Part 1: Exploration

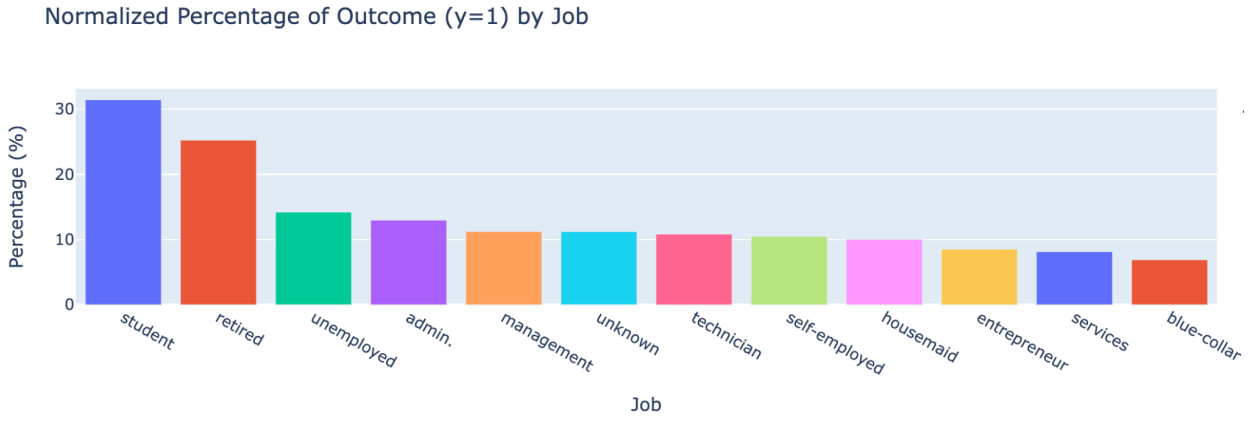
The data set analyzed includes 41,118 entries (rows) and 20 columns, 19 of which are customer features and one of which is the target outcome ‘y’, whether the customer purchased a term deposit or not.

‘Default’, ‘housing’, and ‘loan’ columns have missing values. ‘Default’, whether the customer has credit in default, has 20% of its values missing, thus this column was not used in the analysis. Feature ‘pdays’, the number of days that have passed since last contact of a customer, and ‘previous’, the number of previous contacts, indicate that most customers have not been contacted in previous campaigns.

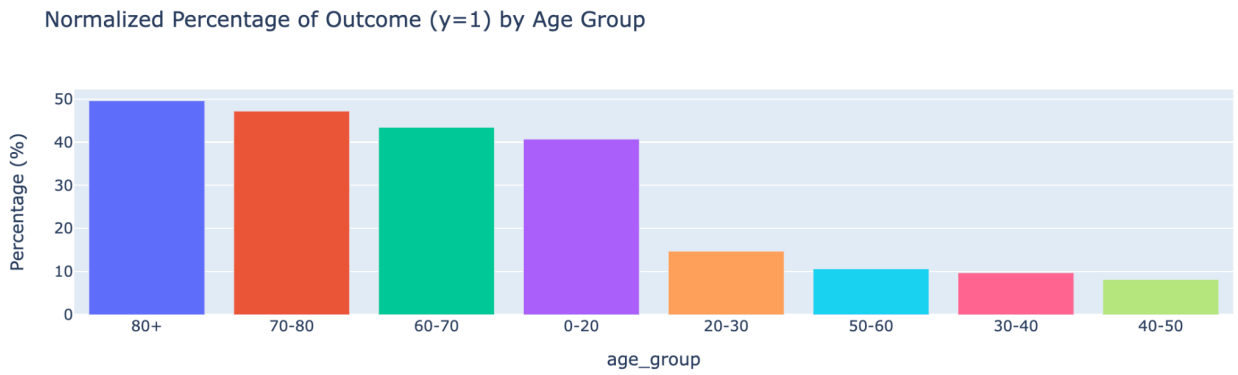
‘Age’ was organized into bins and relabeled ‘age_group’ for ease of analysis.

Each feature was evaluated for variance of ‘y’ by calculating the percentage of ‘y’ = 1 and ‘y’ = 0 within each category of that feature. ‘Job’, ‘campaign’, ‘age_group’ and ‘month’ were chosen for the first round of modeling due to their high variance of outcome ‘y’.

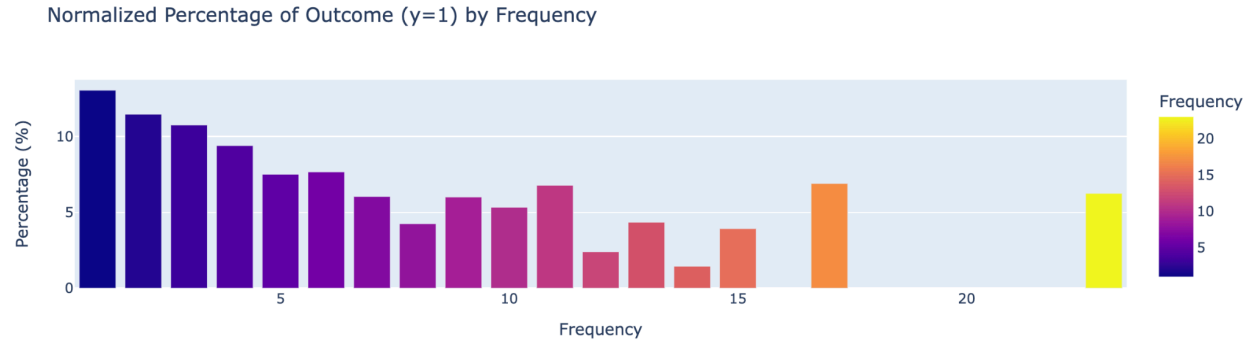
Within ‘job’, the lowest percentage of outcome ‘y’ = 1 (customer bought the term deposit) was 6.89% within blue-collar, whereas within student, ‘y’ = 1 was 31.43%. Retired customers also bought the term deposit at a higher rate of 25.23%.



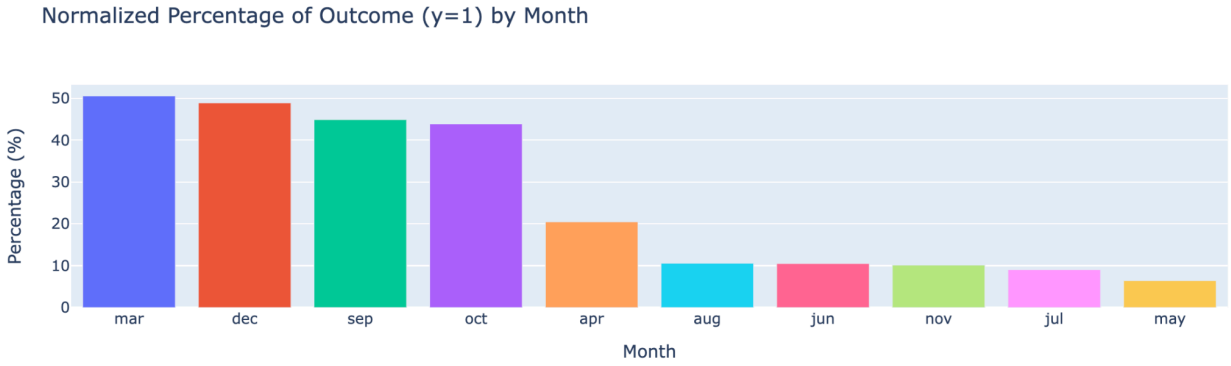
Within ‘age’, the lowest percentage of outcome ‘y’ = 1 was 8.17% and the highest was 49.58%. Older and younger people are the most likely to buy, which correlates to students and retired customers also being the most likely to buy.



Within ‘campaign’, or frequency, the lowest percentage of outcome ‘y’ = 1 was 1.45% and the highest was 13.04%.



Within ‘month’, the lowest percentage of outcome ‘y’ = 1 was 6.43% in May, whereas in March, ‘y’ = 1 was 50.55%.



Part 2: Modeling

Model Details

‘Job’, ‘campaign’, ‘age_group’ and ‘month’ were chosen as the features ‘X’ for the first round of KNN classification modeling, predicting the target ‘y’, whether a customer will purchase a term deposit.

‘Job’, ‘age_group’ and ‘month’ were transformed into dummy variables and reference categories were made for use in the model.

In order to improve the model performance, the next version of the model used all features ‘X’, except ‘default’, ‘housing’ and ‘loan’ due to their missing values. All other categorical features were also transformed into dummy variables with reference categories.

Model Scores

The resulting score using the first round of ‘X’ variables on the training data was 88.82%, only slightly higher than the baseline of 88.73%. The accuracy score on the test data was 88.33%. The recall score was 17.54%.

The second iteration of the model increased performance to 91.45% against the training data. The model was then refined, identifying the ideal n_neighbors (k) and weight (w) parameters of k = 10 and w = distance.

With the inclusion of all features ‘X’ and the tuning of the k and w parameters, performance of the model on test data improved slightly to 88.43%. Recall improved to 29.09%.

The confusion matrices for both models saw the majority of predictions fall into the true negative quadrant, meaning it was predicted the customer would not purchase, and they actually did not.

A false positive means it was predicted that a customer would purchase, but they did not. There are 190 of these in the first model and 290 in the second model.

A false positive in this case could cause inefficiencies as a result of marketing to people expected to purchase, but do not. That money could be spent on other customers who will purchase.

A false negative means it was predicted that a customer would not purchase, but they did. There are 771 of these in the first model and 663 in the second model.

A false negative in this case could represent missed opportunities to market to customers who could be convinced to purchase. We don't know how many more people predicted to not purchase would purchase if they were contacted.

Choosing the Best Model

The second model performed better than the first model, based on:

- Accuracy score & recall scores
- Number of true predictions
- Performance within key demographics

Accuracy & Recall Scores

The second model performed better in both accuracy and recall scores:

Accuracy

Model 1: 88.33%
Model 2: 88.43%

Recall

Model 1: 17.54%
Model 2: 29.09%

Number of True Predictions

The number of true negatives decreased from 7113 to 7013 between the first and second model, but the number of true positives increased from 164 to 272. Thus, the overall number of true predictions increased by 8 between model 1 and model 2.

Performance within Key Demographics

The second model was much better at predicting if younger customers will purchase, which is important given their high conversion rates. The first model predicted 37.5% of the 0-20 age group correctly vs. 70.83% for the second model. The second model also did slightly better at predicting if older customers will purchase, 65.22% for 60-70 age group, 62.5% for 70-80 age group and 48.15% for 80+ age group. The first model predicted the same groups at 52.17%, 57.14% and 48.15%, respectively.

The second model also did better at predicting whether retired customers and customers who are students will purchase, which is also valuable given their higher conversion rates. The first

model predicted retired customers correctly at 72.83% and students correctly at 60.39%. The second model predicted the same groups at 75.07% and 73.38%, respectively.