Client Report - Bank Term Deposits

Marketing Campaign Analysis

Course CSE 450: Landon Davis, **James Hall**, **Brad Strange**, **Jacob Ferris**, **Keaton Kesler**

This report helps a bank determain the likelihood of a customer signing up for a bank term deposit. The data is stored in a bank data CSV file. We then took the features and data and created models and charts to show the diffrences in the probabilities of someone signing up for a term deposit.

Model

I.Model Features

Our model features include education, day of the week, job, marital status, housing, loan, contact, month, and poutcome. Education and day of the week were combined into numeric values, the other features were one-hot-encoded. Default was not included due to having only 3 contacts known to be in default, although evidence suggests that consumers who are known not to be in default are more likely to subscribe to a deposit than those for which default status is not known.

II. Model Chosen

We tested a couple of different models for our predictions: a decision tree, and a gradient boosting classifier. We programmed the decision tree with parameters that made the possibility of overfitting unlikely. To achieve this end, we decided that the tree would only make a branch if the information gain was sufficient and if the number of samples was sufficient to make statistically significant splits likely. The values we used for this purpose were calculated from the number of features and the size and entropy of the dataset. Although this model slightly underperformed the other model, it was used to help us understand relationships in the dataset. Our primary model is a Gradient Boosting Classifier, which combines the results of multiple decision trees into a single model. This model performed the best of all the models we tested. Due to its complexity, this model was also easier to tune the tradeoff of accurately guessing “yes” values and accurately guessing “no” values. After getting the probabilities from the model, we classified all data with at least an 8% probability of being “yes” as yes, and all data points with a less than 8% probability of being “no” as no.

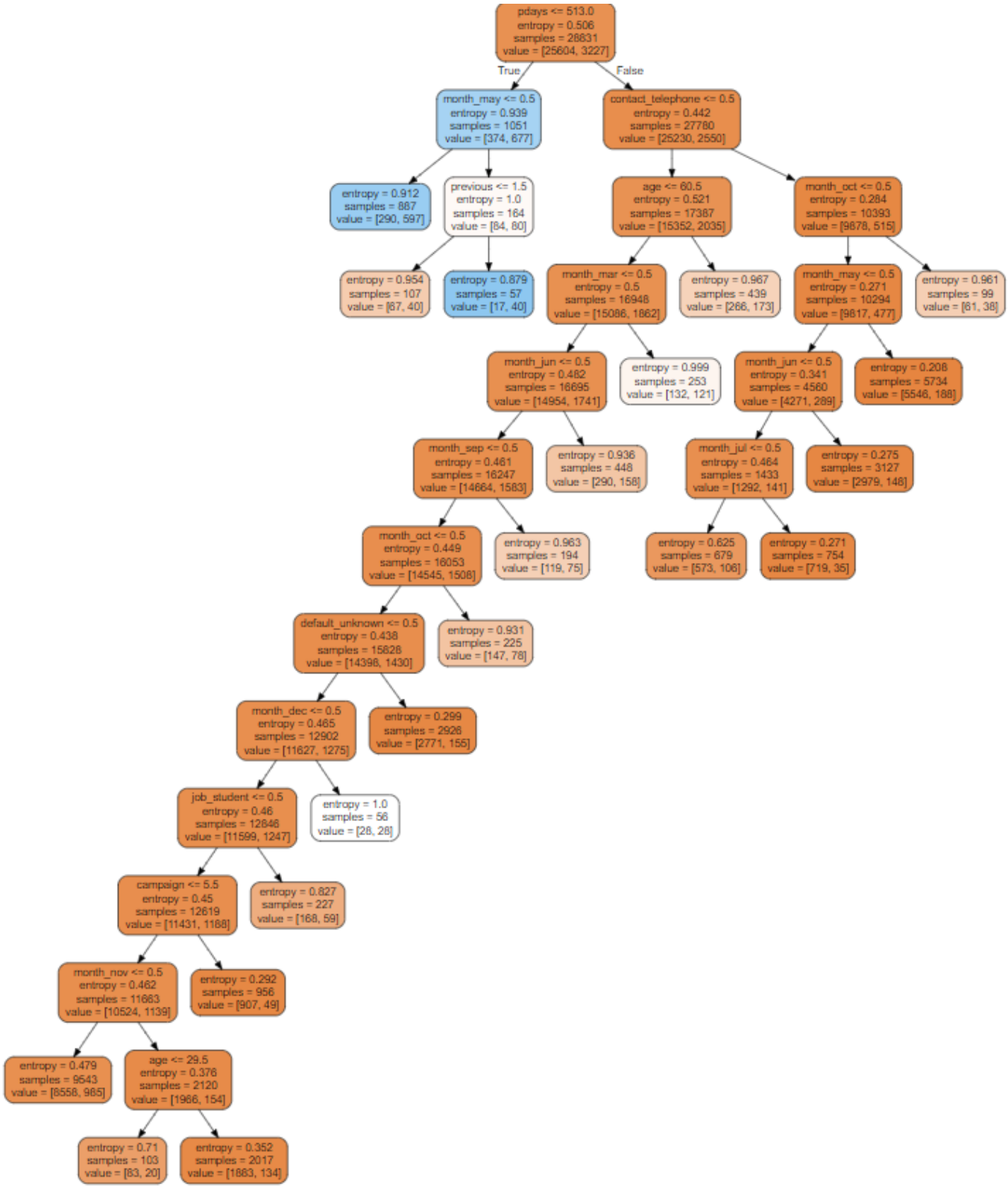
III.Model Testing

We used a 20% / 80% split, with 20% going to the test data. This test set has 930 “yes” values and 7308 “no” values. Our model keeps 78% of the “yes” values, although this number may be off by up to 4 percentage points, based on a 95% confidence interval. Our model successfully throws out 55% of the “no” values, and this figure is unlikely to be off by more than 2 percentage points. Although only 18% of the data selected by our model are “yes” values, this is 1.6 times the rate given by sampling the data at random, and although there is some uncertainty in the dataset, we are extremely confident that this ratio exceeds 1.4. The consistency of the model being accurate in the future depends on the assumption that the data will be about the same as it has been in the past. This may or may not be accurate. We also tested economic conditions. This improved model performance in some cases, but this was not statistically significant.

TECHNICAL DETAILS

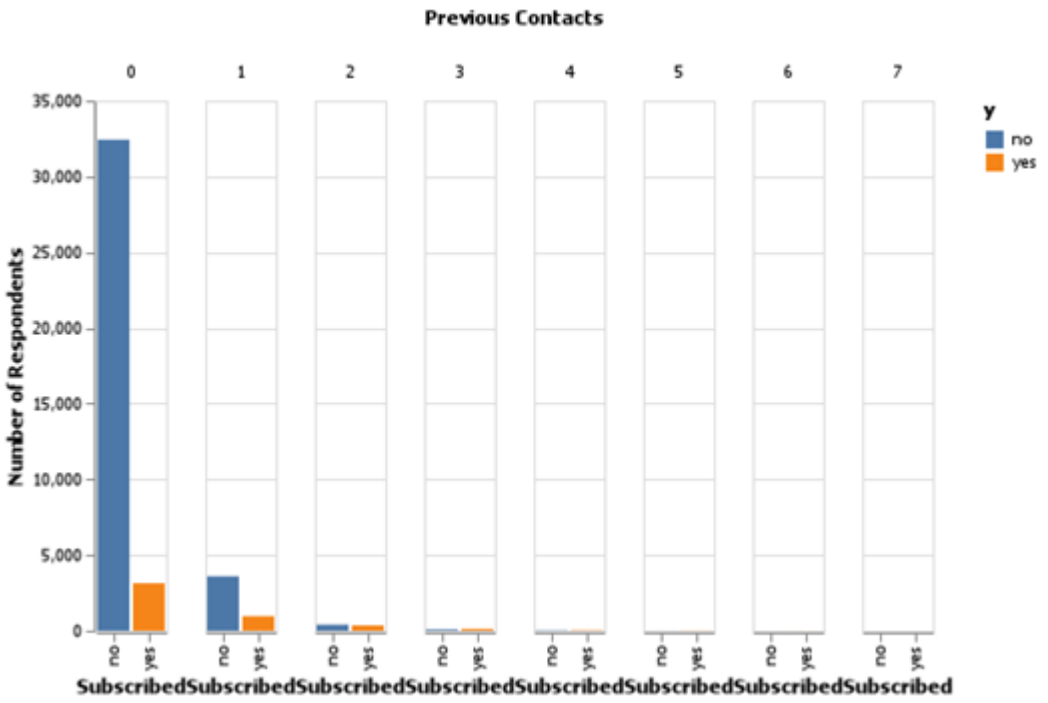
```
campaign = pd.read_csv('https://raw.githubusercontent.com/byui-cse/cse450-
course/master/data/bank.csv')
campaign.info()
features = ['age', 'job', 'marital', 'education', 'default', 'housing', 'loan',
'contact', 'month', 'campaign', 'pdays', 'previous', 'poutcome']
X = pd.get_dummies(campaign[features], drop_first=True)
y = campaign['y']
# split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=1)
# fit decision tree
tree = DecisionTreeClassifier(criterion='entropy', min_samples_split=87,
min_samples_leaf=29,min_impurity_decrease=0.00037, random_state=1)
tree.fit(X_train, y_train)
# predict on test data
y_pred = tree.predict(X_test)
```

Early Version of Decision Tree:



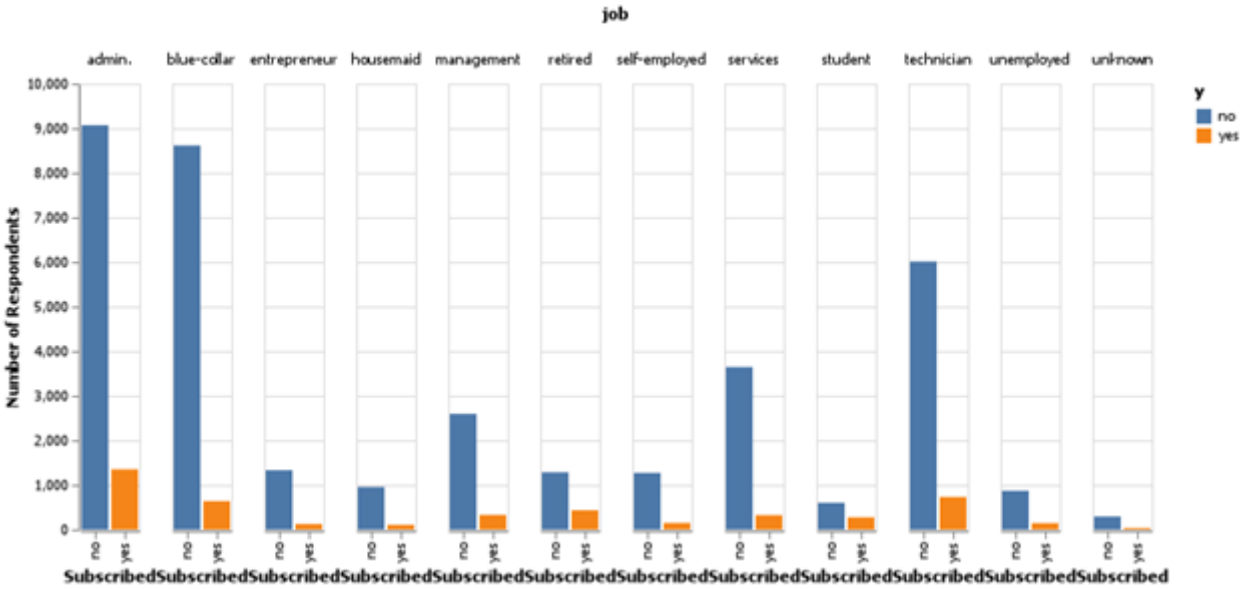
Visualizations

Previously Contacted:



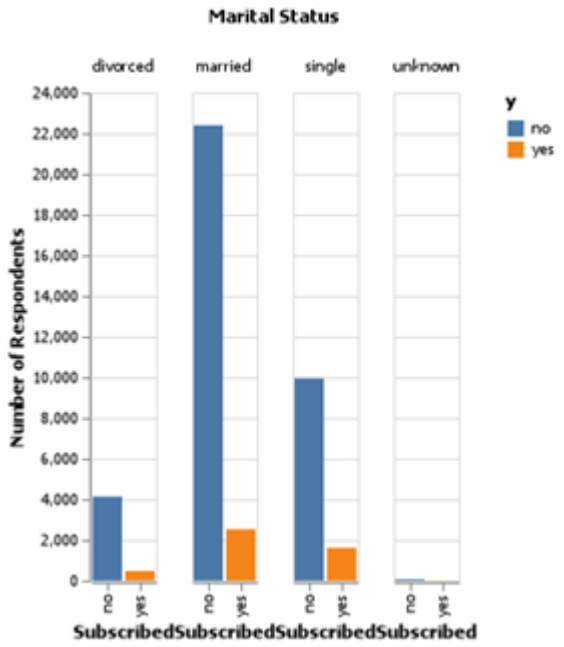
This chart is showing how people who had been contacted a certain amount of times responded to the phone call by subscribing or not. As we can see here, most of this sample had not been contacted during the previous campaign and no one had really been contacted more than 3 times. What we can see here though is that those that were contacted once during the previous campaign were more likely to subscribe then those that had never been contacted.

Job:



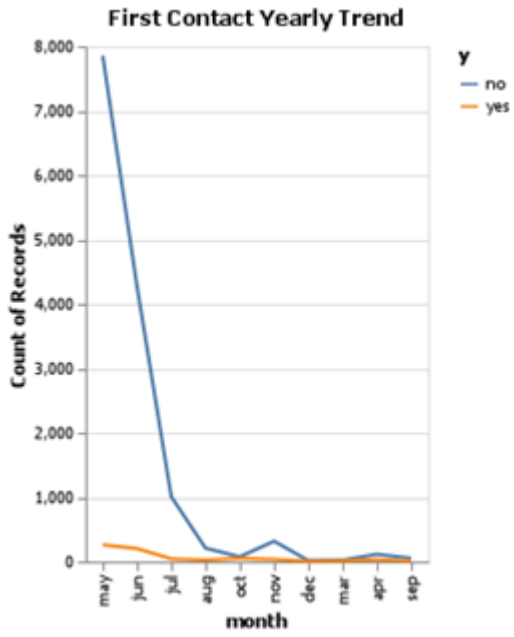
The above chart here displays how people subscribed based on their job. Although students are a smaller sample size, it would appear students and retired groups have the best ratio in subscribing. You may want to avoid entrepreneurs and unemployed people as they have very small subscription rates.

Marital Status:



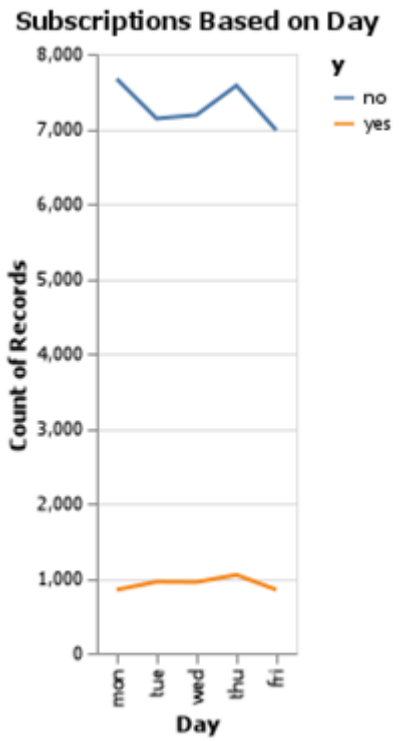
Marital status also seems to have some effect on subscription rates. Single individuals are more likely to subscribe than married or divorced people. They are over twice as likely to subscribe as married customers.

First Contact:

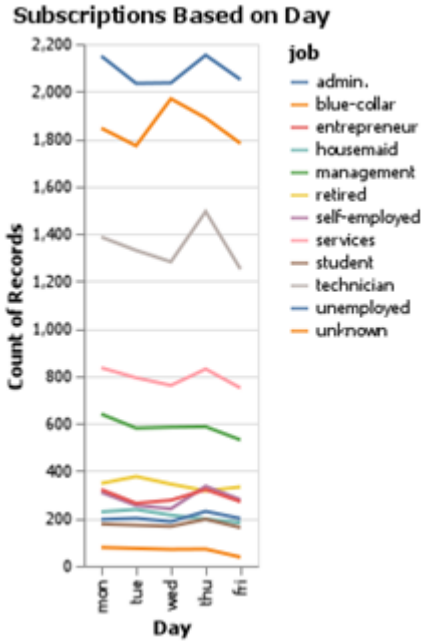


This is a chart showing the subscription rate of new customers that hadn't been contacted and the month they were contacted in. We can see most contacts occur in may and the contacts decline throughout the year. We also seem to get the most yes values in May and June and then it pretty much hits zero as you get later into the year.

Subscription based on Day:



This shows what day of the week customers were contacted on and their subscription rate. We can see a pretty flat line for the yes line with a slight spike on Thursday. There is a spike on Thursday for both lines though meaning just more contacts happen on that day as a whole. So there may not be a correlation between the day contacted and subscription rate but if anything, Thursday would be the best day. If you look at it from a per job basis in the chart below you will see that most jobs do respond better on Thursdays with a dip on Tuesdays and Wednesdays except for a select few like the blue-collar workers.



Discussion Responses

Below are our final answers to the case discussion questions:

Do you think a supervised or unsupervised approach would work best for this situation?

- We wanted to go with a supervised learning approach because we have a target value column which allows us to train our model based off of the previous values.

How much of that data will you use to train your model?

- We used 80% to train our model and 20% to test it. Using a larger portion of data to train our model could influence our model to overfit because of the imbalance in the dataset.

I'm wondering if it's possible for us to see if those results are true for all customers, or if some types of customers respond better on certain days than others?

- Most jobs do respond better on Thursdays with a dip on Tuesdays and Wednesdays except for a select few like the blue-collar workers.

What do you think we might need to do for this project in order to be compliant with GDPR regulations?

- Our data does seem to be GDPR compliant. Our data is anonymous and cannot be connected to an individual. If the dataset contained names, we would want to remove them before sharing that dataset or model with anyone.

APPENDIX A (PYTHON CODE)

Python Notebooks:

Below are Github Gist links to the notebooks we used during this case study:

- https://colab.research.google.com/gist/jamesbhall423/e757aebcdb4d90cd8089cd91c72d0a9e/copy-of-starter_bank.ipynb#scrollTo=_QF7W0nwOVyo
- https://colab.research.google.com/gist/jamesbhall423/2d3f75f5ad96f0c526aaadf5690d4c25/copy-of-starter_bank.ipynb#scrollTo=up5qYnnm9XLi

Economic Test Gists:

- https://colab.research.google.com/gist/jamesbhall423/668ad906cf2d64675bee9e8efe849a74/copy-of-starter_bank.ipynb#scrollTo=g-b0ouV81DFZ
- https://colab.research.google.com/gist/jamesbhall423/1d610c116e70f8ba497ea2feddd4914d/copy-of-starter_bank.ipynb#scrollTo=I9h73x0PhfJi

Retest Gist:

- https://colab.research.google.com/gist/jamesbhall423/2b7abe1a3da9d6b032fa3bbec7ccfd95/copy-of-starter_bank.ipynb#scrollTo=SybIFUItrLZ0

CSV link:

- https://colab.research.google.com/drive/1dTmigW1F0ovhseYnp1vIu9ebq8hv_oSb

Model Persist Link:

- https://drive.google.com/uc?export=download&id=13ewxymF0a-8whbotghcHQ8e-lS5J9C5C