CIND119 Final Project:
Finding an Effective Telemarketing Strategy using the Bank Dataset

Joshua Joachimpillai JJoachimpillai@ryerson.ca
Ramello Peralta ramello.peralta@ryerson.ca

Summary

Our objective for this project was to provide an effective strategy for the client to successfully recruit more customers to subscribe to their long-term deposit accounts using data analytics and data science concepts. We were provided the bank dataset with variables describing the customer's socioeconomic status (eg. marital, job, education, loan, housing), as well as some variables relating to previous telemarketing campaigns (ie. contact, campaign, previous, duration). This was a supervised classification task in which the outcome is defined by the class variable "y", with a target outcome of "yes" (the client subscribed to a long-term deposit account) or "no" (they did not). We used Decision Trees and Naïve Bayes analyses on our data and compared mainly precision and AUROC metrics to see which model performs better for this case. Our analyses showed that both models performed well, but the decision tree was slightly better with AUROC and precision. Our top five most important attributes were duration, month, marital, balgroup (the balance a customer had) and job. We provide the client with insight and recommendations to their strategy based on our exploratory data analysis and model output.

Workload distribution

Member Name	List of Tasks performed
Joshua Joachimpillai	● **Variable Analysis** ● **Writeup** ● **Tableau**
Ramello Peralta	● **Variable Analysis** ● **Writeup** ● **Python coding**

Data Preparation

Attribute Type

Nominal attributes consisted of the following: “job”, “marital”, “default”, “housing”, “loan”, “contact”, “month”, “poutcome”, “y”. Quantitative attributes included “age”, “balance”, “day”, “duration”, “campaign”, “pdays”, and “previous”. Initially, the only ordinal attribute was education, but balance was discretized into an ordinal variable named “balgroup” in data processing.

Missing Values

This dataset contained “unknown” values which meant the client was not contacted and the information was not yet known. We treated these unknown values as NAs.

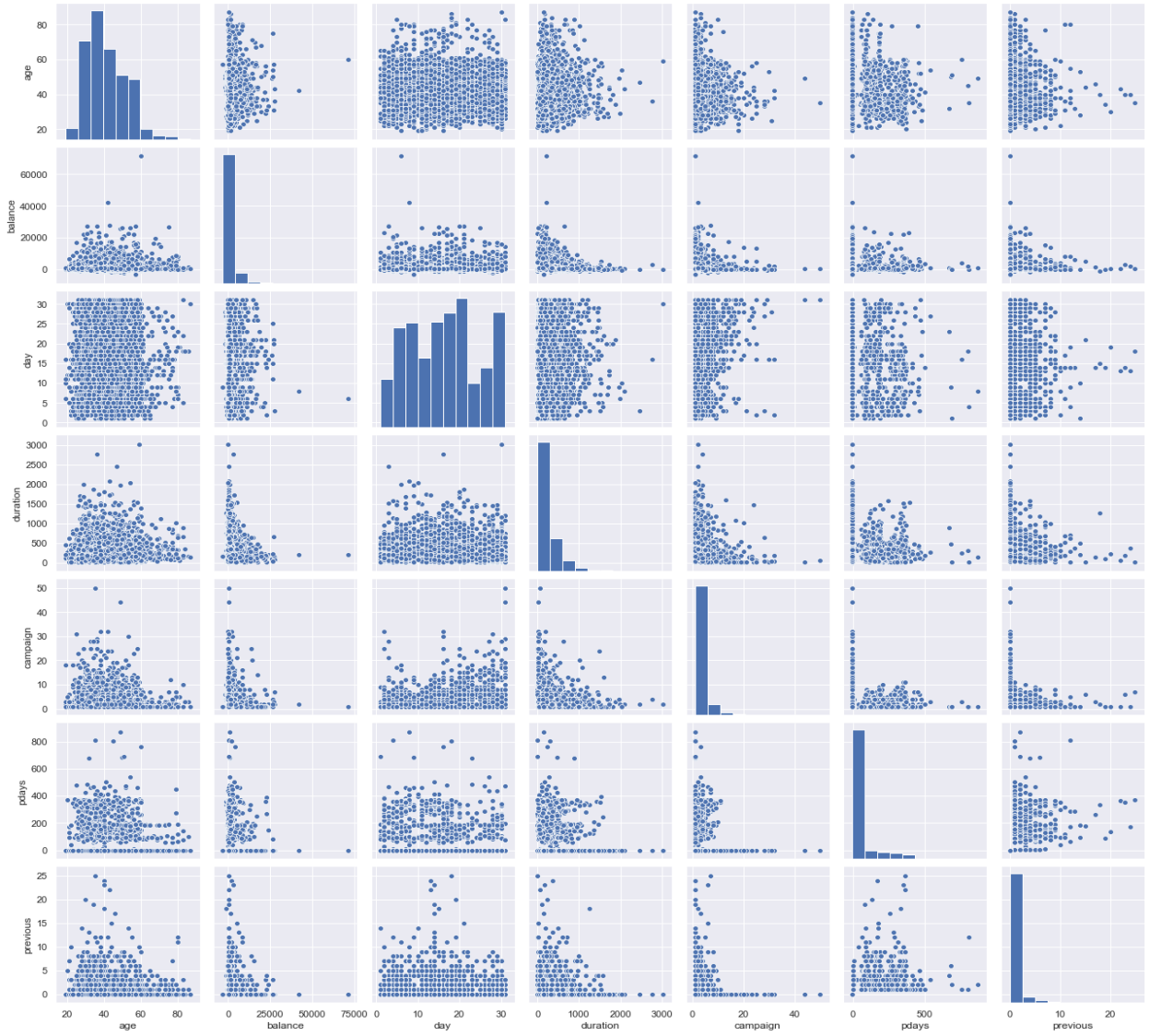
Descriptive Statistics (Min, Max, Mean)

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
age	4521	19	87	41.17	10.576
balance	4521	-3313	71188	1422.66	3009.638
day	4521	1	31	15.92	8.248
duration	4521	4	3025	263.96	259.857
campaign	4521	1	50	2.79	3.110
pdays	4521	-1	871	39.77	100.121
previous	4521	0	25	.54	1.694
Valid N (listwise)	4521				

Outliers

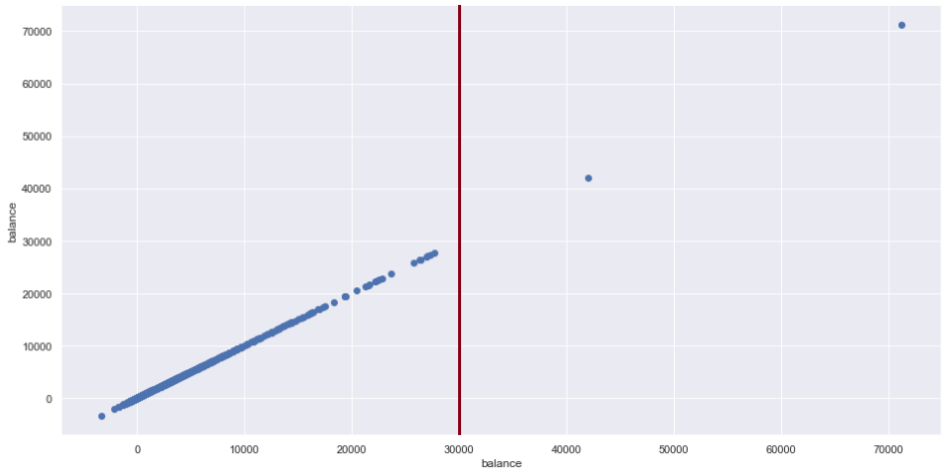
Distributions of each quantitative variable were observed in following pairplot:



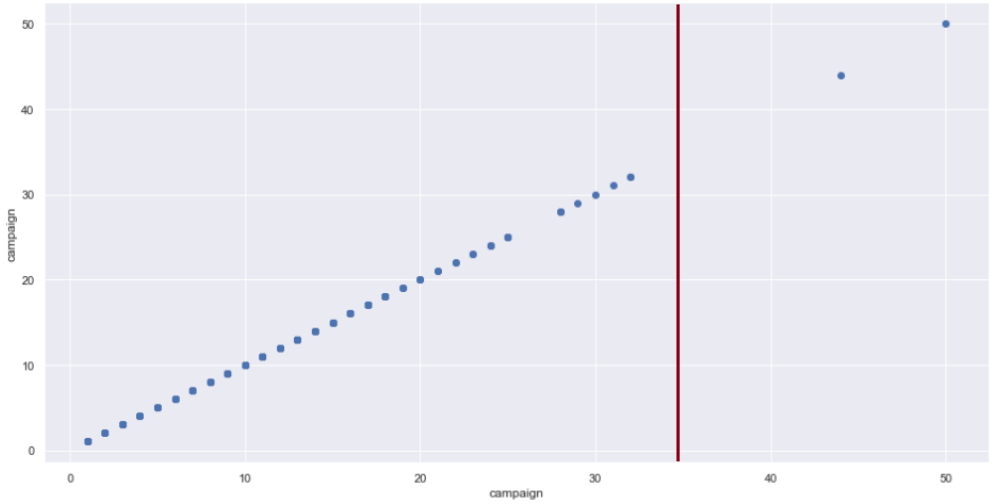
It is apparent that most of our variables are not normally distributed, and some variables have prominent outliers (eg. balance).

Outliers were removed based on visually identifying significantly differing data points. We chose to minimally remove outliers on the top-end to preserve the integrity of each variable and prioritize retaining as many records as possible.

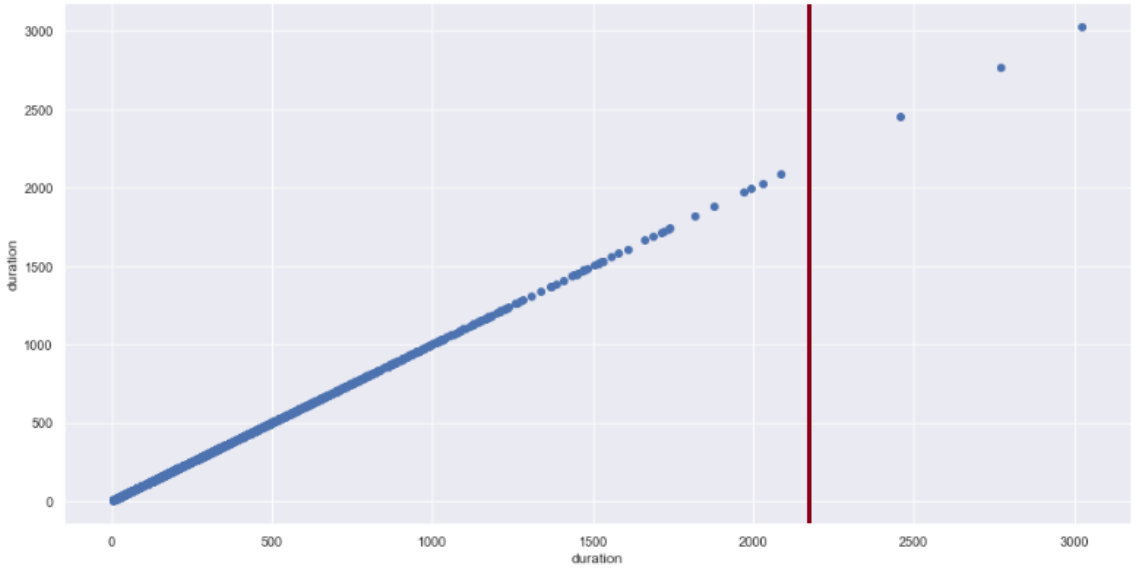
Balance:



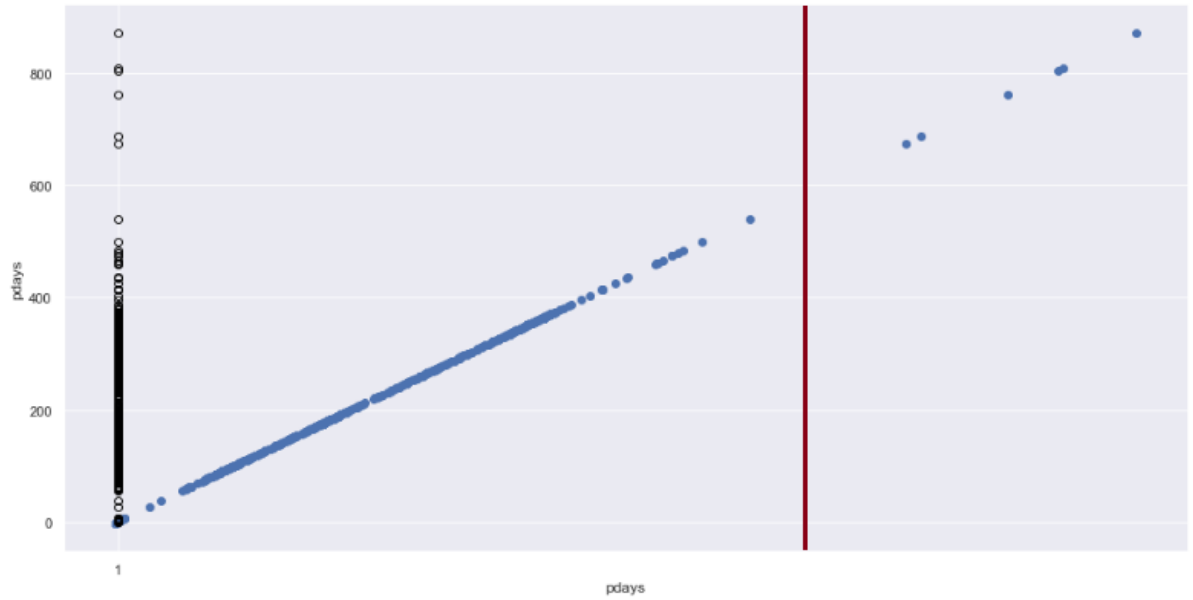
Campaign:



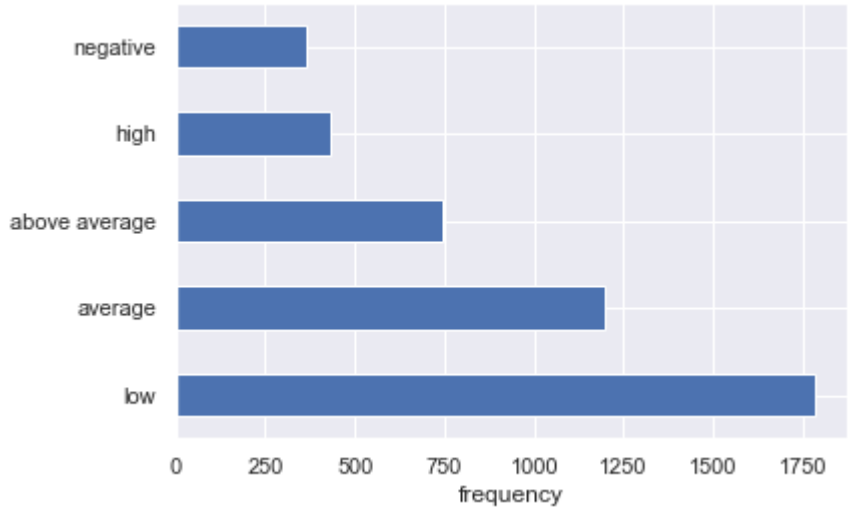
Duration:



pdays:



Discretizing balance



Balance was discretized into 5 categories based on quartiles and the mean. This was done to eliminate the negative values (necessary for Naive Bayes) as well as for simplicity in choosing groups for the telemarketers to focus. The new variable was named “balgroup”.

Old value (balance)	New variable (discretized)
Less than $0	**Negative**
$0-$399.99	**Low**
$400-$1399.99	**Average**
$1400-$3999.99	**Above Average**
$4000+	**High**

Criteria for excluding attributes

In this dataset, “unknown” is used to denote that a customer has not yet been contacted, therefore some variables will not have any value. In our case, these values can be considered NA. Some attributes (like poutcome) had higher correlations with our dependent variable, but were still removed because +80% of the values were unknown. We cannot impute NA values for such a large portion of the dataset without changing the integrity of the data. We also did not want to remove each row containing an unknown value as it would drastically reduce the size of our dataset. Hence, the solution was to drop the columns.

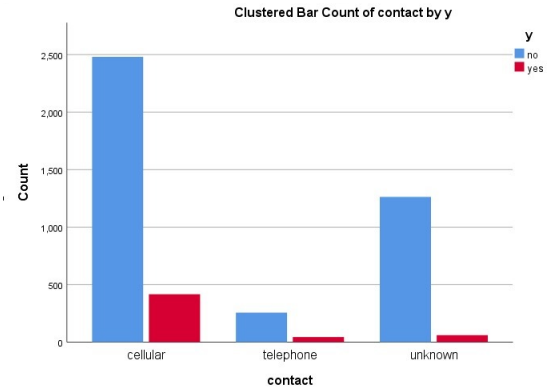
Removing Pdays & Poutcome

```
In [4]: bank.poutcome.value_counts()
Out[4]:
unknown    3705
failure     490
other       197
success     129
Name: poutcome, dtype: int64
```

```
In [5]: bank.pdays.value_counts()
Out[5]:
-1      3705
 182      23
 183      20
 363      12
 92       12
        ...
 222       1
 210       1
 206       1
 162       1
 28        1
Name: pdays, Length: 292, dtype: int64
```

pdays was removed as 82% of the data were unknown (unknown was numerically encoded as -1 in this variable). “poutcome” was removed for the same reason.

Removing Contact




```
In [7]: bank.contact.value_counts()
Out[7]:
cellular     2896
unknown      1324
telephone     301
Name: contact, dtype: int64
```

Contact (whether the call was made on a cell or landline) was removed as 30% of the values were NA as well. Also, there was no discernible pattern or correlation in contact compared to y.

Removing Previous

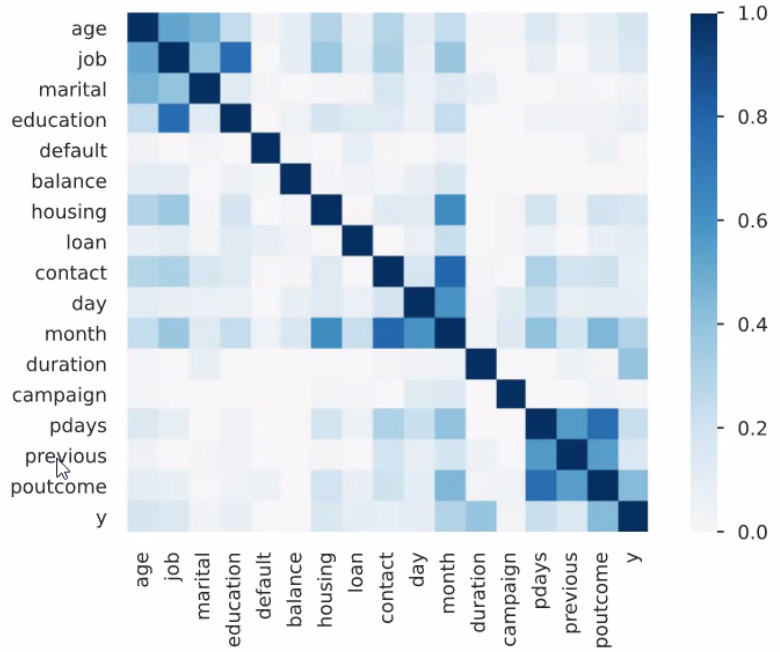
```
In [24]: sum(df.previous==0)
Out[24]: 3698
```

Similar to pdays, previous has NA values encoded as 0 (as in, these customers were not previously contacted in the last campaign) which comprised a majority of the column, hence it was removed.

For other variables like job or education, we simply removed the records with unknowns as it resulted in only a ~5% reduction of our dataset. Only 225 entries were removed from this procedure.

Bivariate Analysis

Looking for patterns in correlation between y and the independent variables



Duration and poutcome have the highest correlation to the class variable.

Correlations

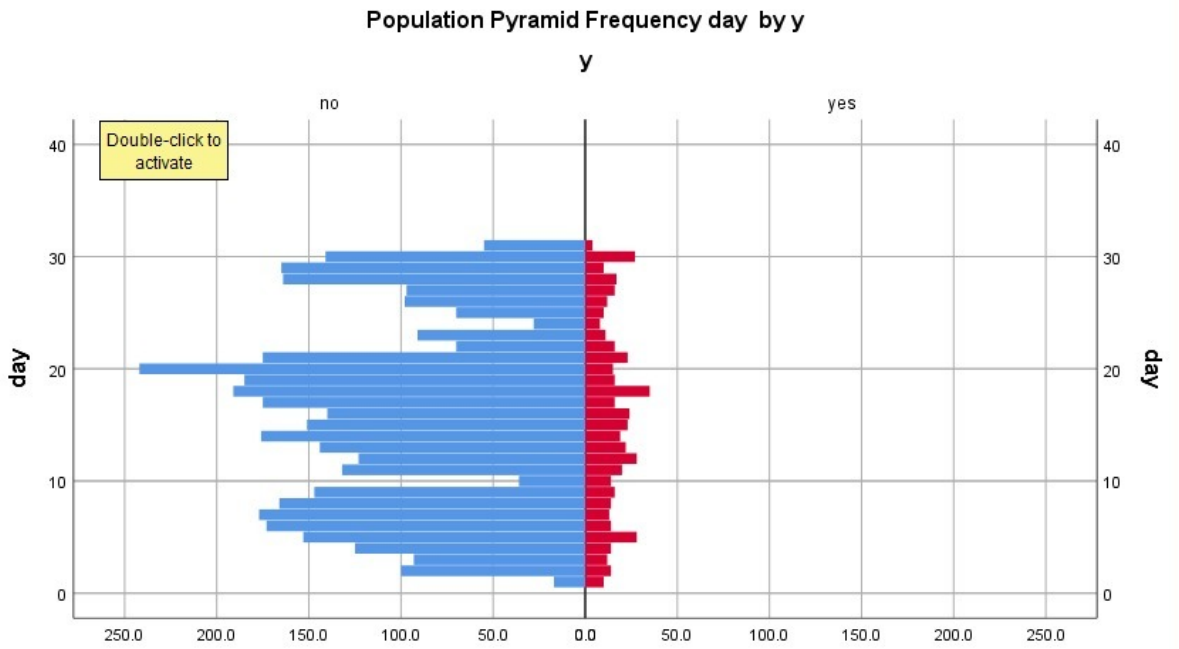
		age	balance	day	month	duration	campaign	pdays	previous	Y numeric
age	Pearson Correlation	1	.084**	-.018	.074**	-.002	-.005	-.009	-.004	.045**
	Sig. (2-tailed)		.000	.230	.000	.874	.729	.550	.813	.002
	N	4521	4521	4521	4521	4521	4521	4521	4521	4521
balance	Pearson Correlation	.084**	1	-.009	.100**	-.016	-.010	.009	.026	.018
	Sig. (2-tailed)	.000		.560	.000	.284	.502	.526	.078	.229
	N	4521	4521	4521	4521	4521	4521	4521	4521	4521
day	Pearson Correlation	-.018	-.009	1	.080**	-.025	.161**	-.094**	-.059**	-.011
	Sig. (2-tailed)	.230	.560		.000	.098	.000	.000	.000	.450
	N	4521	4521	4521	4521	4521	4521	4521	4521	4521
month	Pearson Correlation	.074**	.100**	.080**	1	.000	.059**	-.112**	-.037*	.023
	Sig. (2-tailed)	.000	.000	.000		.985	.000	.000	.012	.117
	N	4521	4521	4521	4521	4521	4521	4521	4521	4521
duration	Pearson Correlation	-.002	-.016	-.025	.000	1	-.068**	.010	.018	.401**
	Sig. (2-tailed)	.874	.284	.098	.985		.000	.485	.224	.000
	N	4521	4521	4521	4521	4521	4521	4521	4521	4521
campaign	Pearson Correlation	-.005	-.010	.161**	.059**	-.068**	1	-.093**	-.068**	-.061**
	Sig. (2-tailed)	.729	.502	.000	.000	.000		.000	.000	.000
	N	4521	4521	4521	4521	4521	4521	4521	4521	4521
pdays	Pearson Correlation	-.009	.009	-.094**	-.112**	.010	-.093**	1	.578**	.104**
	Sig. (2-tailed)	.550	.526	.000	.000	.485	.000		.000	.000
	N	4521	4521	4521	4521	4521	4521	4521	4521	4521
previous	Pearson Correlation	-.004	.026	-.059**	-.037*	.018	-.068**	.578**	1	.117**
	Sig. (2-tailed)	.813	.078	.000	.012	.224	.000	.000		.000
	N	4521	4521	4521	4521	4521	4521	4521	4521	4521
Y numeric	Pearson Correlation	.045**	.018	-.011	.023	.401**	-.061**	.104**	.117**	1
	Sig. (2-tailed)	.002	.229	.450	.117	.000	.000	.000	.000	
	N	4521	4521	4521	4521	4521	4521	4521	4521	4521

**. Correlation is significant at the 0.01 level (2-tailed).

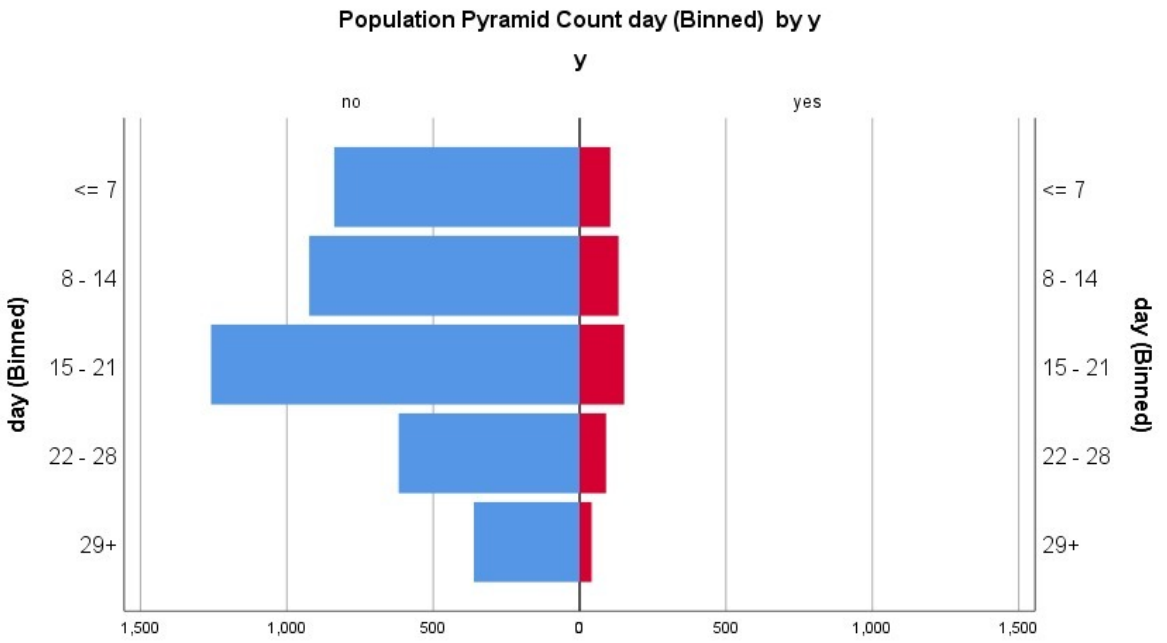
*. Correlation is significant at the 0.05 level (2-tailed).

Age, campaign, duration, pdays and previous have a significant correlation with “y” (whether or not the client subscribed). The correlation between campaign and y is negative--the more potential subscribers were contacted, the less likely they were to sign up.

Days & Subscription

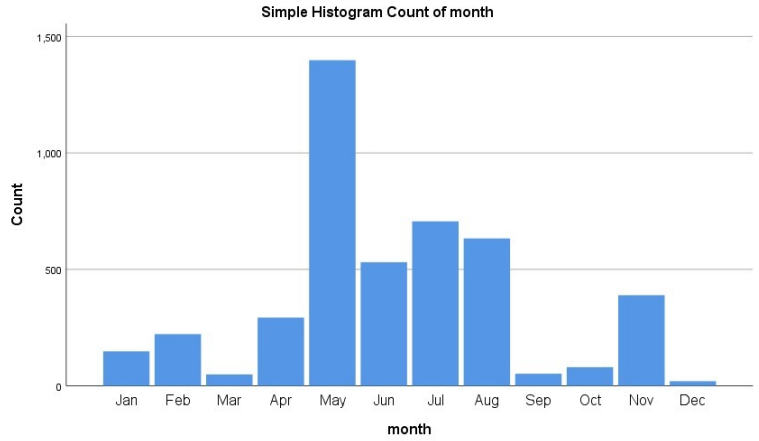


“Days” was dropped as there was no discernible relationship between days and y. Analysis of day’s relationship with y showed no observable pattern.

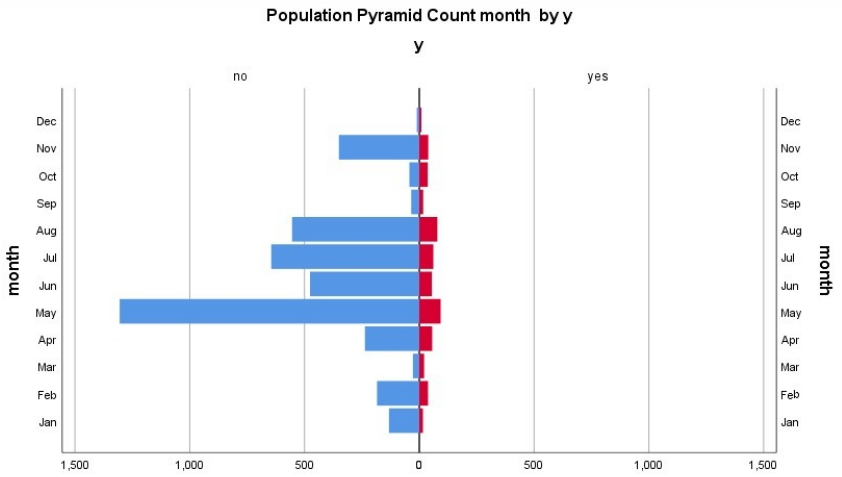


Further discretizing days into “weeks”, there is still no significant pattern.

Month & Subscription

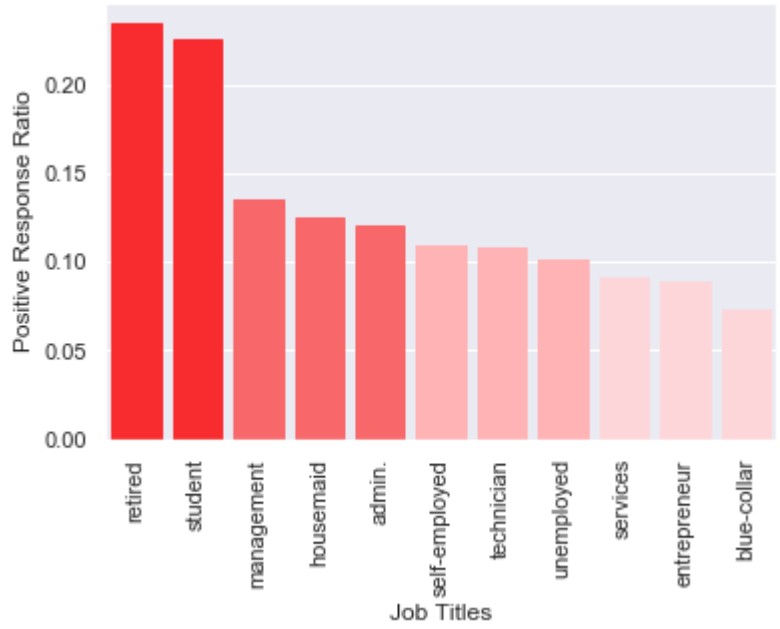


Contacting clients was not uniform over the months; there was a higher push in May.

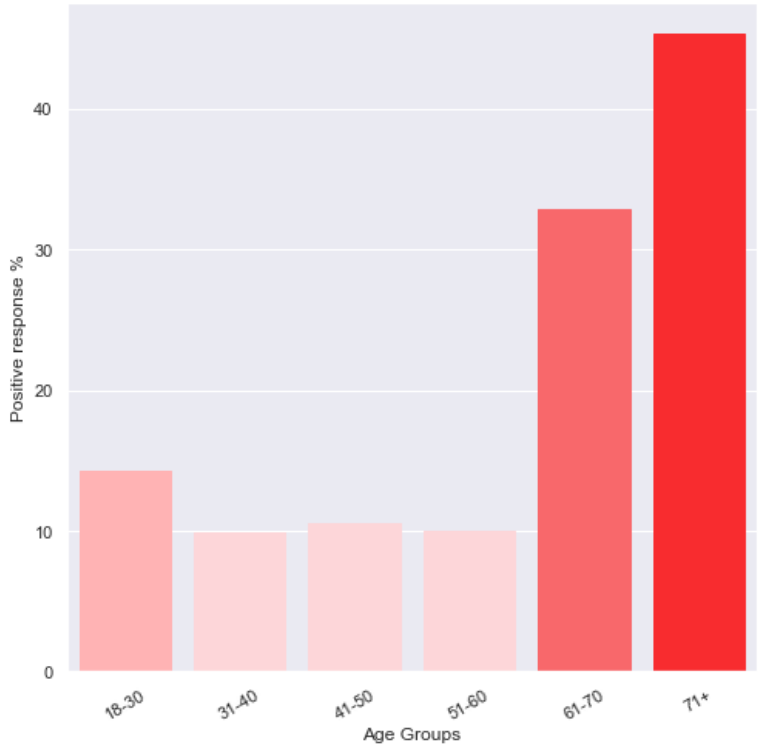


From this chart we can see that in the months of March, Oct, Nov and Dec there is around an equal proportion of yes/no, which may be a significant pattern indicating that these months may be the most fruitful for signups. However, there was a lower number of total attempts for these months, suggesting more attempts could be made in order to confirm.

Employment & Subscription

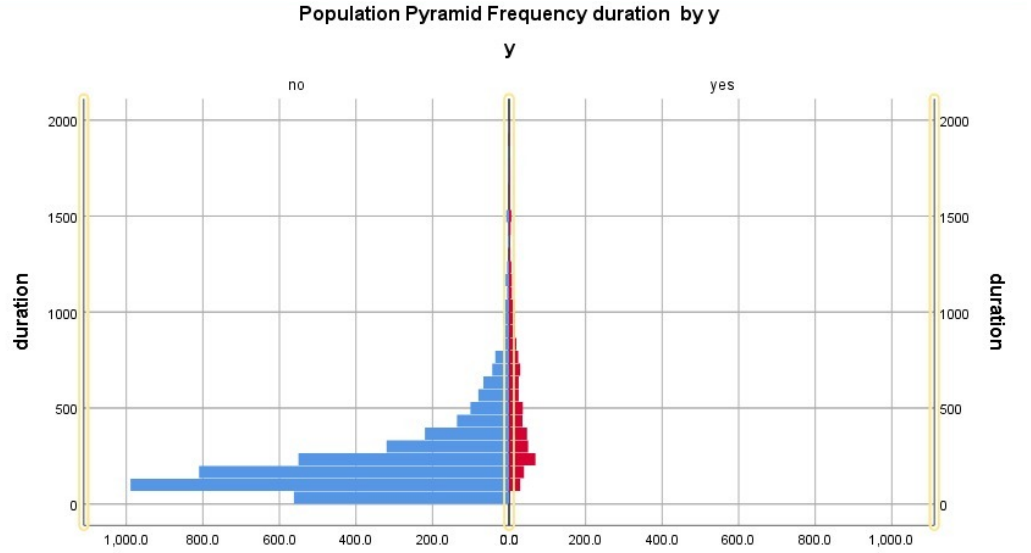


From our analysis on employment, the two most likely subscribers to a bank account are retirees and students. From this information, we can also associate other factors related to job title, such as age.



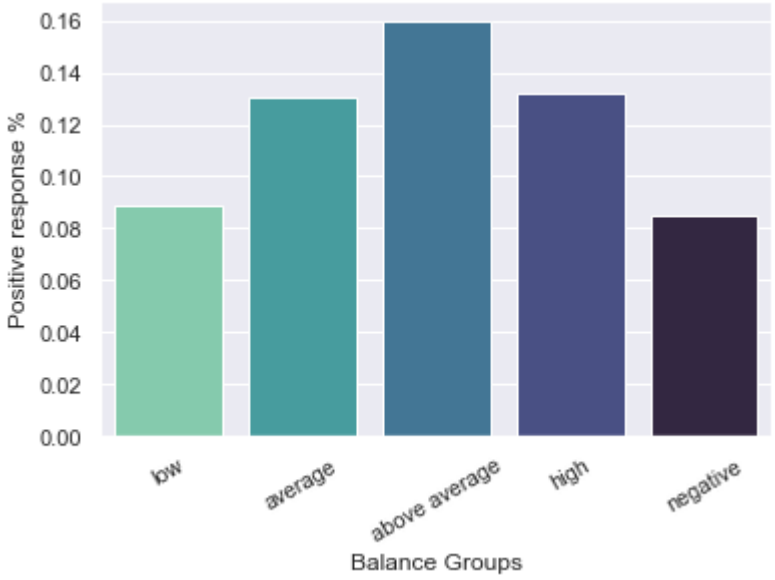
Closely matching the pattern for employment, student-age recipients from 18-30 and retirees from 60+ are also prime targets for this telemarketing strategy.

Duration & Subscription



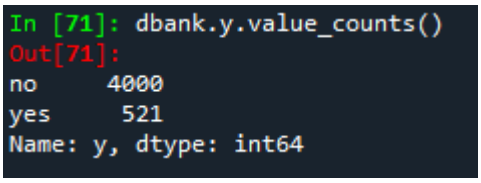
Majority of the calls ended within 5 minutes, however at the 15 minute (1000s) mark, we have almost an equal proportion of recipients subscribing to the bank account. This might mean that the clients were more engaged and interested in the idea and asked questions which extended the time to this point.

Balance & Subscription



After discretizing our balance groups, it is easier to see that most in the above average range (1400-4000 a year) were the most likely to subscribe to a deposit.

Imbalanced class distribution



Our dataset was heavily imbalanced, meaning the “no” class occurred in much greater proportion compared to the minority positive class (y = “yes”). This skewness can affect model output and predictive power for the “yes” class. Accuracy was avoided as the main metric because it may be misleading as our model was trained on an imbalanced dataset. Higher model accuracy may mean that our model would drastically overpredict the majority class (y = “no”). Our main approach for this case is to include AUROC as an additional evaluation metric. Other options we discussed to deal with this imbalance data was downsampling, but we did not want to dramatically change the size of our dataset by downsampling.

ROC curve is plotted on TPR vs FPR. Area Under ROC(AUROC) determines the degree to which a model is able to properly discriminate between the positive and negative class. AUROC is more sensitive to imbalanced datasets and makes it a better suited metric for our case.

Changes we made on our dataset:

- Removal of fringe outliers
- Removal of records with unknowns in job and education
- Removal of variables: pdays, poutcome, contact, previous, days

Predictive Modeling

In this project we used Decision Trees and Naive Bayes as our classifiers. We did a simple train-test split for our dataset.

Decision Trees

Decision trees model a categorical response which makes it a good classification tool for our case. A decision tree starts with a feature on the root node, and “grows” based on GINI in our algorithm. The GINI index is a measure of “impurity” of each node. Each next feature added is based on the lowest GINI score. The tree optimizes for higher information gain/lower GINI scores for better separation of classes.

Naive Bayes

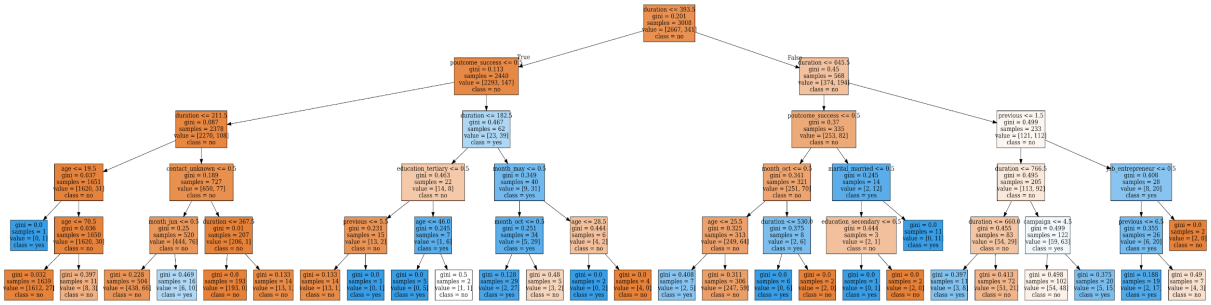
Naive Bayes is another classification algorithm that relies on conditional probabilities. This score is calculated based on relative frequencies of occurrence. For the case of quantitative data, it is discretized via the algorithm.

Both classification tools are useful, however the main drawback in Naive Bayes is that our dataset violates the independence assumption. Our variables are not independent. In fact, many variables are dependent on pdays ; whether or not the client has been contacted before (pdays== -1) affects the values of other columns such as duration and campaign.

Performance Measures

AUROC and precision were our main metrics for the models.
AUROC, as described above, is more sensitive to imbalanced data thus it serves as a more reliable metric than accuracy.

We also valued precision because from our understanding (with little to no domain experience in this area) that false positives would have a higher cost to the business. More false positives would result in more financial resources spent for workers to contact predicted positive individuals who will not end up subscribing to a long term bank account (ie false positives). It is in the business’ best interest to avoid spending too much on clients who will not end up signing up for an account.



Metrics before data processing

NAIVE BAYES

	precision	recall	f1-score	support
no	0.94	0.80	0.86	1207
yes	0.25	0.56	0.35	146
accuracy			0.77	1353
macro avg	0.60	0.68	0.61	1353
weighted avg	0.86	0.77	0.81	1353

ROC AUC Score = 0.768

DECISION TREE

	precision	recall	f1-score	support
no	0.92	0.95	0.94	1207
yes	0.46	0.35	0.40	146
accuracy			0.88	1353
macro avg	0.69	0.65	0.67	1353
weighted avg	0.87	0.88	0.88	1353

ROC AUC Score = 0.749

AFTER DATA PROCESSING:

NAIVE BAYES

	precision	recall	f1-score	support
no	0.94	0.82	0.88	1137
yes	0.32	0.63	0.43	153
accuracy			0.80	1290
macro avg	0.63	0.72	0.65	1290
weighted avg	0.87	0.80	0.82	1290

ROC AUC Score = 0.806

DECISION TREE

	precision	recall	f1-score	support
no	0.90	0.98	0.94	1137
yes	0.62	0.22	0.33	153
accuracy			0.89	1290
macro avg	0.76	0.60	0.63	1290
weighted avg	0.87	0.89	0.87	1290

ROC AUC Score = 0.821

The decision tree with our finalized dataset has a roc score of 0.821, an increase of almost 0.07 from the original data. Precision also increased by almost 0.17 in the decision tree after data processing, and only increased 0.03 in Naive Bayes.

Conclusion and recommendations

Our modelling suggested that decision tree based classification was the best model. The decision tree had the best precision metrics (tree=0.62, Naïve Bayes=0.32), and AUROC score (tree=0.821, Naïve Bayes=0.806). Naïve Bayes also performed well as a classifier based on performance metrics. However, we are cautious to accept it as a good estimator because the data are not independent. With the decision tree model focused on precision, we can be assured that the company will spend less money and resources on low-likelihood subscribers.

From the decision tree:

The most important features from the tree were duration, month, marital and job. These are the main features that the bank should focus on to get the best results.

Our recommendations for the bank based on our exploratory analysis and modelling output--

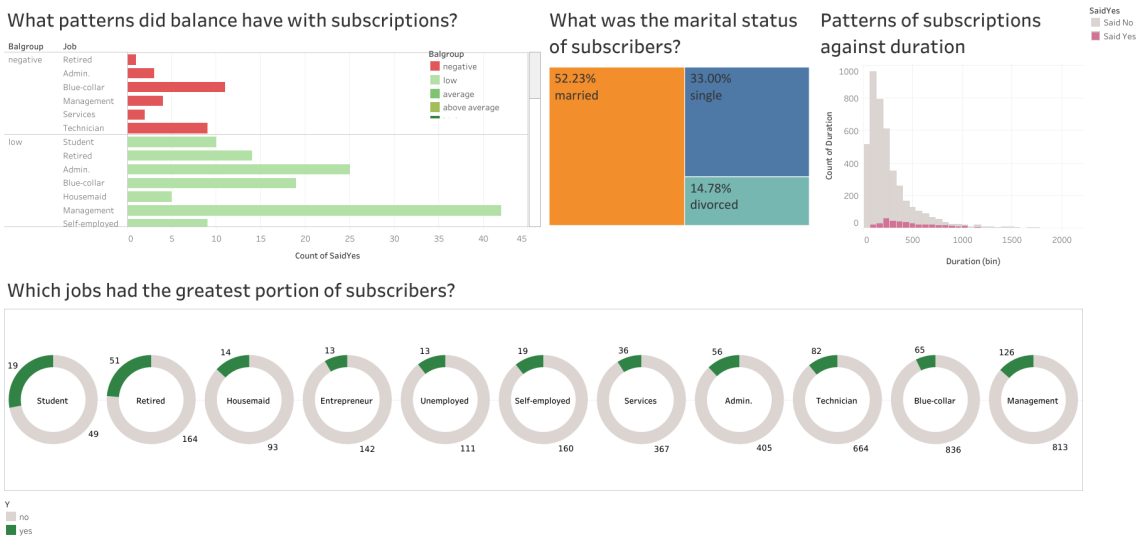
The target profile for a customer should be:

- Clients who show interest in the phone call, thereby having a call duration of ~15 minutes should be prioritized for additional contact and/or services
- Best months for contact are in Mar, Apr, Sep, Oct, and Dec
- Either 18-30 OR 60+ age
- Students or retired
- Single or Married

Individuals fitting these criteria are high priority targets for the bank. The model we have implemented will help the business narrow their focus on these important individuals, thereby lowering costs by avoiding low-likelihood subscribers.

Tableau

https://prod-useast-b.online.tableau.com/t/jjssite/views/CIND119RamelloJoshua/BankMarketingDashboard?:showAppBanner=false&:display_count=n&:showVizHome=n&:origin=viz_share_link



End of CIND119 Final Project.