Banco Federal Das Finanças

Apresentado por

Samuel Mickelsen

Luke Briggs

Jarom Bradshaw

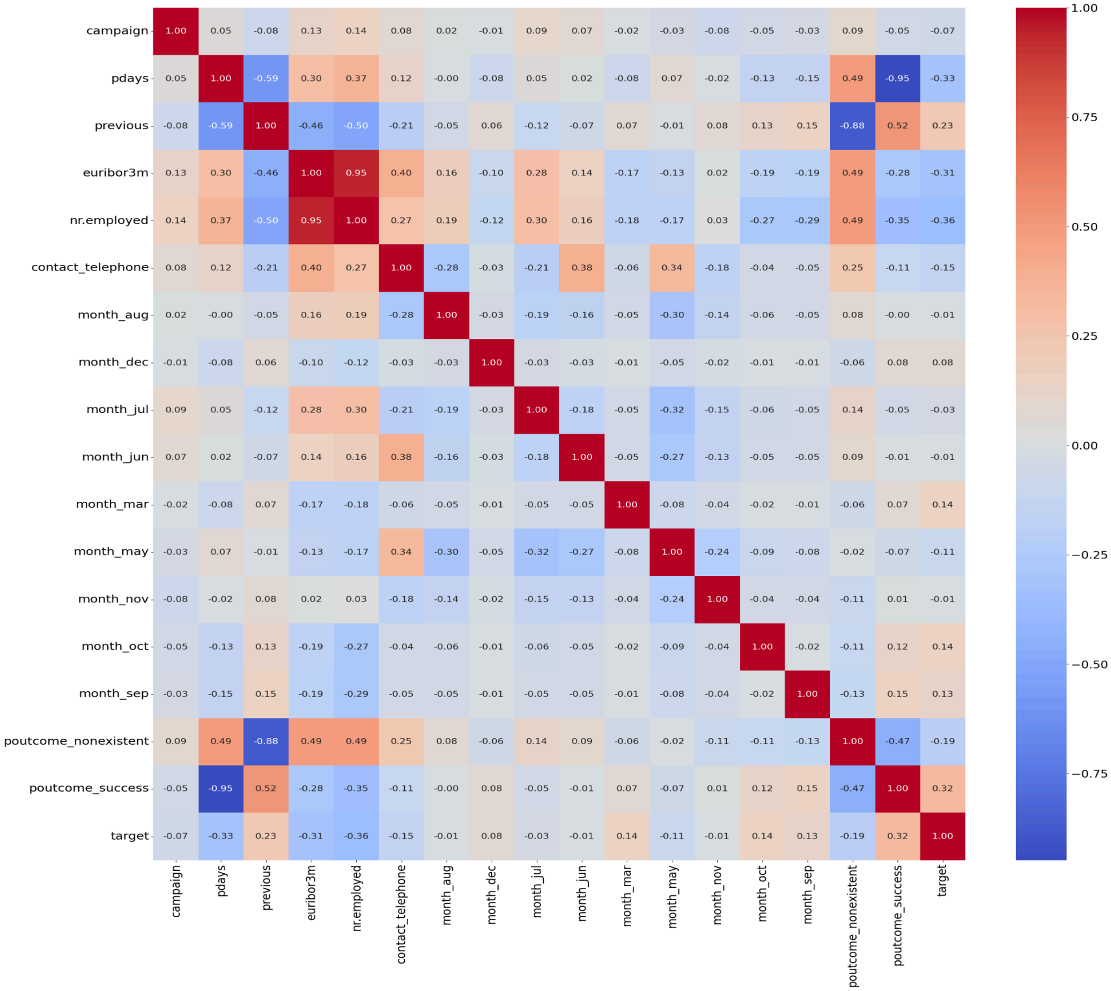
Scott O'Connor

Tyler Hilton

The purpose of this study was to create a model that predicts a person's probability of making a purchase for a new checking's account. Ideally, this model will help us decide on who and where to focus our marketing. We can report, our model is very able to predict based on the data who is most likely to create a new account with the bank.

The most influential variables in correctly predicting a person's purchase decision were these features: (features = ['contact', 'month', 'campaign', 'pdays', 'previous', 'poutcome', 'euribor3m', 'nr.employed']) Specifically, people tended to make a purchase when they had previously made one in the last campaign(poutcome). Therefore, maintaining old customers is important and helps bring more people to make a purchase then new customers. This aligns with the values of Banco Federal of having a personal approach and maintaining contact with those who have had positive results in the past.

The following is a small heat map of how strongly each of our main features impacts the model's ability to predict accurately. We used a larger version, housing all of the possible features, in order to narrow down the important points. The points that are closest to 1 and -1 have a strong correlation with our given target, so we spent a bit of time figuring out the most valuable features. The features that we discussed above seem to have the strongest correlation to the data and gave us a much stronger recall value when predicting those who will say yes to opening an account during a phone call:



(data in figure rounded to the second decimal place)

We have eliminated some of the outlier data points because it appears those points were anomalies or errors and really do not belong to the population being studied. I tested 3 separate models for the rest of the data: binning the data (with bins of equal size, bins were defined using an scikit dfunction), and were also tested under continuous, and binary methods. We found the most accurate way of testing was balancing the data and using a decision tree.

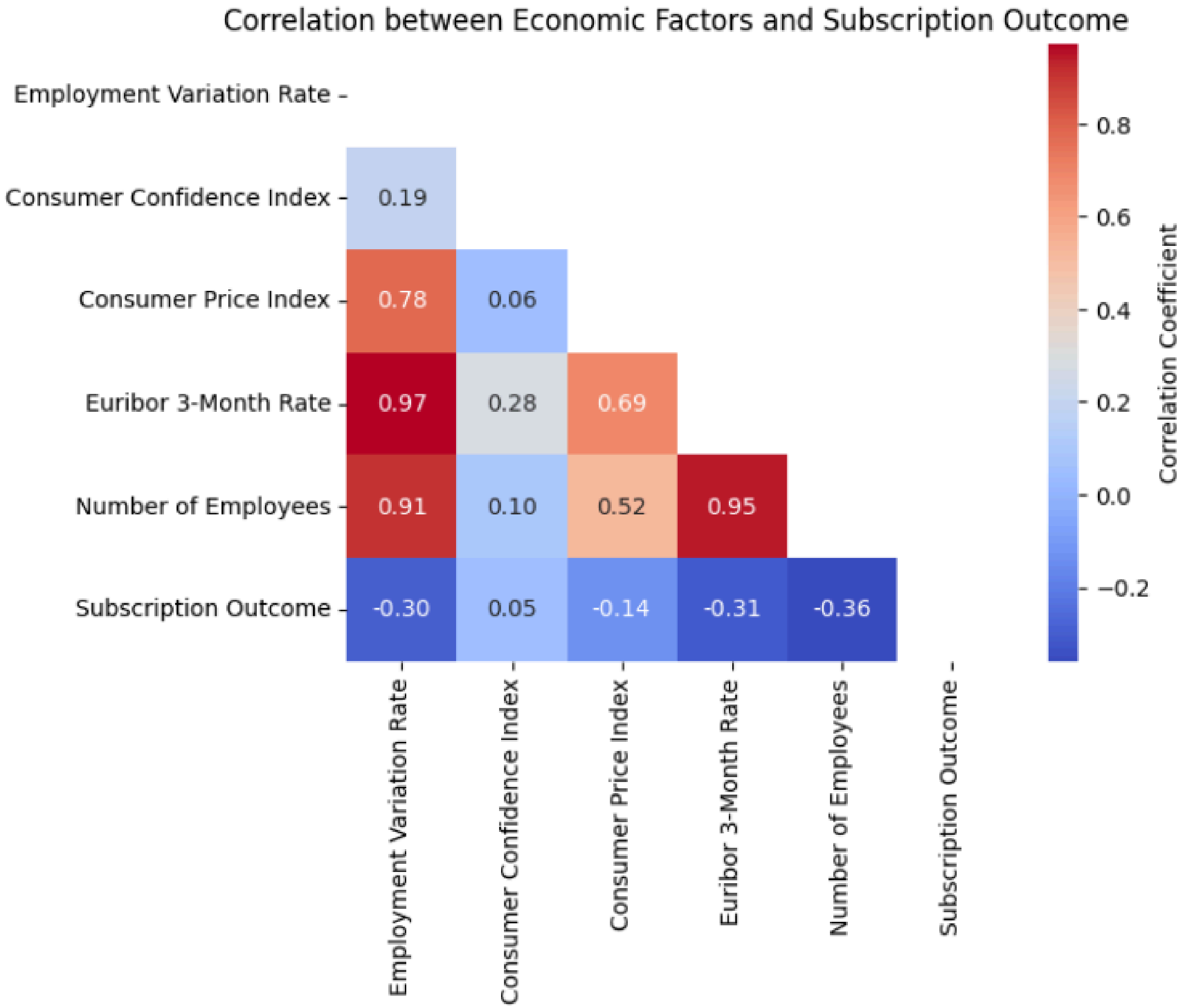
Below is a summary of the models' performance.

The features above	Overall Accuracy	Recall (what % of those who made a purchase were correctly identified by the model)	Precision (what % of purchase predictions were correct)
Continuous	.61	.52	.06
Binned	.70	.55	.27
Binary	.63	.43	.70

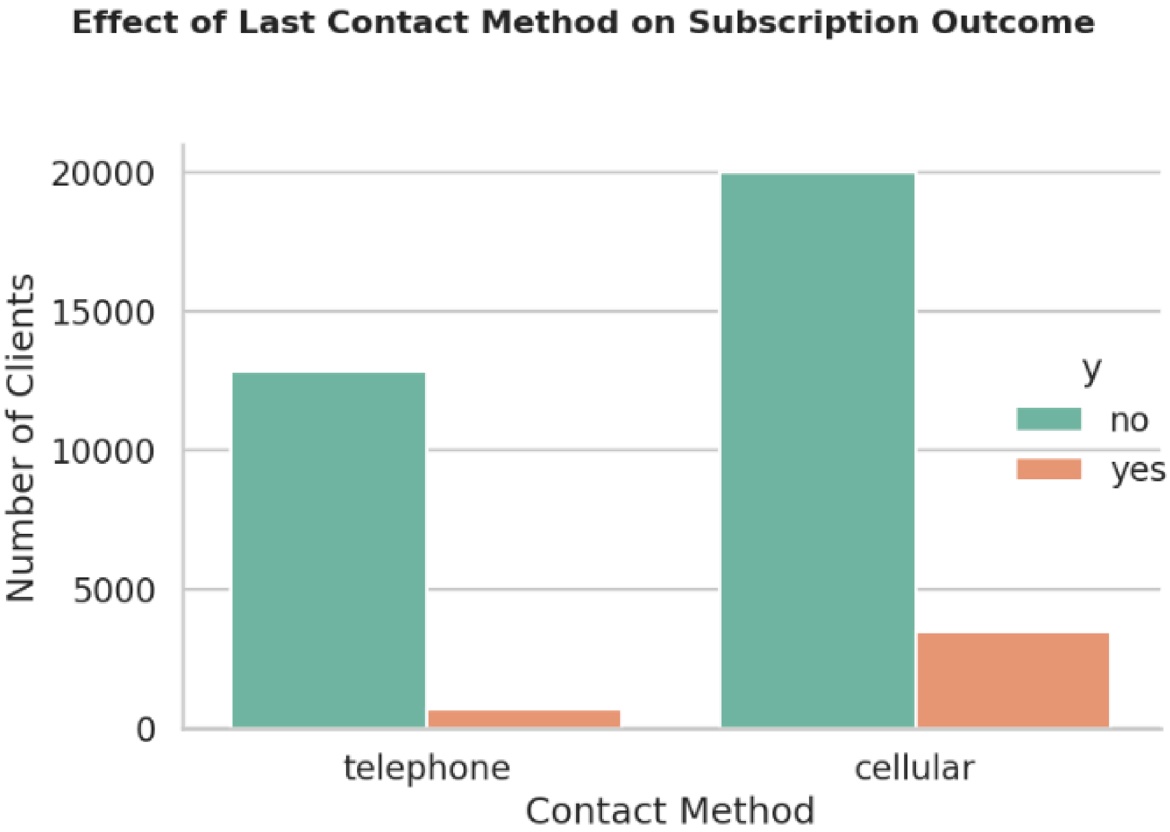
Based on these results, we moved forward with the model that uses binned values of the previously mentioned data.

Recommendations

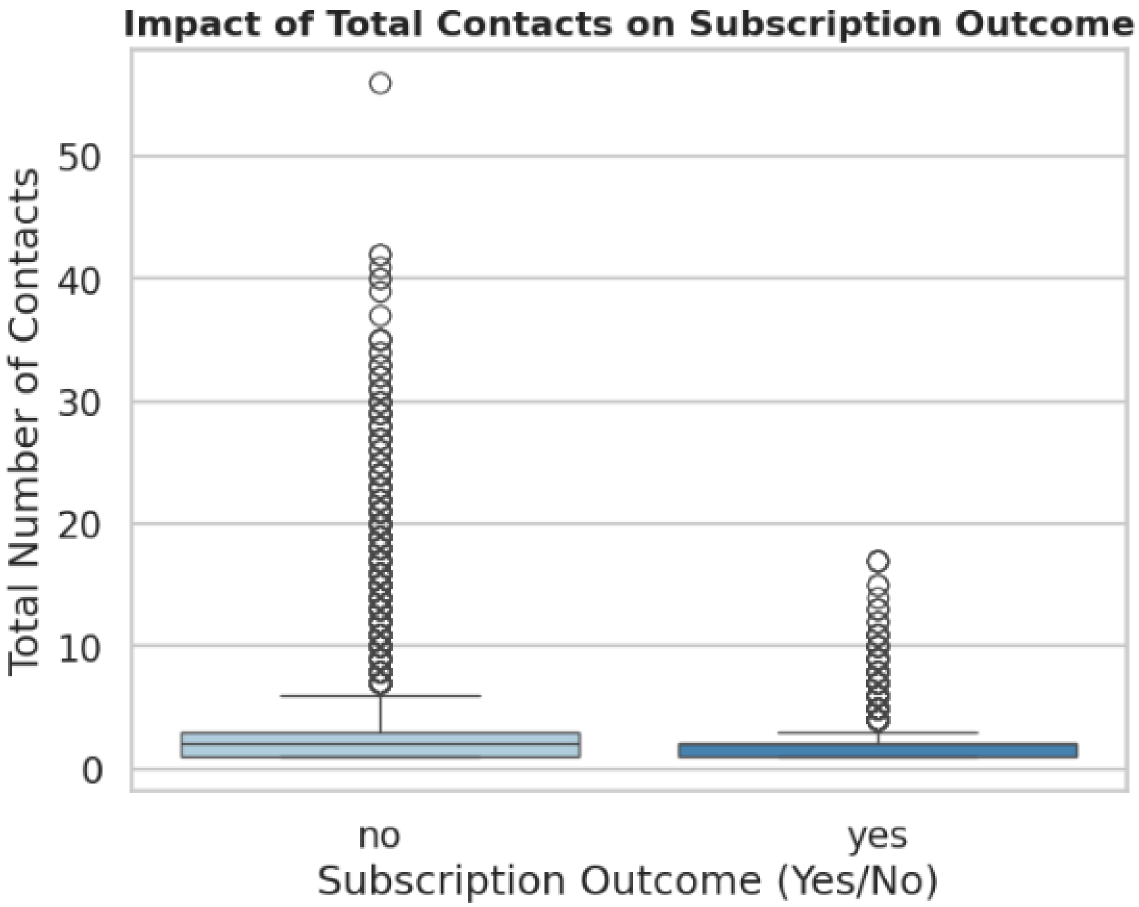
In this section, we provide our key recommendation: to leverage the highly accurate model we have trained. Given its strong predictive capabilities, the model offers a reliable approach to identifying high-potential subscribers, allowing you to optimize outreach efforts and improve overall conversion rates. We encourage the adoption of this model as the primary tool for targeting prospects efficiently and effectively. The trained model is designed to effectively predict the likelihood that reaching out to a potential subscriber will be worthwhile, utilizing a set of key features. Below are some of the critical features integrated into the model to ensure accurate predictions:



The above heat map includes several key economic features. These variables are shown in relation to subscription outcomes, with color intensity indicating the strength of their correlation. This allows for a clear visual representation of how economic factors, like inflation trends, interest rates, employment levels, and consumer sentiment, impact the likelihood of subscription conversions.



The chart above illustrates that outreach via cell phone significantly increases response rates, with nearly 25% of these responses resulting in subscriptions. This demonstrates that contacting prospects through a cell phone is far more effective than using a home-line.



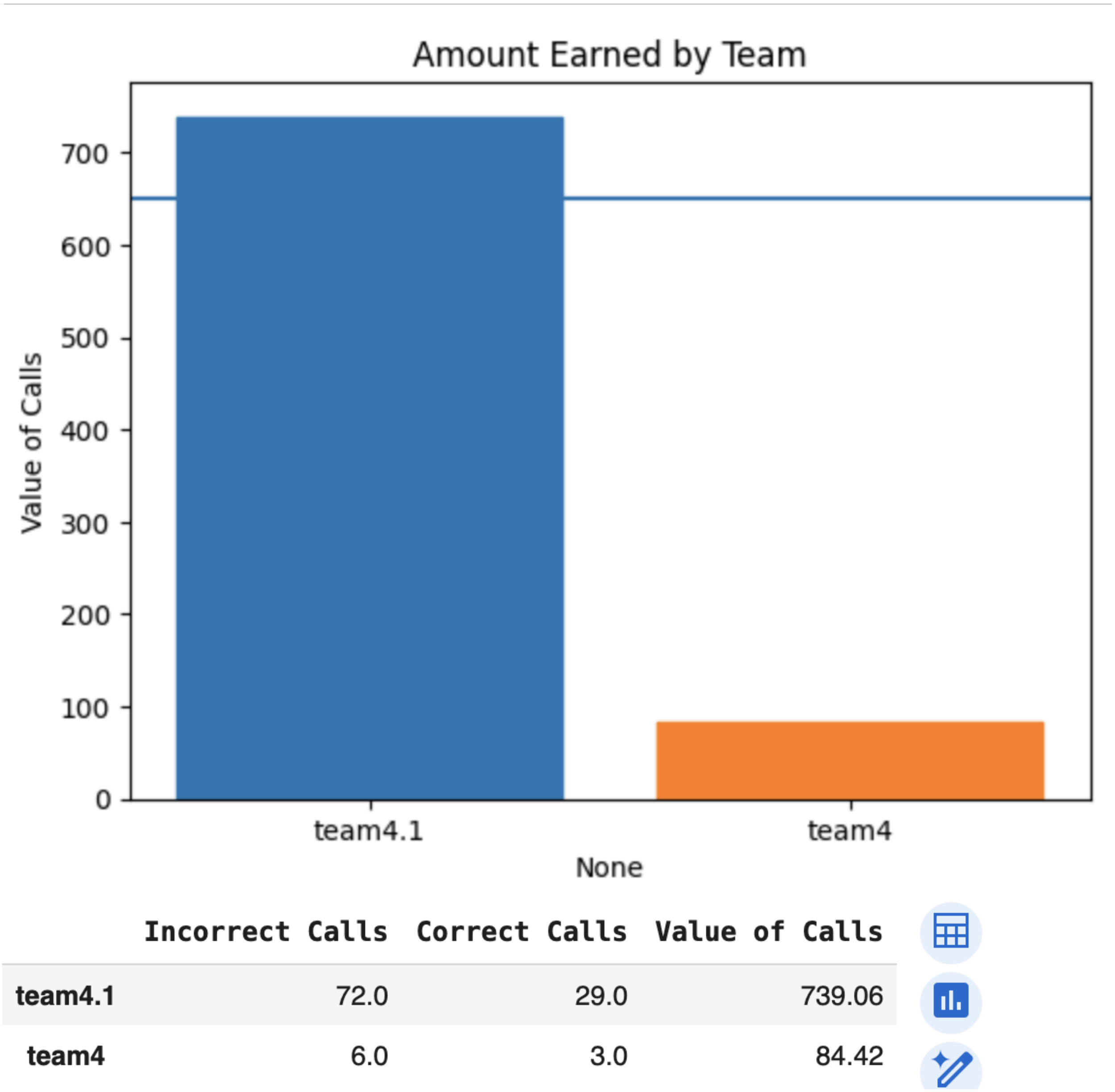
Increased contact frequency with potential subscribers does not necessarily correlate with a higher likelihood of conversion. It is essential to avoid wasting time reaching out to individuals who have already declined multiple times. Instead, the focus should be on those who have previously shown interest or have only been contacted a limited number of times.

Methods

Three separate models were created based on how outcome or y-variable was treated. Not all available variables were included in the model because they did not have a significant relationship on the outcome of how accurately we could predict which people would say yes to open an account if we called them.

A random forest model was used in each case, with a max-depth of 5 layers to prevent overfitting and to keep things simple. Various values for this parameter were attempted, but somewhere near 5 seemed to hit highest recall levels without overfitting. No appreciable difference was noticed for values of 4 or 6 when using 5-fold cross-validation. Over sampling was used on the training dataset in order to not bias the model towards a “no” prediction.

Simple decision trees were also attempted but consistently performed worse (5-20% less recall of “yes, purchase made”).



	Incorrect Calls	Correct Calls	Value of Calls
team4.1	72.0	29.0	739.06
team4	6.0	3.0	84.42

the

How we made the most correct model

At the start we took the raw data and ran it through a few tests to see what we were dealing with. After documenting all our findings, we prepared our fist ML model. This first model was not special. As expected, it was really good at predicting correctly the customer who would not create a new account, (around 98% precision) but when it tried to select customers who would like to create a new account it was beyond bad, only getting one out of four. This changed how we looked at the data.

Two of our findings made the biggest difference in making our model better. The first was to balance out the data. Because the dataset contained about 6x more data about the "no" cases we needed to level that out, to save ourselves from over-training the model. The second big difference was made when we limited our feature selection to only include data correlated with the outcome we were looking for. This refined us from 19 features to just 8. The combination of these two changes, along with a few others, increased our models ability to predict correctly from only 25% of the time to 69% of the time! An extremely noticeable difference shown by the chart above.

The process of getting our model as good as it is now took a lot more steps than we are able to comfortably list here. Let's leave it to this, in an effort to create the ultimate model best suited for this task, we had to fail a lot. We had to build a lot of models. We had to test many different new tools. In short we went above the standard to create an unusually good model for our clients.

Questions from the team discussion section

1. Based on your initial analysis of the data, your team feels:

A. simple 80/20 split will provide us with enough to accurately train and test our model.

2. Based on your initial analysis of the data, your team feels:

B. This is anonymous data, so we should be just fine.

3. Is there a significant relationship between the day we contact someone or the month we contact someone and the success rate of outcome?

C. There does not seem to be any significant correlation between the day or month someone was contacted and the success of the outcome. We do hypothesize that the time of day is possibly more important than the day or month that a call is made.