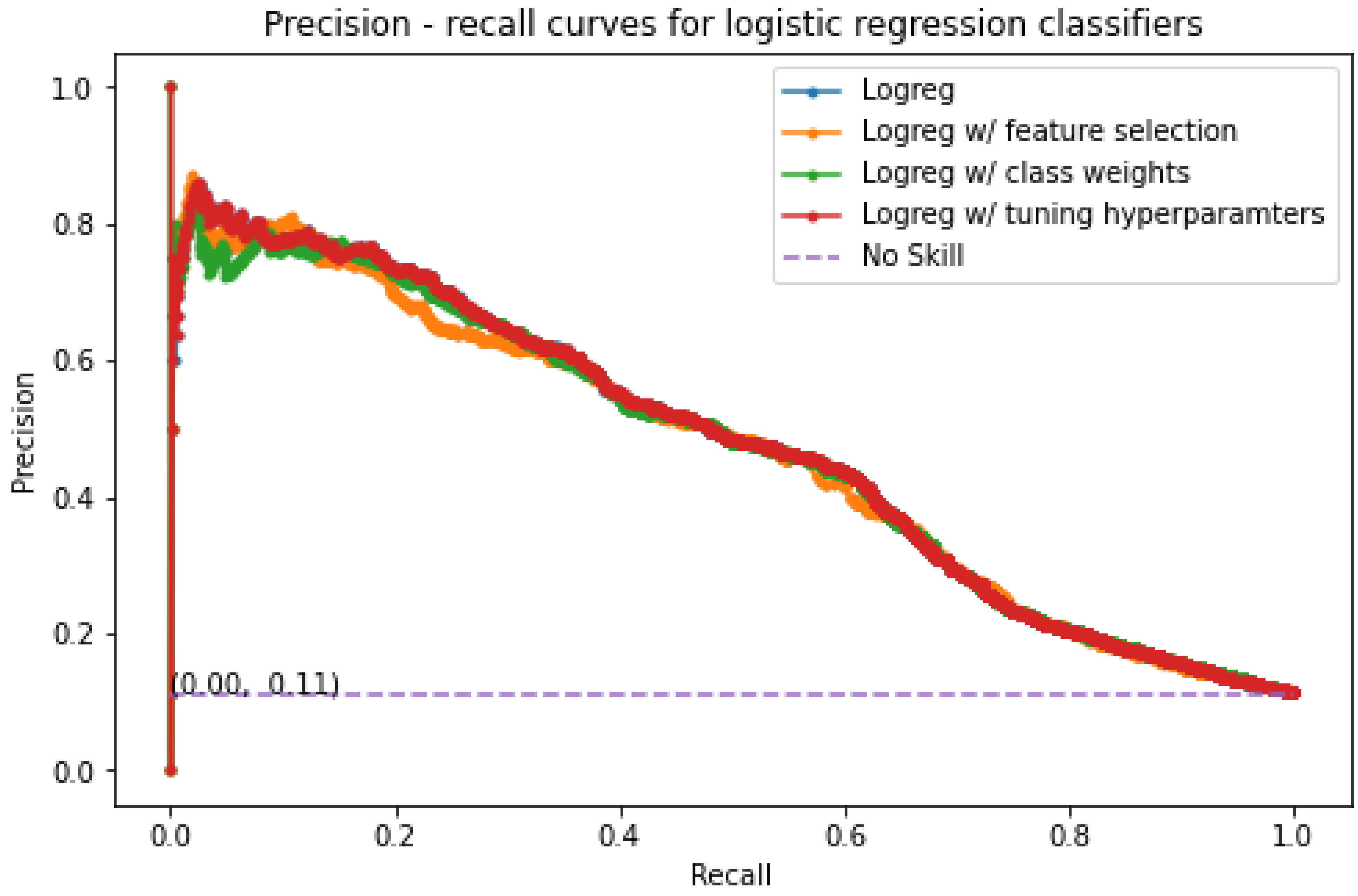
MODEL DEVELOPMENT | SELECTION | APPLICATION

Performance of baseline model

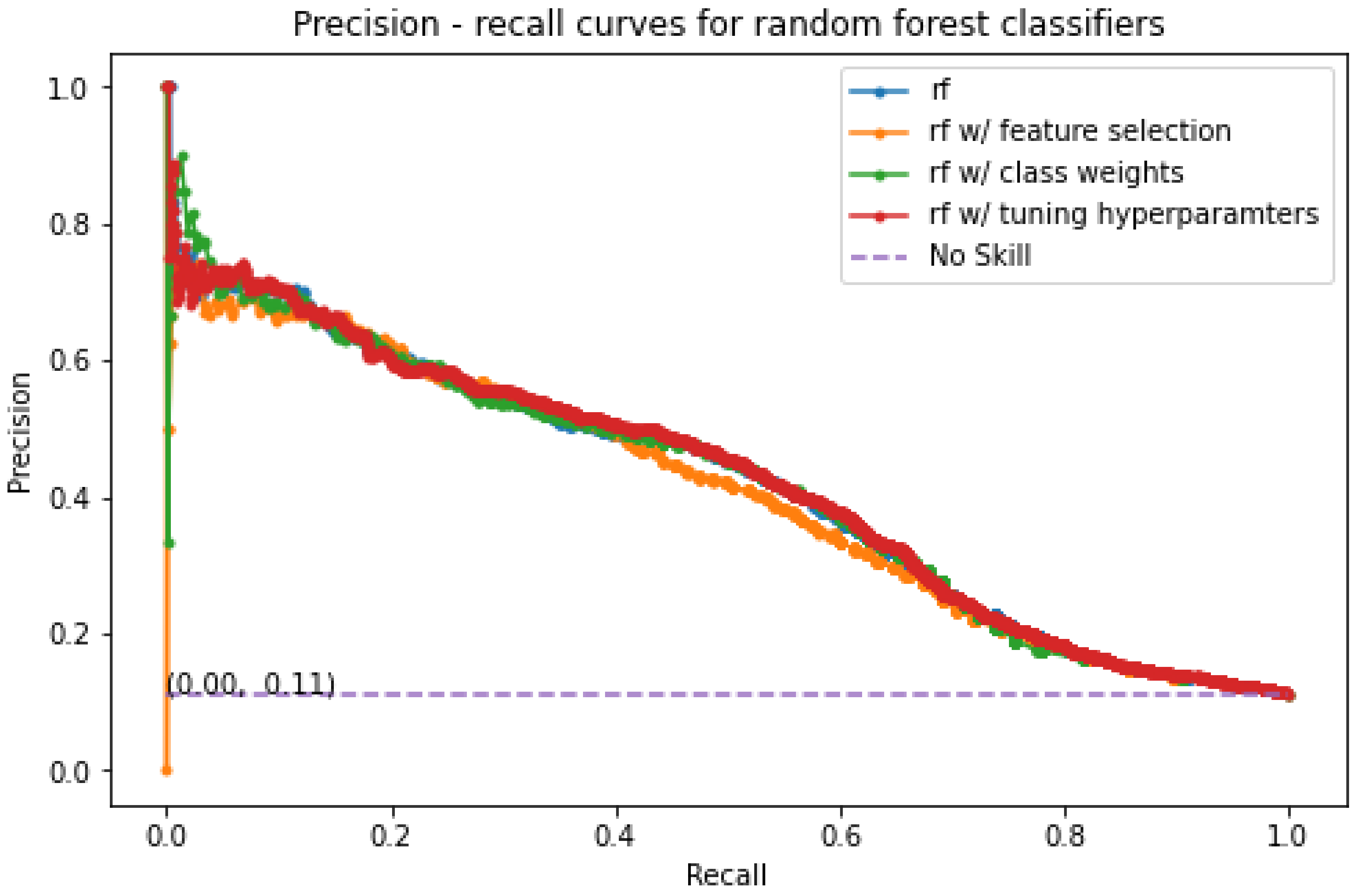
Model	Class	Recall	F1	F2	AUC-PR	AUC-ROC
Logistic regression	0	0.99	0.95		0.48	0.81
	1	0.25	0.37	0.29		

Log

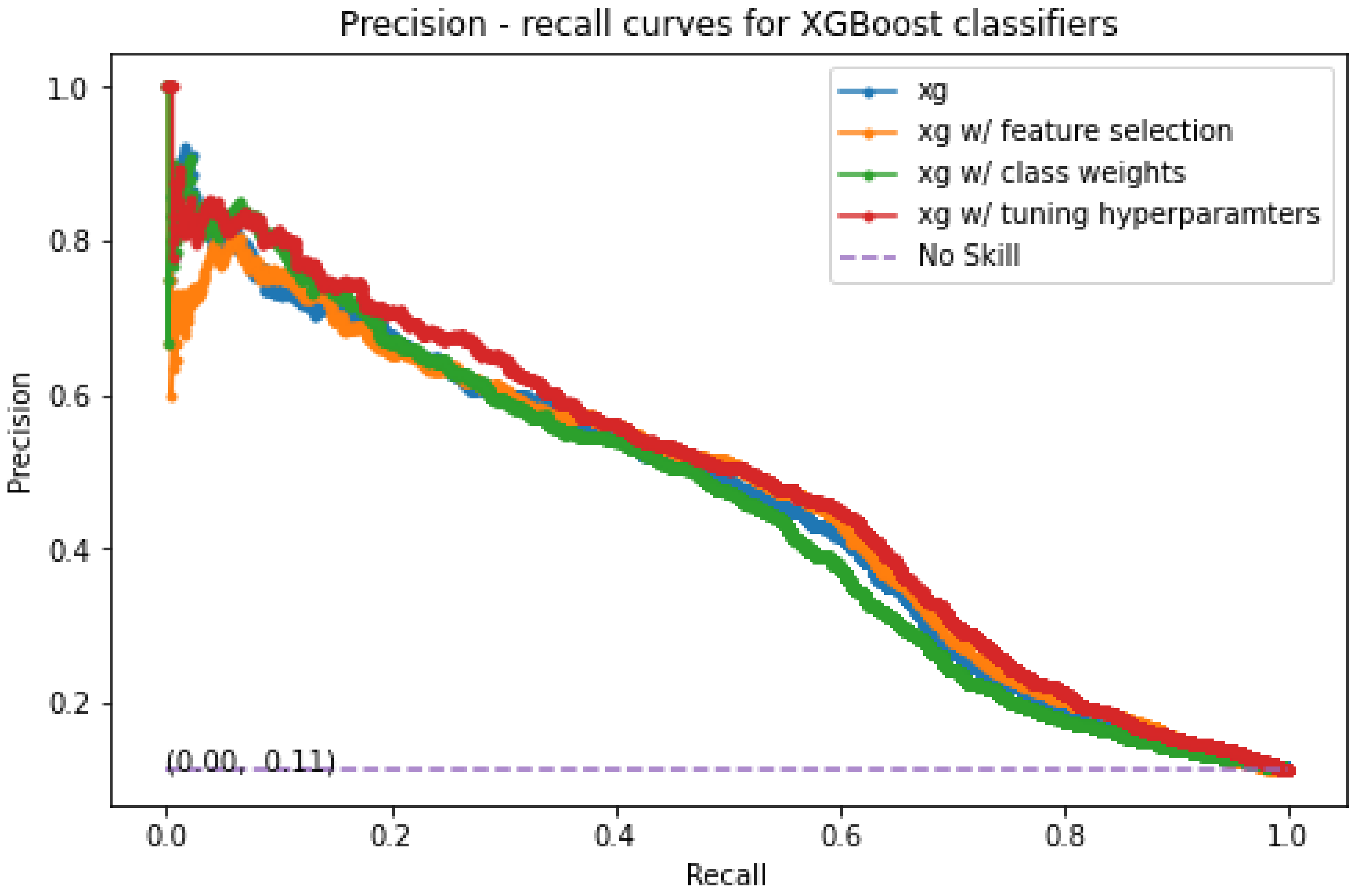
Logistic regression model - logreg	AUC-PR CV scores	AUC-PR test scores
Logreg	0.45	0.48
Logreg w/ class weights	0.45	0.48
Logreg w/ hyperparameter tuning	0.45	0.47
Logreg w/ feature selection	0.41	0.47



Logistic regression optimization results



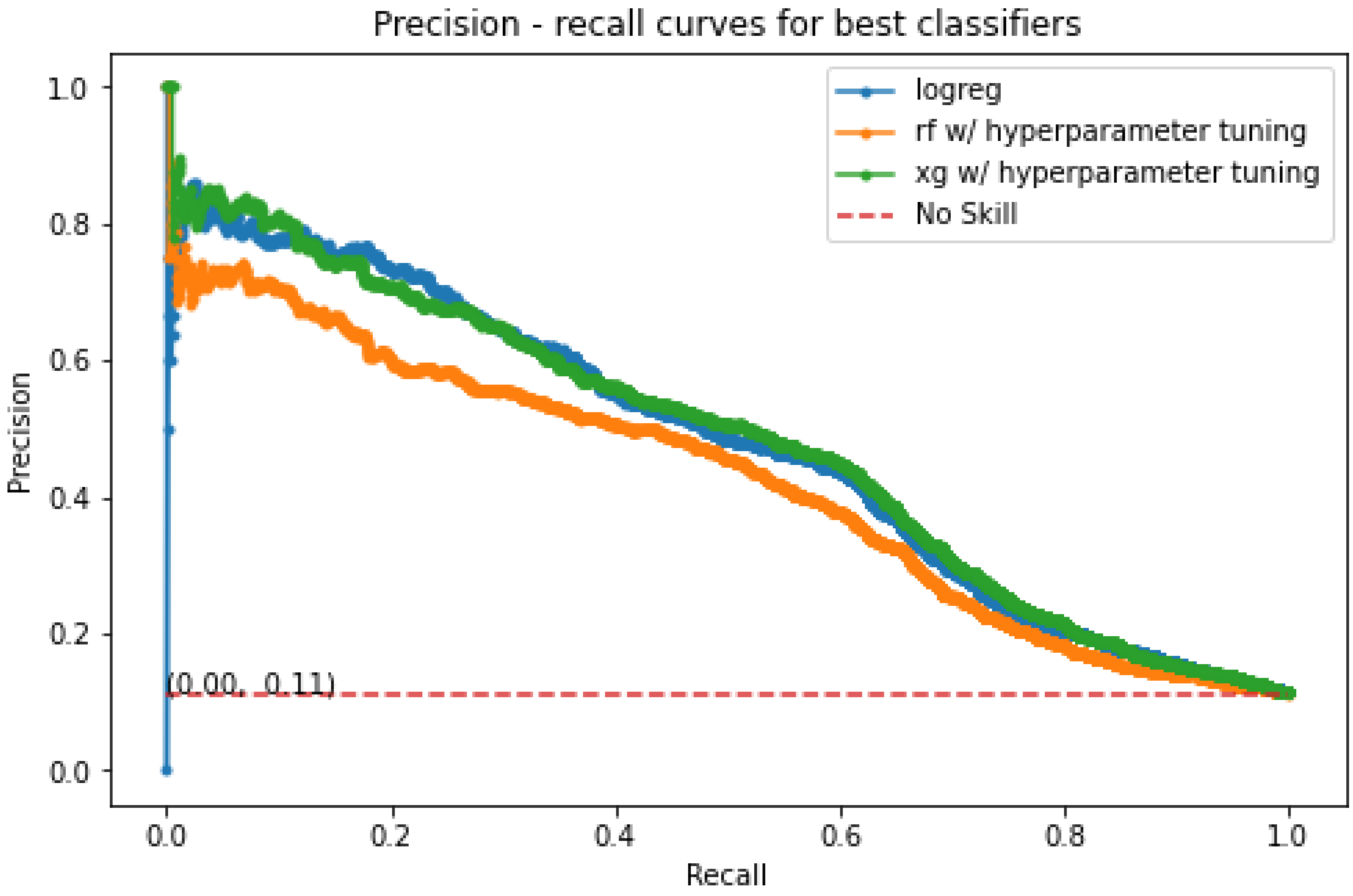
Random forest optimization results



XGBoost optimization results

Best models from algorithms

Best models	AUC-PR CV scores	AUC-PR test scores
xg w/ hyperparameter tuning	0.46	0.48
Logreg	0.45	0.48
rf w/ hyperparameter tuning	0.41	0.42



Comparison of algorithm performance

Results from dataset balancing with SMOTE

SMOTE undersampling	AUC-PR CV scores	AUC-PR test scores
xg w/ hyperparameter tuning	0.93	0.48
Logreg	0.72	0.47
rf w/ hyperparameter tuning	0.94	0.41

Performance summary

Model	Class	Recall	F1	F2	5-fold CV : AUC-PR	AUC-PR	AUC-ROC
XGBoost w/ tuning	0	0.98	0.95		0.46	0.48	0.81
	1	0.27	0.39	0.31			
Logistic regression	0	0.99	0.95		0.45	0.48	0.81
	1	0.25	0.37	0.29			
Random forest w/ tuning	0	0.97	0.94		0.41	0.42	0.78
	1	0.31	0.40	0.34			

Model prediction results for 20 clients

	1	2	3	4	5	6	7	8	9	10
Age	39	29	50	40	34	29	28	30	54	43
Response	No	No	No	No	No	No	Yes	No	Yes	No
Predicted	No	No	No	Yes	No	No	No	No	No	Yes

Unable to predict positive class

Conclusion / Recommendation

Determined most important factors for rent	Predicted response with XGBoost model
• Unclear factors	• AUC-PR: 0.48

Assumptions/Limitations/Opportunities

High uncertainty around data

Additional feature engineering may be beneficial

Higher compute capabilities for hyperparameter tuning

Questions