phData Case Study

(Business Version)

Michael Rowlands

Tax Company's Challenge

For years, Tax Company has been unable to identify leads that will result in sales of your software.

This has resulted in thousands of dollars in unnecessary labor and advertising costs.

Tax Company has created a dataset containing two years of customer information and if you were successful at selling to each customer.

phData's Solution

Investigate the dataset, and determine if a machine learning approach is viable.

If so, create a model to predict if a lead will convert.

Preliminary Assumptions

The dataset is historical so we assume future data is similar to the historical data.

Customers only appear once in the dataset.

Tax Company has no prior knowledge on what features are predictive of sale.

There is a fixed profit for sale (true positives) and fixed loss for attempted sales (false positives).

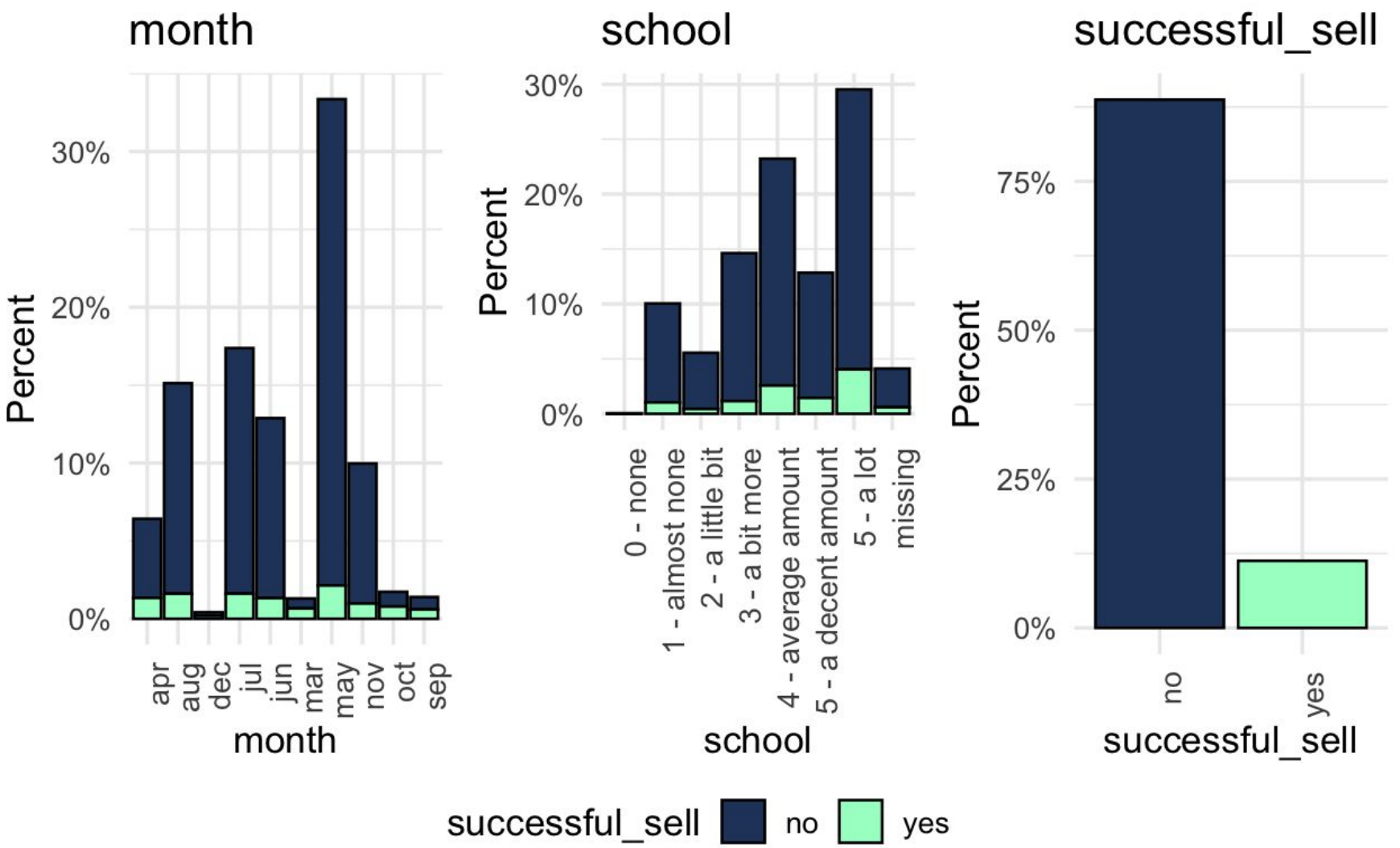
Exploratory Analysis

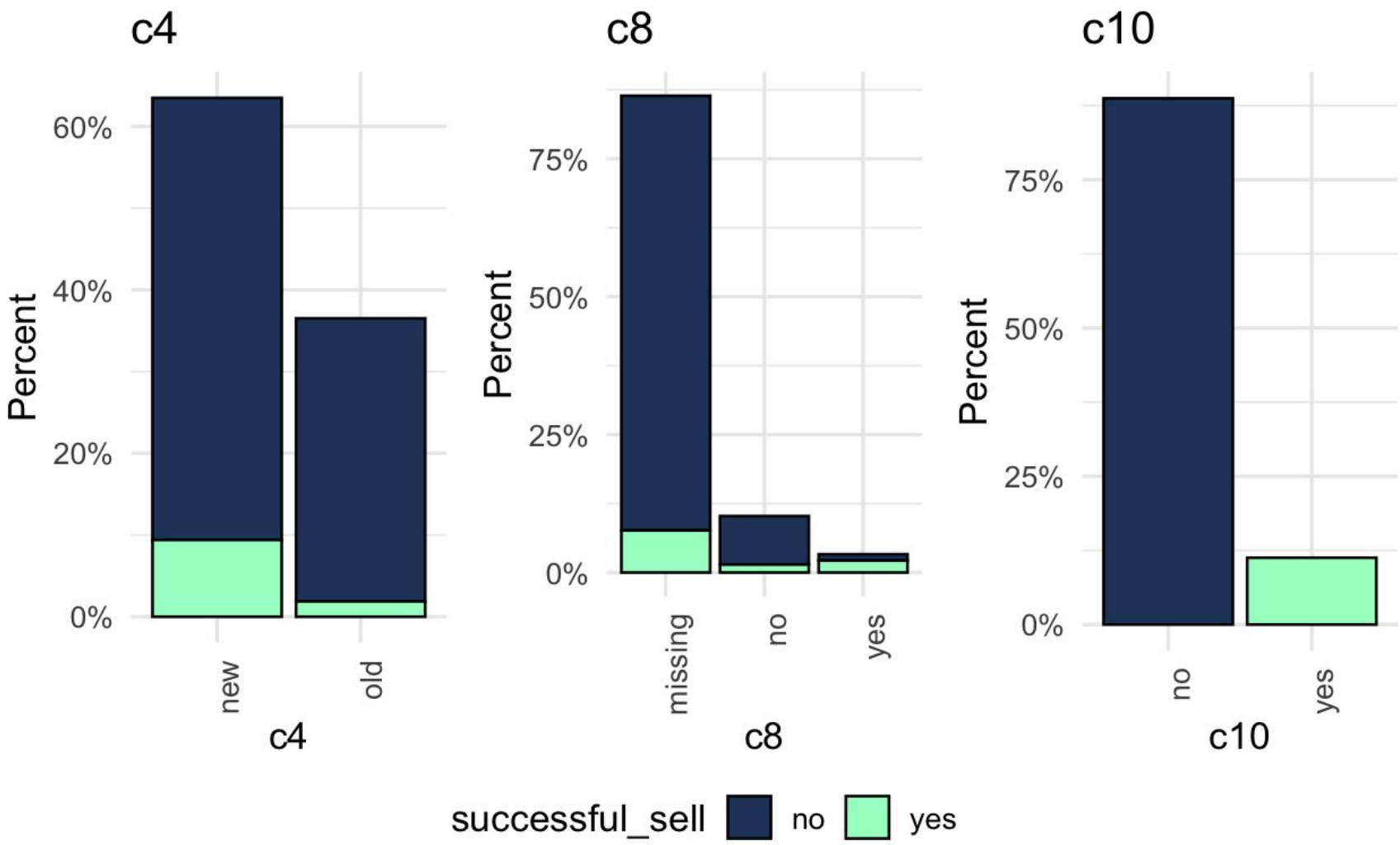
Basic Dataset Info

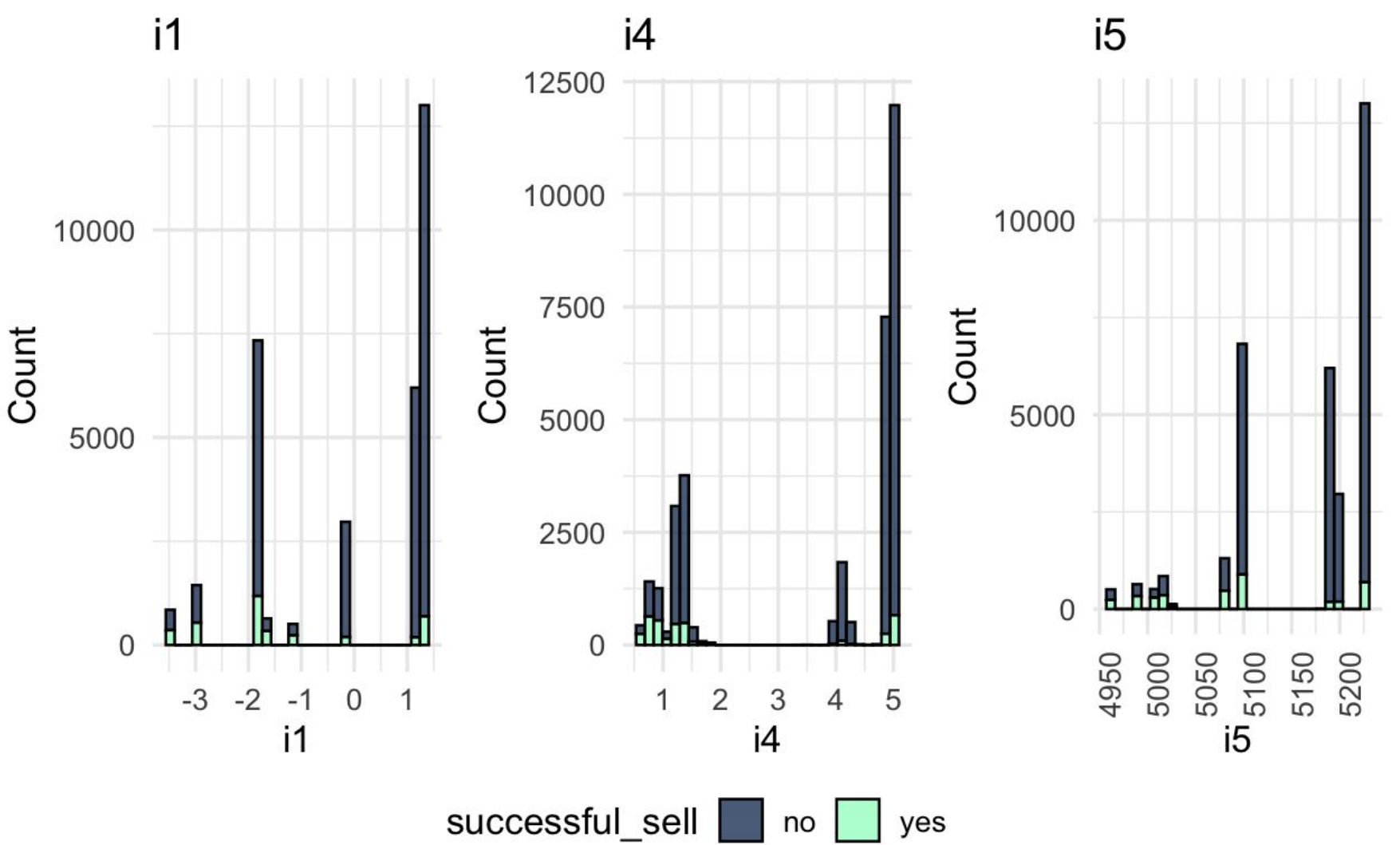
Full dataset contains 41,188 sale attempts and 22 variables of customer information along with if the sale attempt was successful

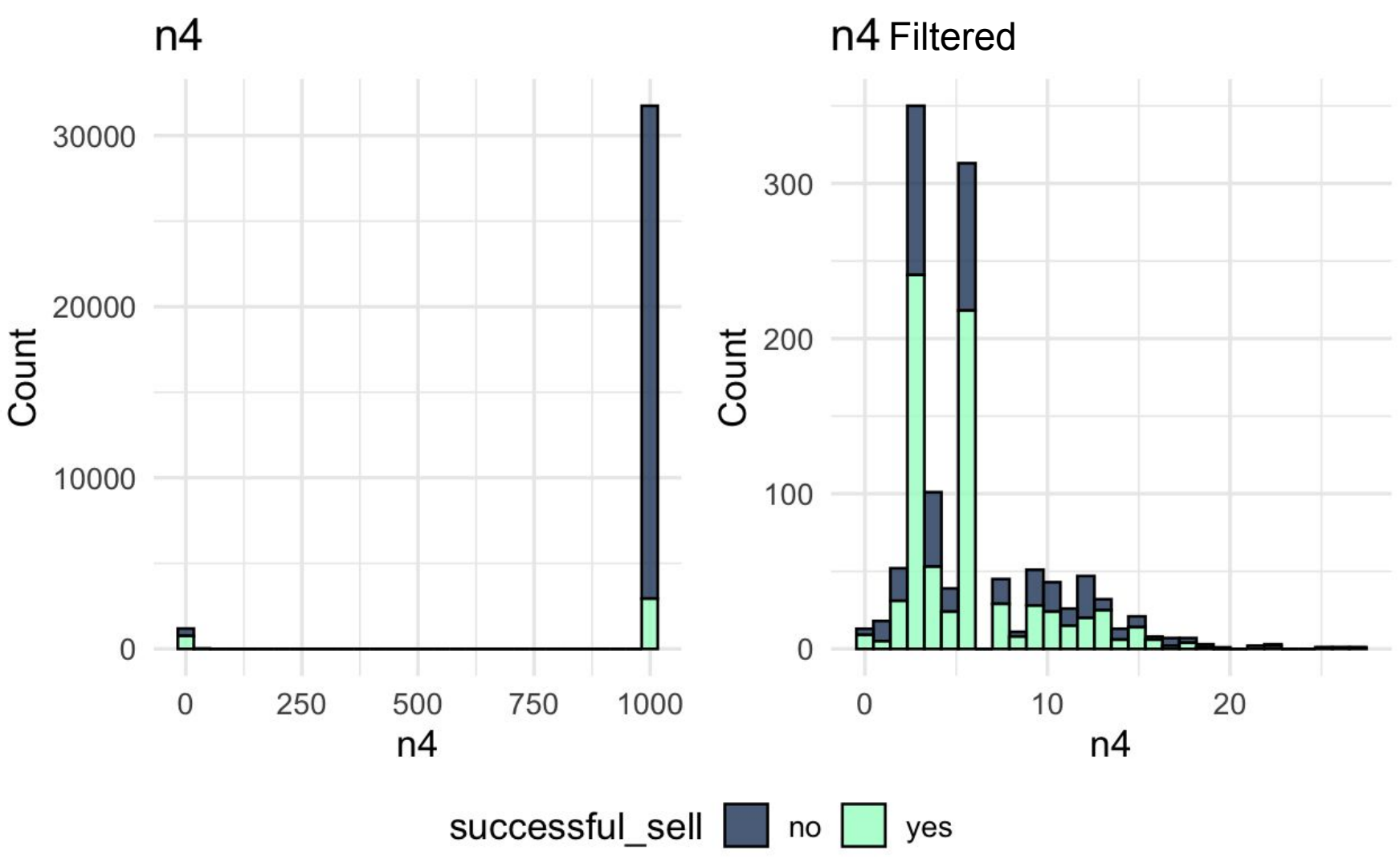
Dataset contains information like customer age, employment type, marriage status, and a lot of variables with undescriptive names like “c3”.

We put aside 20% of the dataset to use as a benchmark for how the model we build will perform on new data

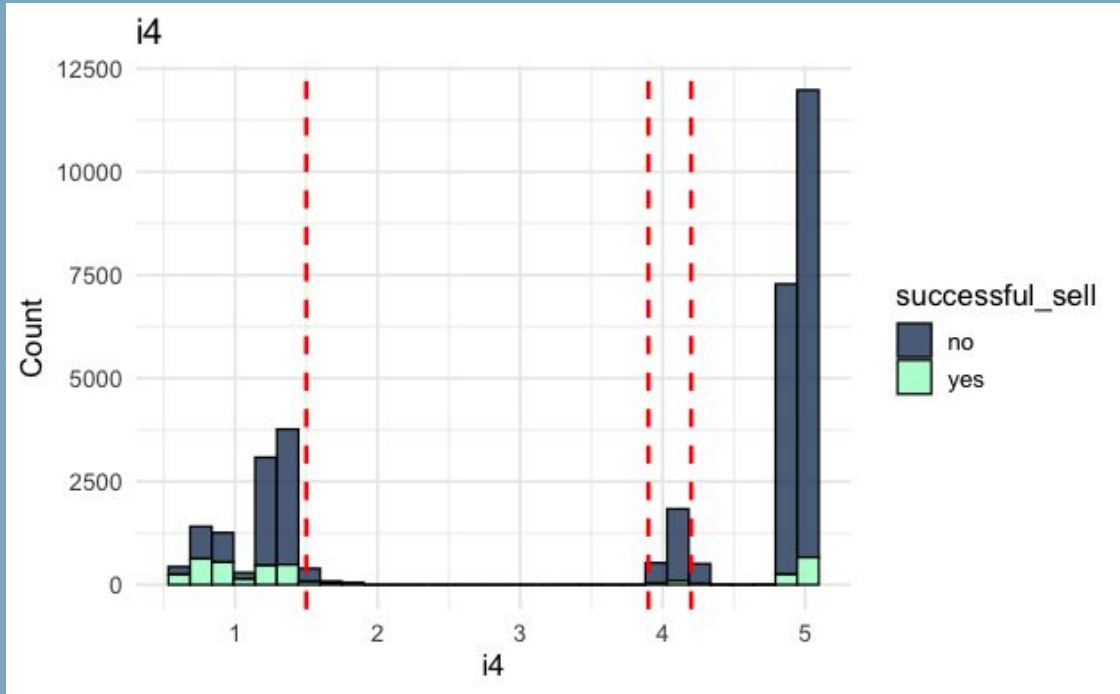

month
Percent
30%
20%
10%
0%
apr
aug
dec
jul
jun
mar
may
nov
oct
sep
month
school
Percent
30%
20%
10%
0%
0 - none
1 - almost none
2 - a little bit
3 - a bit more
4 - average amount
5 - a decent amount
5 - a lot
missing
school
successful_sell
Percent
75%
50%
25%
0%
no
yes
successful_sell
successful_sell
no
yes


c4
60%
40%
20%
0%
Percent
new
old
c4
c8
75%
50%
25%
0%
Percent
missing
no
yes
c8
c10
75%
50%
25%
0%
Percent
no
yes
c10
successful_sell
no
yes


i1
Count
10000
5000
0
-3
-2
-1
0
1
i1
i4
12500
10000
7500
5000
2500
0
1
2
3
4
5
i4
i5
10000
5000
0
4950
5000
5050
5100
5150
5200
i5
successful_sell
no
yes


n4
30000
20000
10000
0
Count
0
250
500
750
1000
n4
n4 Filtered
300
200
100
0
Count
0
10
20
n4
successful_sell
no
yes

We prepared the data to work well in a variety of different types of models



Modeling and Results

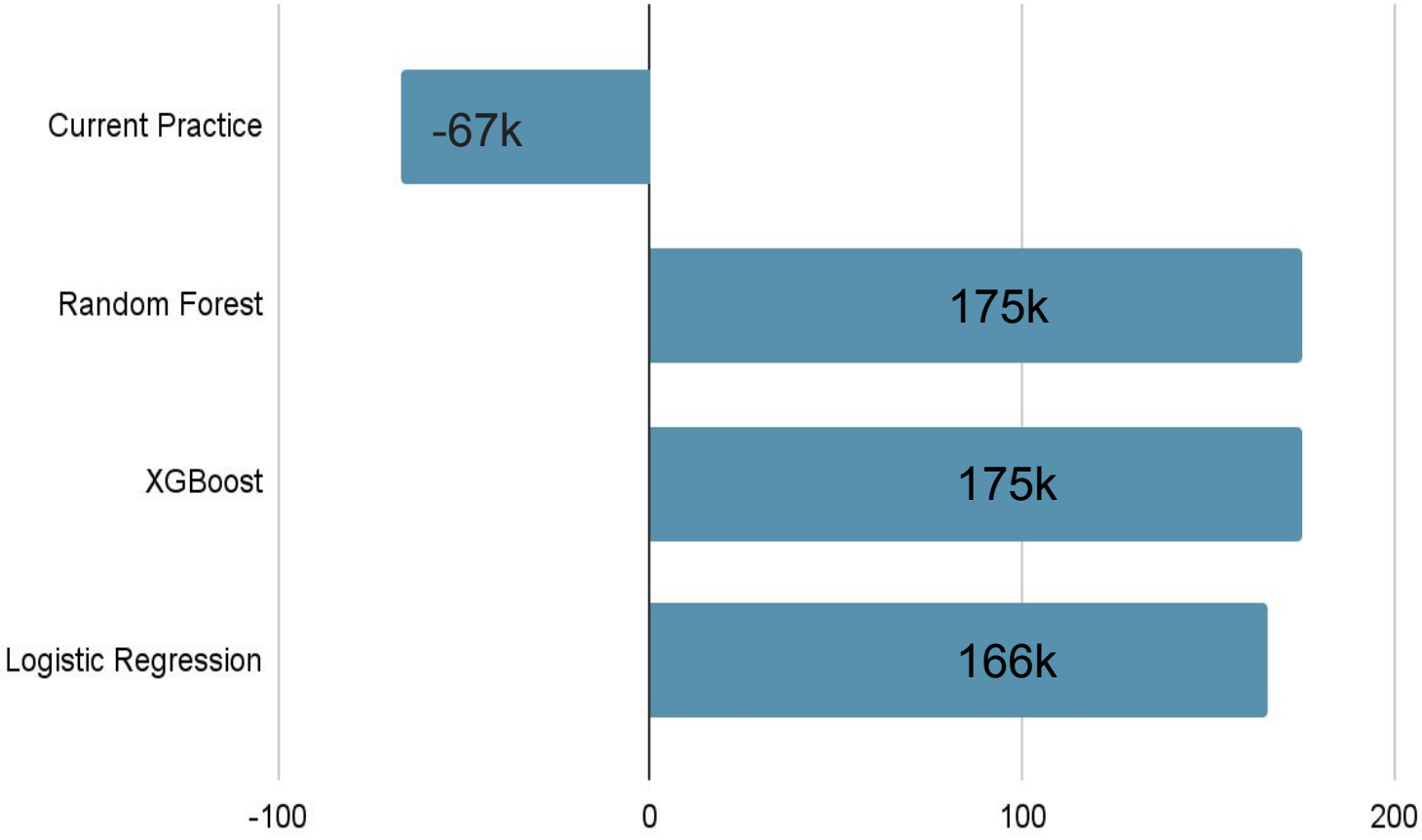
We evaluated the performance of three different models along with a “naive” model which captures the current operating practice of Tax Company

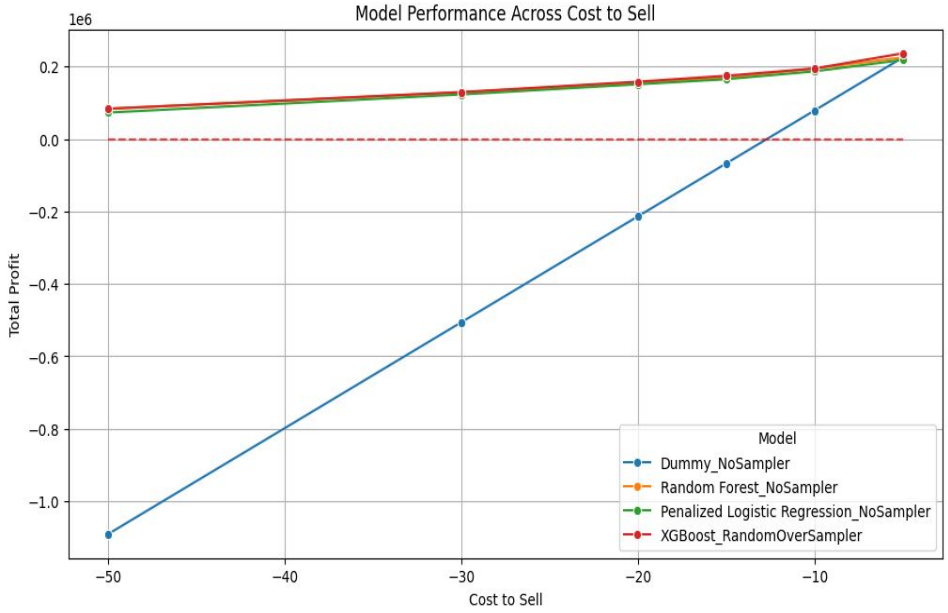
Performance of each model was optimized to maximize profit according to your current profit per sale and the cost of an unsuccessful attempt

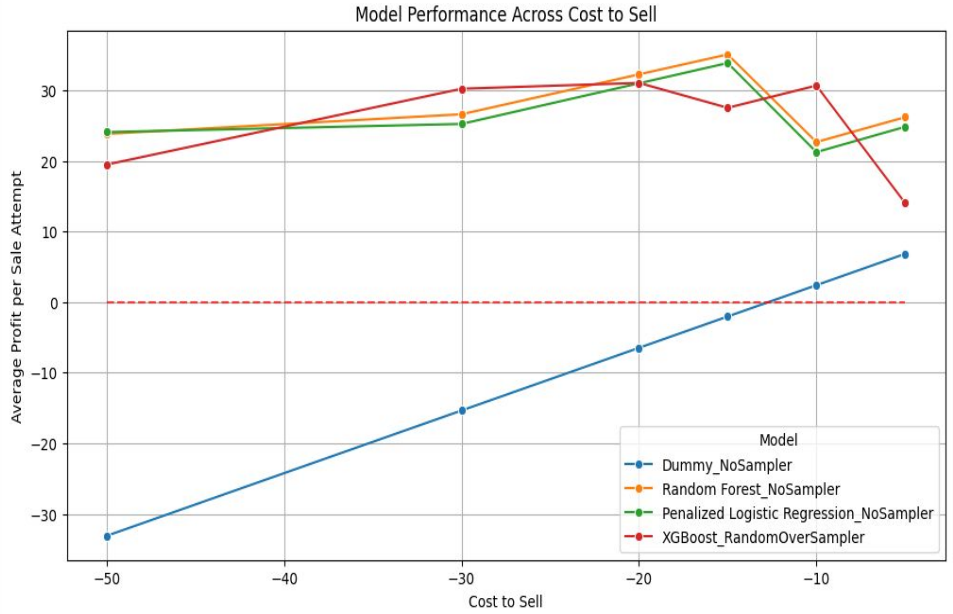
The models we used were:

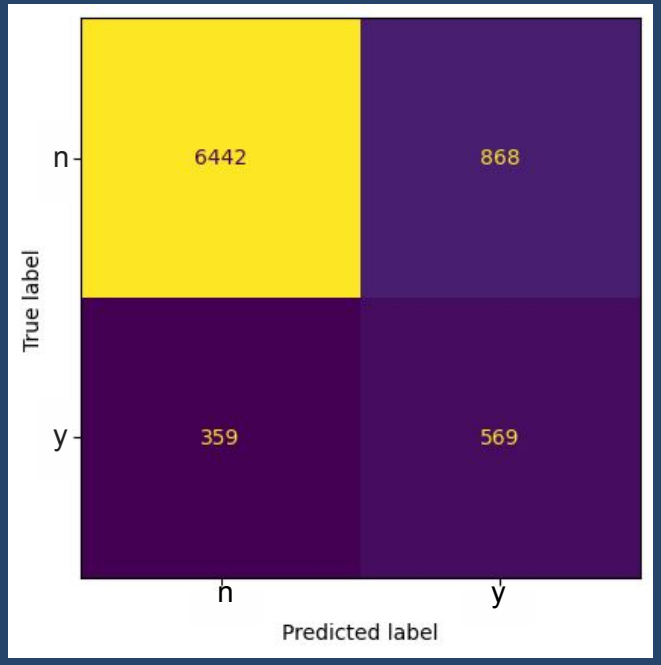
1) Logistic Regression
2) Random Forest
3) XGBoost

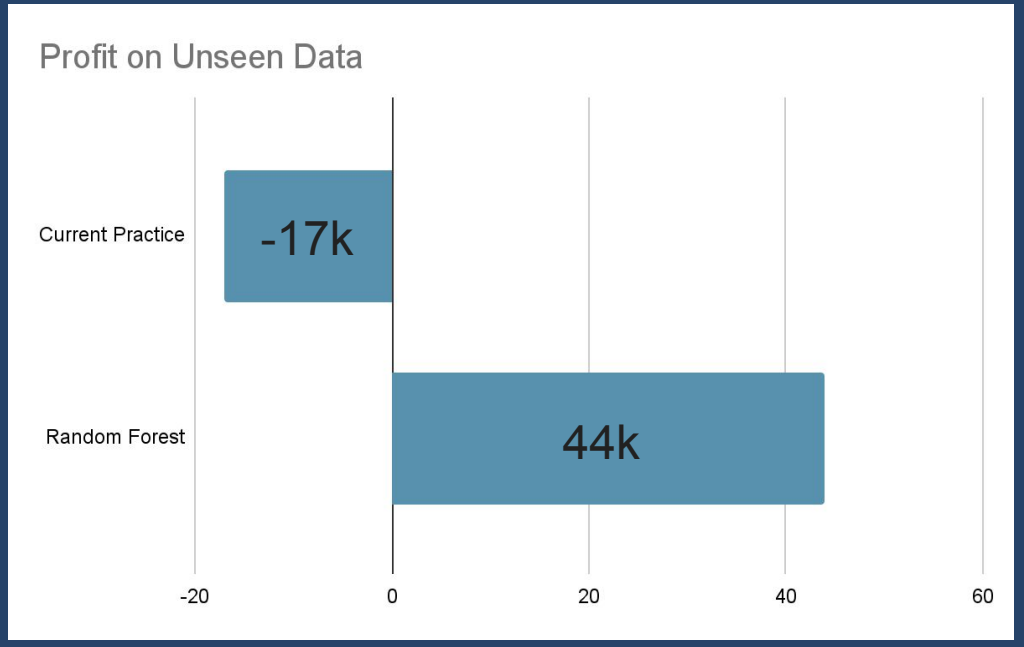
Profit on Training Set




Model Performance Across Cost to Sell
1e6
Total Profit
Cost to Sell
Model
Dummy_NoSampler
Random Forest_NoSampler
Penalized Logistic Regression_NoSampler
XGBoost_RandomOverSampler


Model Performance Across Cost to Sell
Average Profit per Sale Attempt
Cost to Sell
Model
Dummy_NoSampler
Random Forest_NoSampler
Penalized Logistic Regression_NoSampler
XGBoost_RandomOverSampler


6442
868
359
569
n
y
True label
n
y
Predicted label

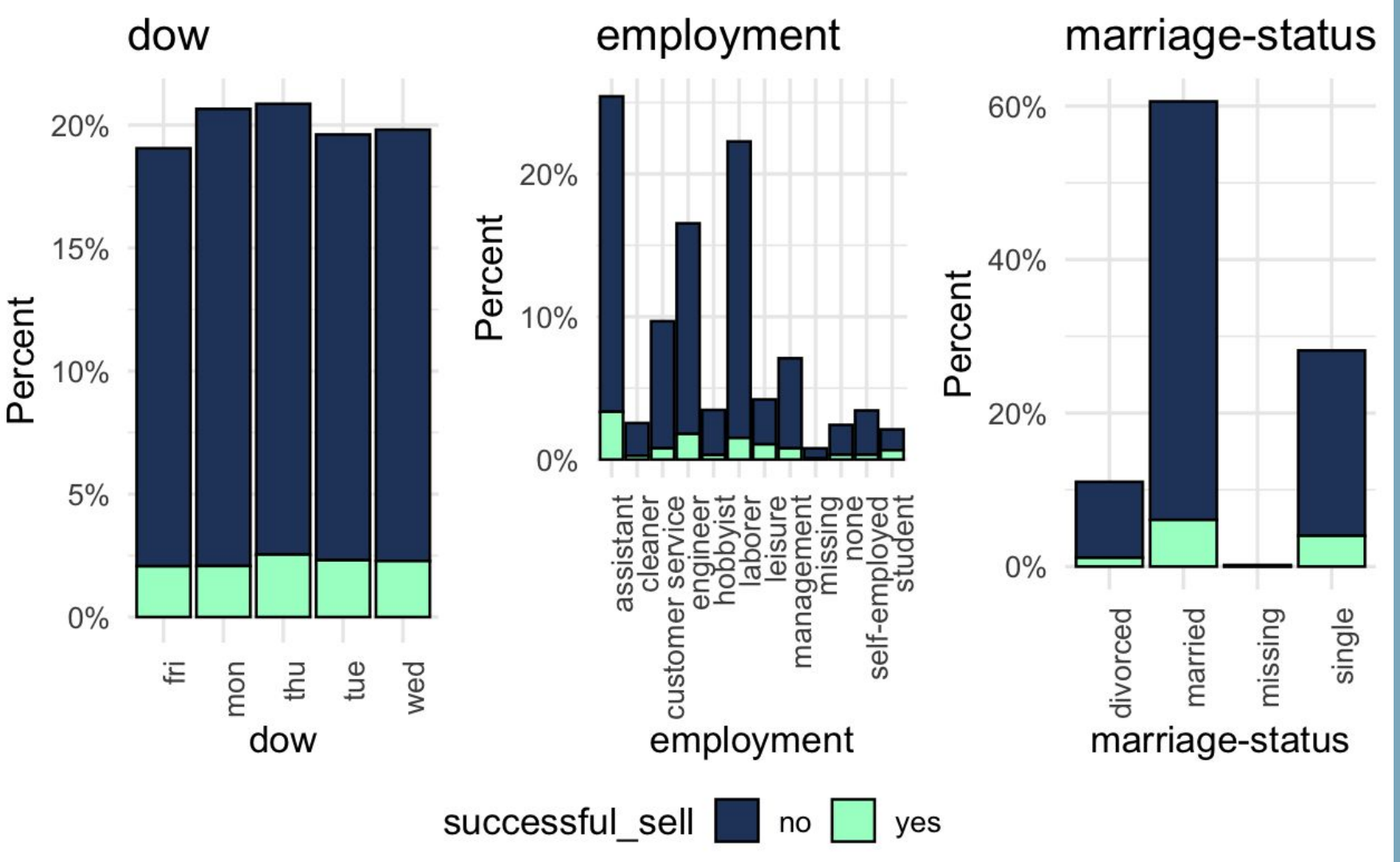

Profit on Unseen Data
Current Practice
-17k
Random Forest
44k
-20
0
20
40
60

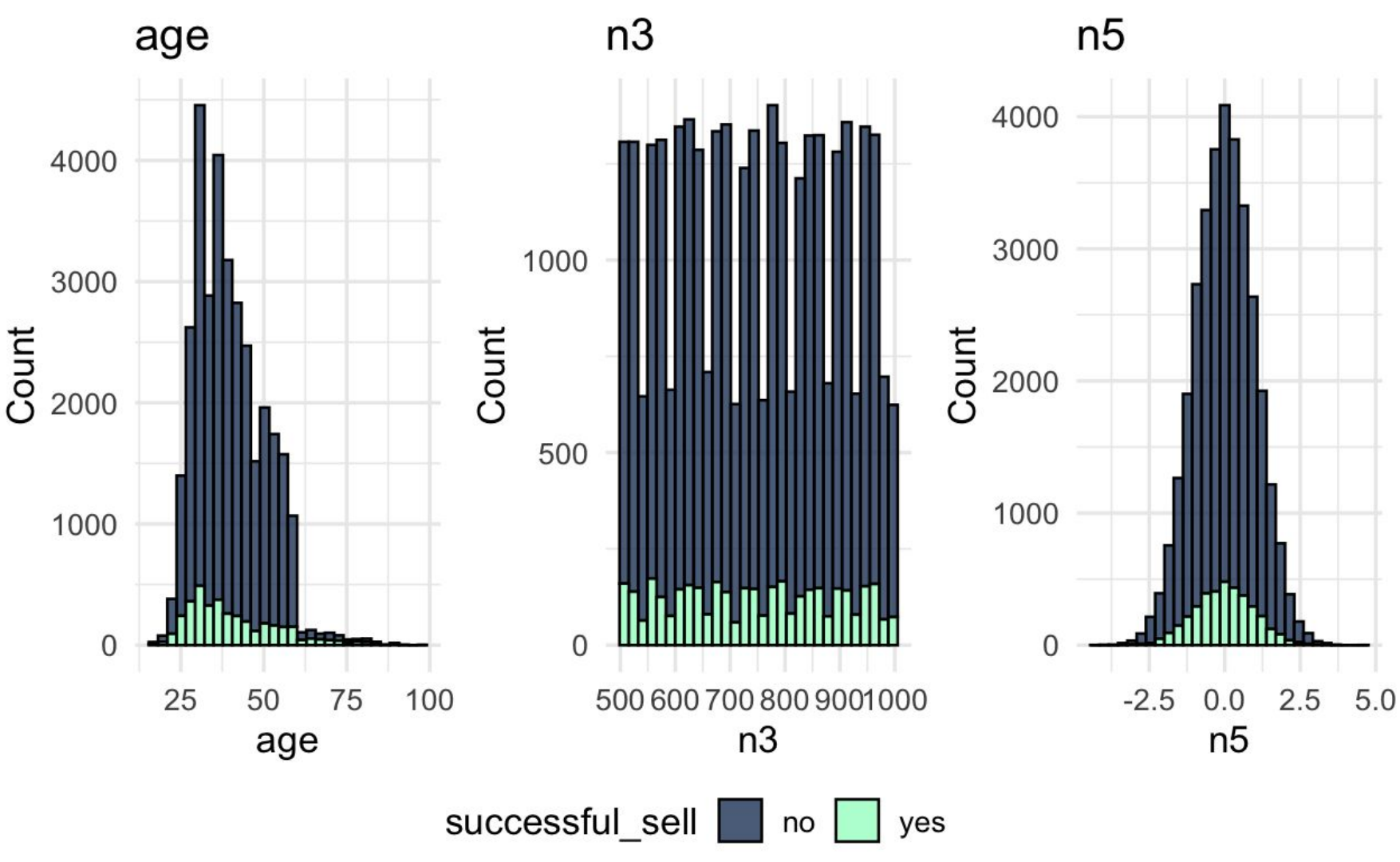
Next Steps and Places for Improvement

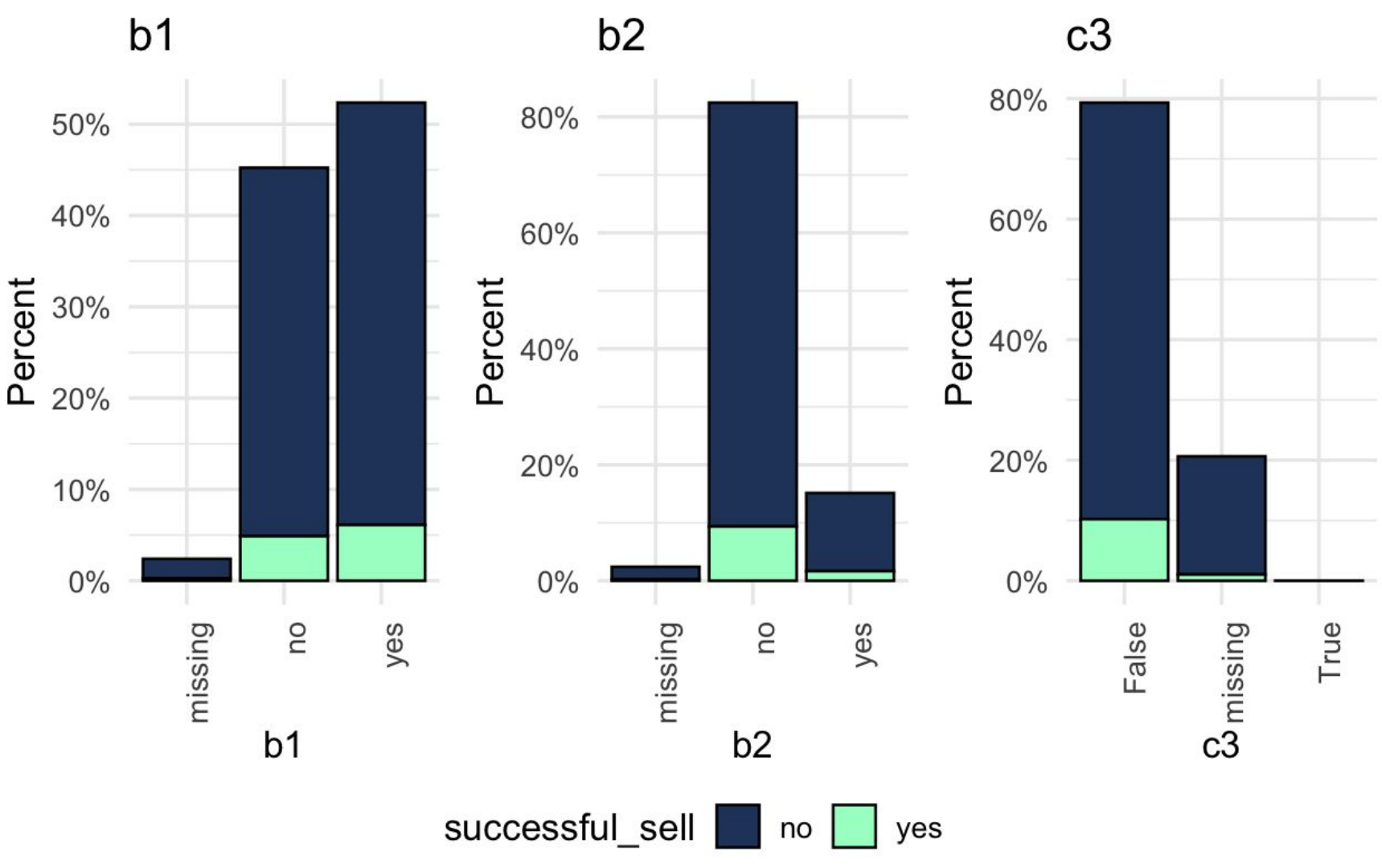
- Deploy and utilize the model so you can use it to generate likely leads
  - phData will monitor performance to ensure likely leads are identified

- Look to utilize domain knowledge and other data sources if available in future modeling
  - If we can leverage strong domain knowledge, we can construct our models to be even better

Questions???


dow
employment
marriage-status
Percent
20%
15%
10%
5%
0%
fri
mon
thu
tue
wed
dow
assistant
cleaner
customer service
engineer
hobbyist
laborer
leisure
management
missing
none
self-employed
student
employment
60%
40%
20%
divorced
married
missing
single
marriage-status
successful_sell
no
yes


age
Count
4000
3000
2000
1000
0
25
50
75
100
age
n3
Count
1000
500
0
500
600
700
800
900
1000
n3
n5
Count
4000
3000
2000
1000
0
-2.5
0.0
2.5
5.0
n5
successful_sell
no
yes


b1
50%
40%
30%
20%
10%
0%
Percent
missing
no
yes
b1
b2
80%
60%
40%
20%
0%
Percent
missing
no
yes
b2
c3
80%
60%
40%
20%
0%
Percent
False
missing
True
c3
successful_sell
no
yes