Bank Marketing Campaign

Uday Singh

16-March-2023

Agenda

- Problem Statement
- Approach
- EDA
- EDA Summary
- Recommendations

Background

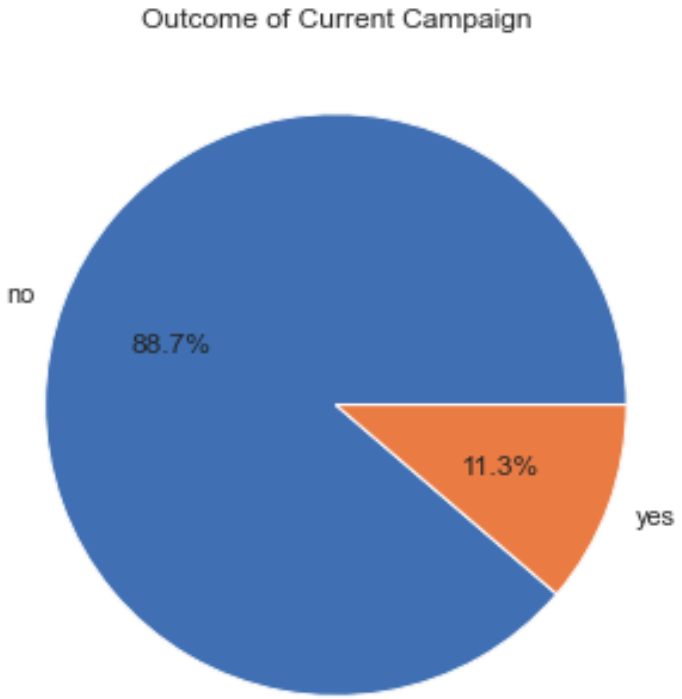
ABC Bank is planning to introduce a new term deposit product for its customers, and they need to develop a model that can predict whether or not a customer will buy their product. This will help the bank to focus its marketing efforts on those customers who are more likely to buy, resulting in a more efficient and effective sales process.

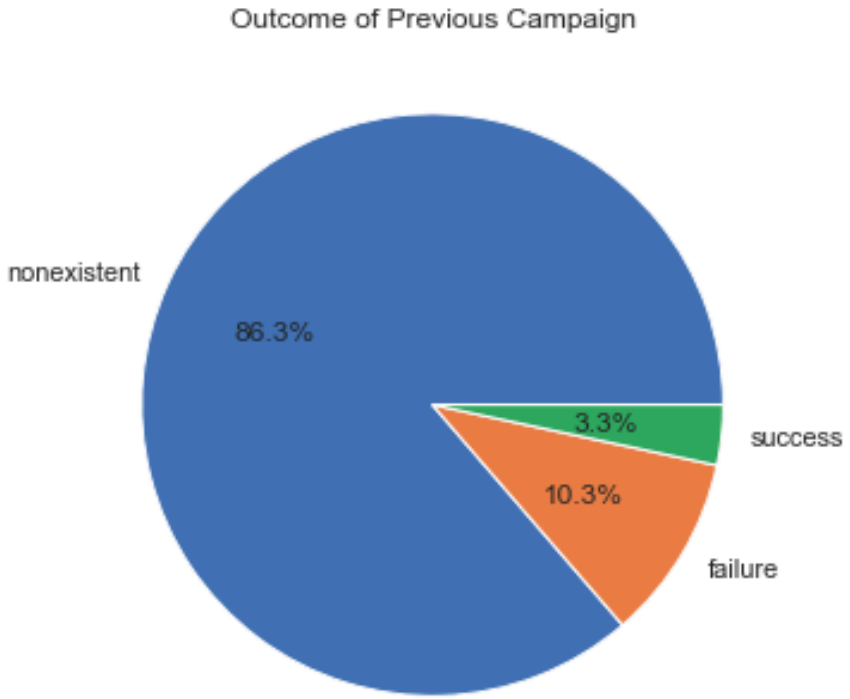
Approach:

- Prepare and clean data
- Explore data to understand trends and relationships
- Create visualisations
- Examine distribution of data
- Provide initial recommendations on the basis of EDA discoveries
- Suggestions for models

Feature	Description
age	The age of the customer (numeric)
job	The type of job of the customer (categorical: 'admin.', 'blue-collar',...)
marital	The marital status of the customer (categorical: 'divorced', 'married',...)
education	The level of education of the customer (categorical: 'basic.4y', 'basic.6y',...)
default	Whether the customer has credit in default (categorical: 'no', 'yes', 'unknown')
housing	Whether the customer has a housing loan (categorical: 'no', 'yes', 'unknown')
loan	Whether the customer has a personal loan (categorical: 'no', 'yes', 'unknown')
contact	The communication type used to contact the customer (categorical: 'cellular', 'telephone')
month	The month of the year when the customer was last contacted (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
day_of_week	The day of the week when the customer was last contacted (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
duration	The duration of the last contact in seconds (numeric)
campaign	The number of contacts performed during this campaign and for this client (numeric)
pdays	The number of days that passed by after the customer was last contacted from a previous campaign (numeric; 999 means the customer was not previously contacted)
previous	The number of contacts performed before this campaign and for this client (numeric)
poutcome	The outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')
emp.var.rate	Employment variation rate - quarterly indicator (numeric)
cons.price.idx	Consumer price index - monthly indicator (numeric)
cons.conf.idx	Consumer confidence index - monthly indicator (numeric)
euribor3m	Euribor 3 month rate - daily indicator (numeric)
nr.employed	Number of employees - quarterly indicator (numeric)
target	Whether the customer subscribed to a term deposit (binary: 'yes', 'no')

Response Variable

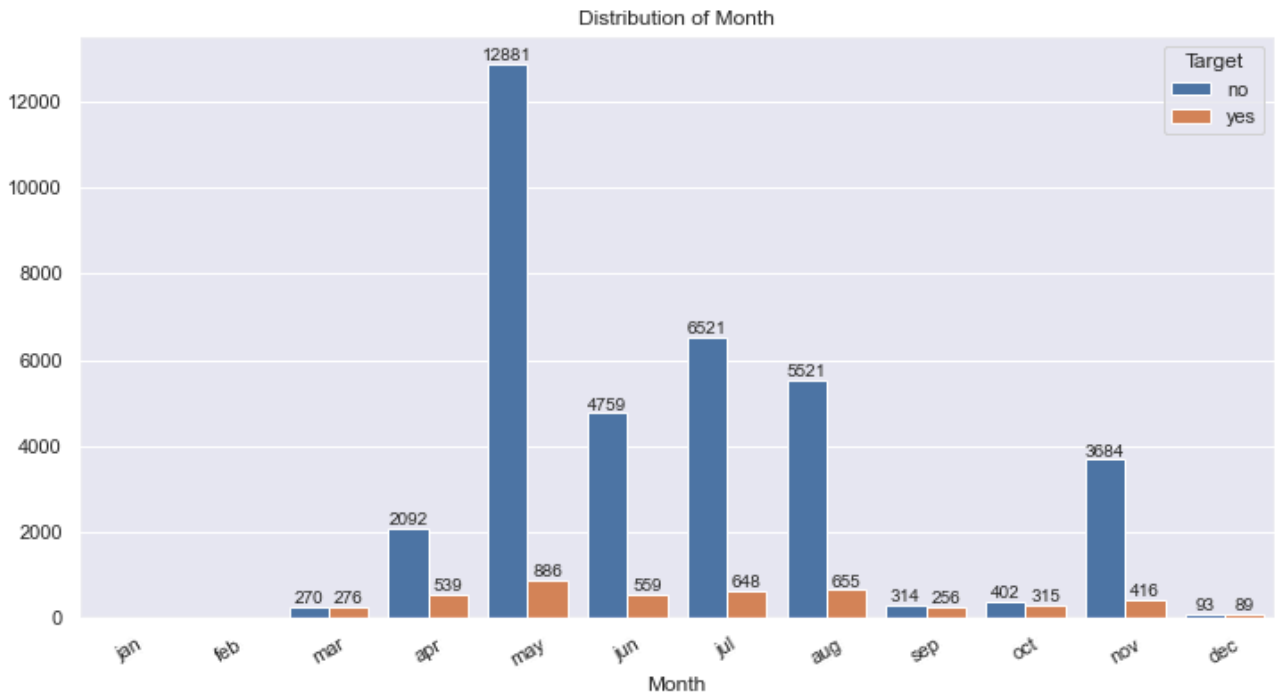


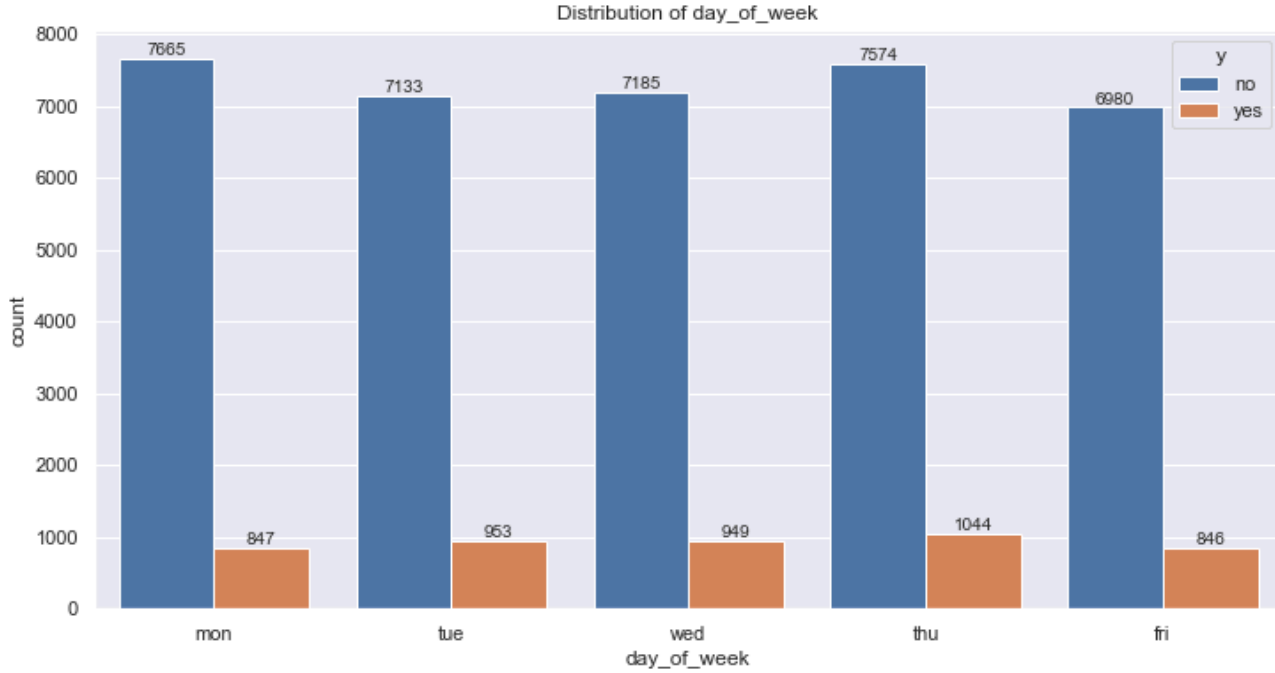


- For previous campaign, most (86.3%) of responses are unknown.
- Negative responses (no) were about 3 times as frequent as positive (yes).
- For the current campaign, all responses are known.
- 88.6% are negative and 11.3% are positive responses.
- Almost 8 times as many negative responses.
- Both campaigns were either unsuccessful in reaching the customer, or in convincing them to invest in a term deposit.

Aside from the immediately obvious conclusions we can draw from the charts, it also Reveals that the data is heavily imbalanced. This means that we might have to employ Special techniques when training our models in order to ensure accuracy.

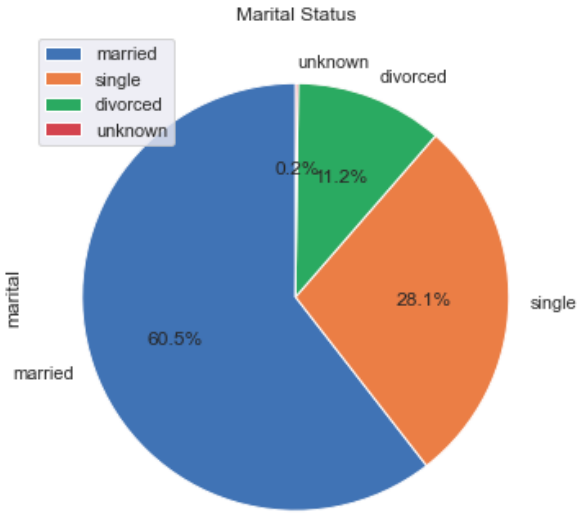
Time Variables



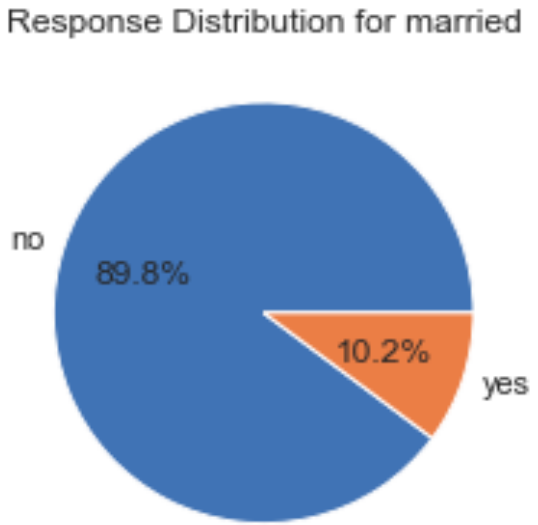


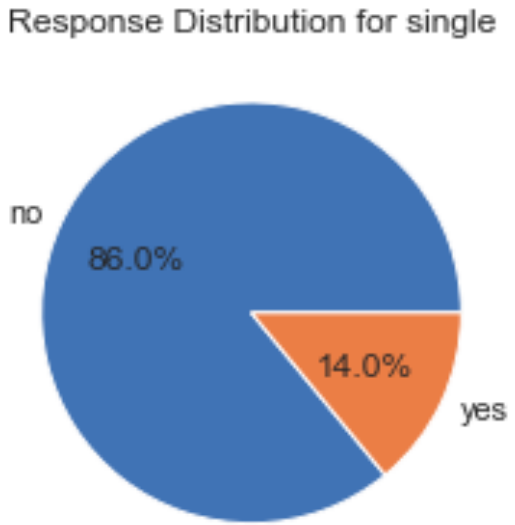
- No contact was made in the months of January and February.
- Summer months (May - August) have highest number of responses.
- Contact was made only on weekdays. Responses are approximately uniformly distributed across the days.
- This information can be used to optimise future marketing campaigns to focus more on the months with higher response rates.

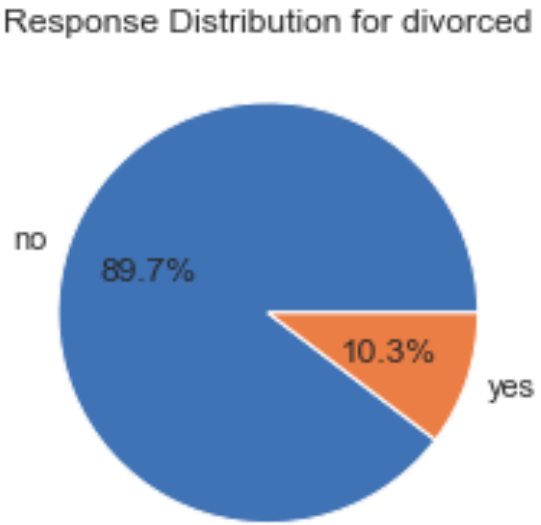
Demographic Data - Marital

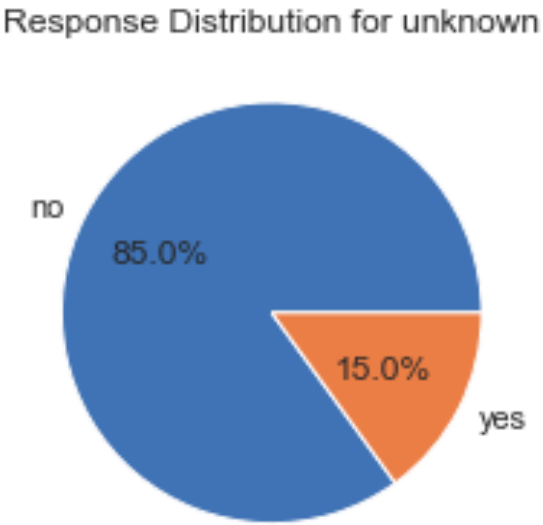


- We can see percentages for each category: Married, Single, Divorced, Unknown.
- Most (60%) clients are married, 28% single and 11% divorced (or widowed).
- There is no significant difference in the response to campaign when comparing the categories.
- We can conclude that the marital status of the client does not have a very strong influence on the outcome of the campaign.

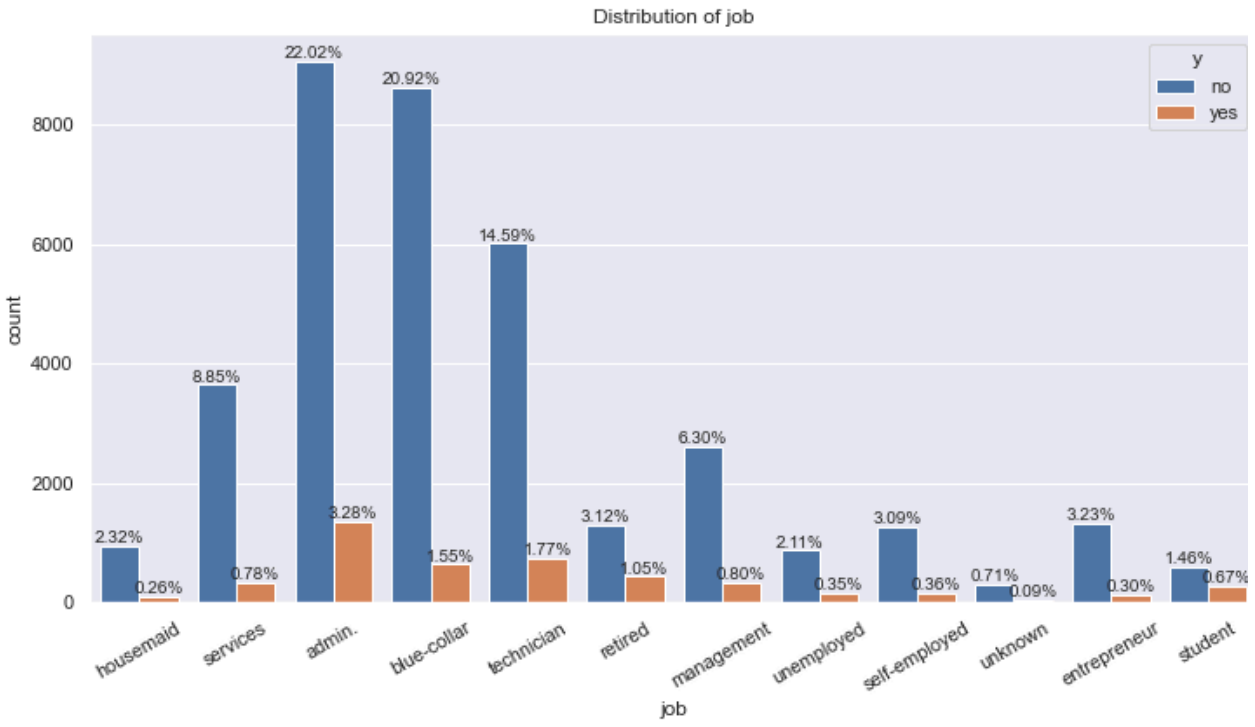


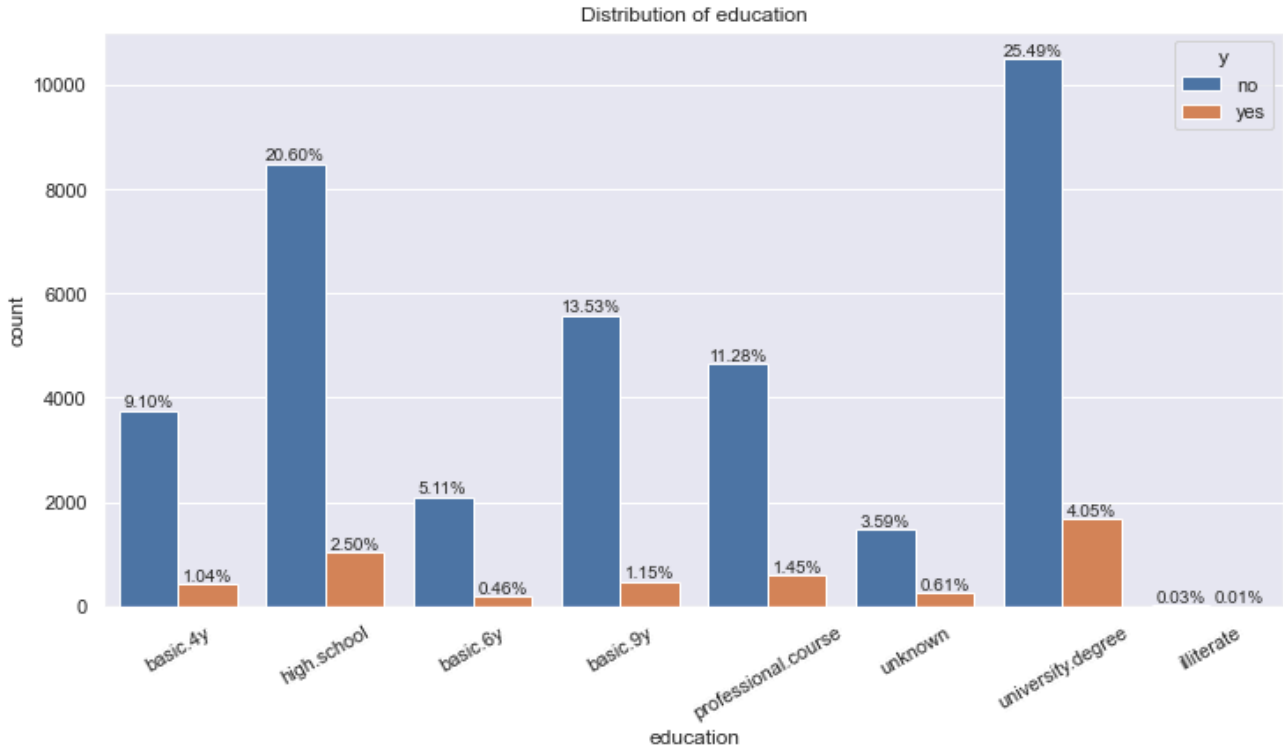






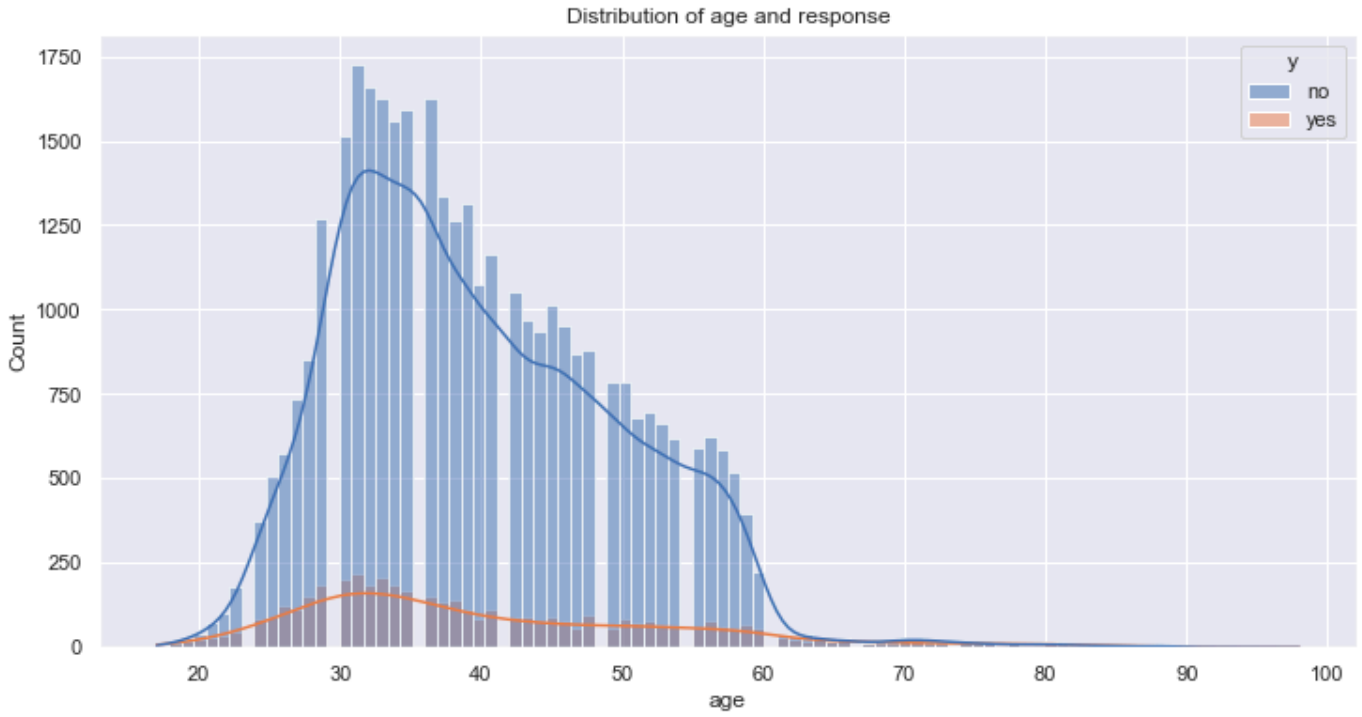
Demographic Data - Job and Eduction

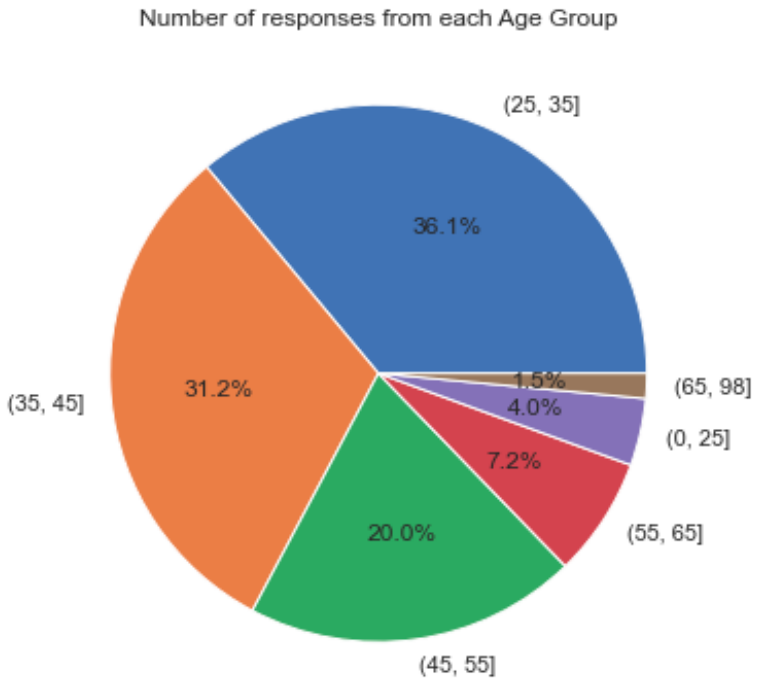




- Highest level of responses were from ‘admin’, ‘blue-collar’ and ‘technician’ professions. For each profession, the proportion of negative responses is far greater than positive.
- In Education, highest number of responses were from high-school and university clients. The response proportions seem to be in line with the trend seen in the data; the percentage of ‘no’ is much higher.
- The charts show us which professions and education levels are more likely to respond to the campaign.

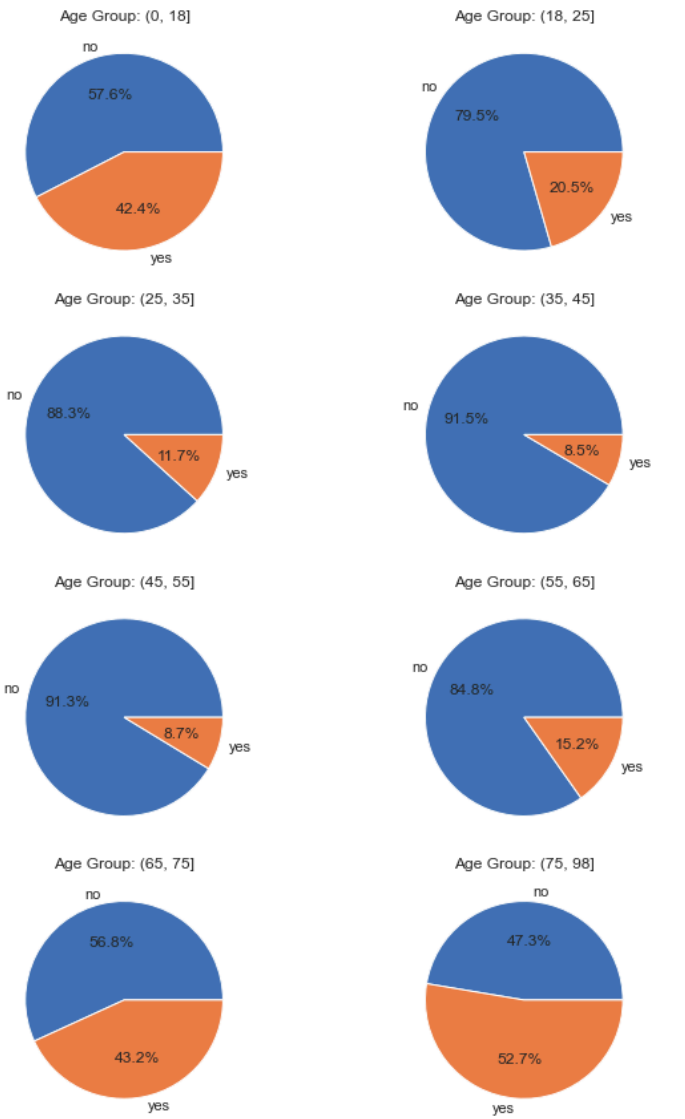
Demographic Data - Age





- The distribution of Age can be seen in the histogram. The probability function is approximately tri-modal with modes around 30, 45 and 60.
- We can break the age into subgroups and examine the counts in each. The majority of responses seem to be from the groups 25-35 and 35-45. The ages 0-25 have been clubbed into one group, and the same for 65 and above.

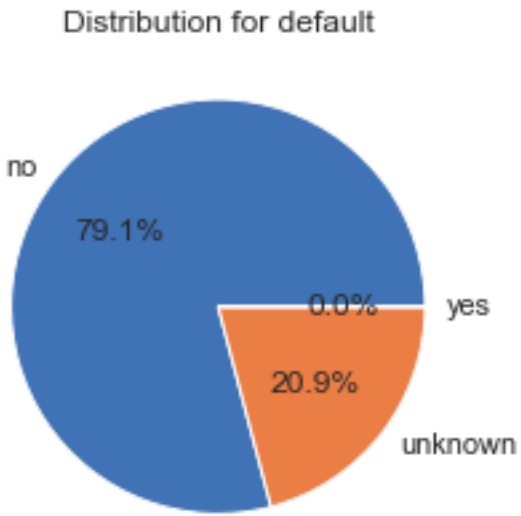
Demographic Data - Age

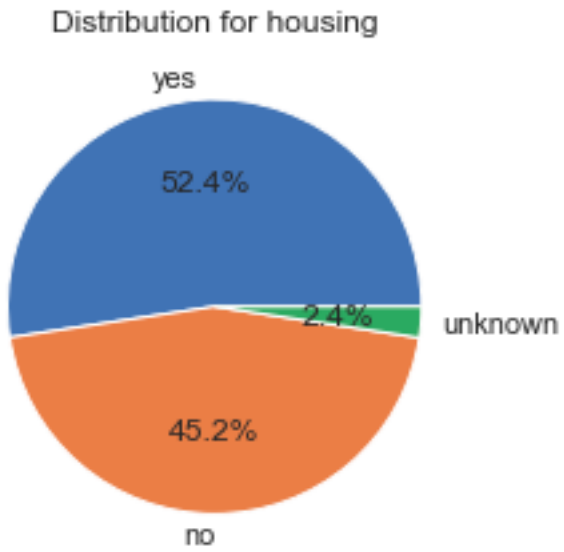


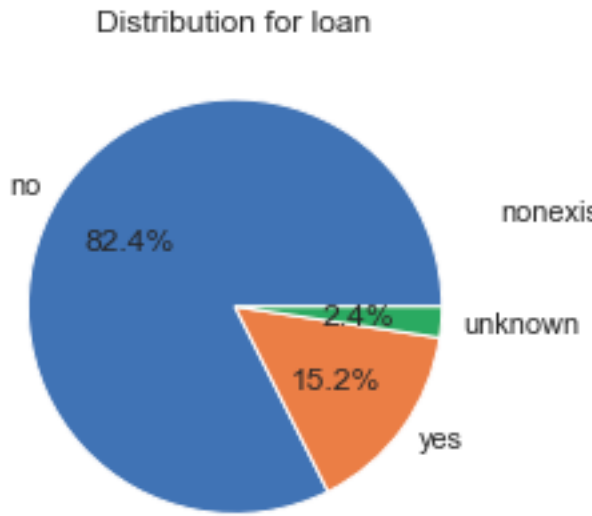
Observations for each pie chart:

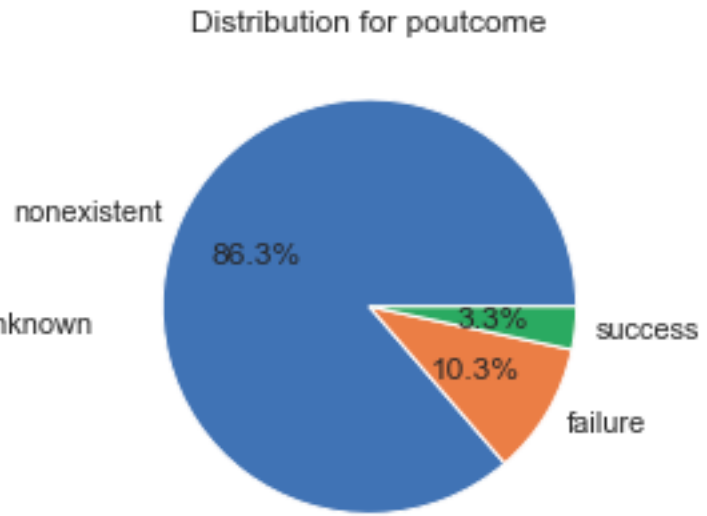
- Age Group: (0, 18]: The proportion of yes to no is 42% to 58%. The minimum age in the dataset is 17.
- Age Group: (18, 25]: 80% of the responses were negative. This might make sense people in this age group may not have the resources to invest in term deposits.
- Age Group: (25, 35]: 89% of the responses were negative.
- Age Group: (35, 45]: A slightly higher proportion of entries in this age group did not subscribe (91%) to the term deposit.
- Age Group: (45, 55]: Percentage in this age group did not subscribe to the term deposit is about the same as the last group (91%).
- Age Group: (55, 65]: Slightly better response; 15% yes and 85% no.
- Age Group: (65, 75]: Response is split; 57% no and 43% yes. This indicates that clients in the older age group are more likely to invest in term deposit.
- Age Group: (75, 98]: Only age group with higher percentage of positive responses. 47% no and 53% yes. The maximum age in the dataset is 98.
- It seems that people over 65 are the best demographic for the bank to advertise this product.

Financial Indicators - Client



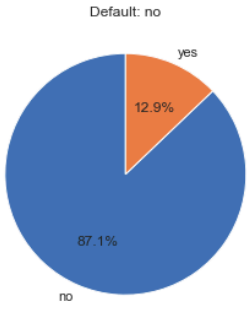


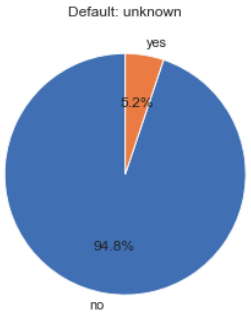


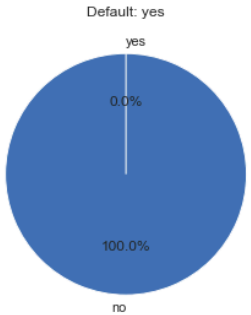


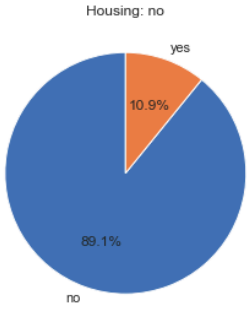
- These are the variables that offer information about the financial status of the client, as well as their history with the bank's other products: 'default', 'housing', 'loan' and 'poutcome'.
- Only 3 clients had previously defaulted on a loan; almost 0%.
- Majority (52%) clients have a housing loan.
- Majority (82%) client do NOT have a personal loan.
- Although majority (86%) outcomes are not known for the previous campaign, the negative responses (10%) were about three times as many as positive (3.3%).

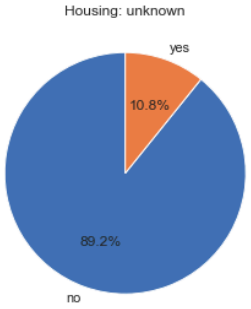
Financial Indicators - Client

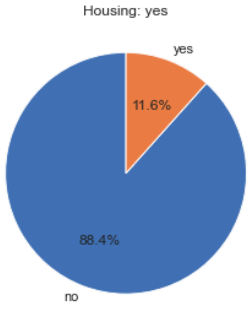


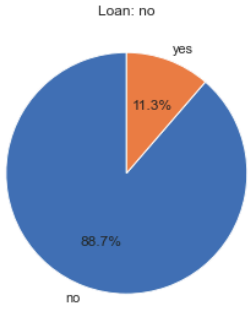


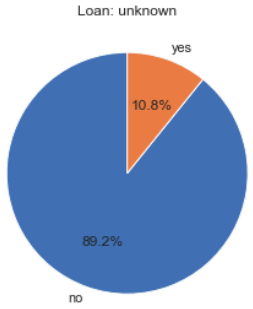


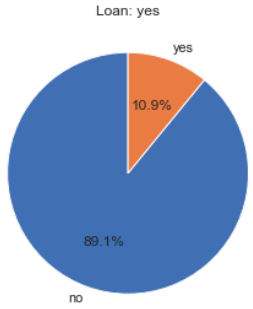


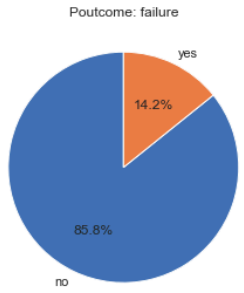


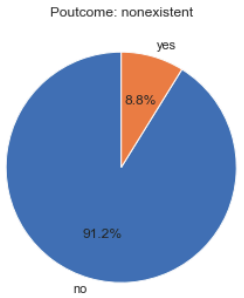


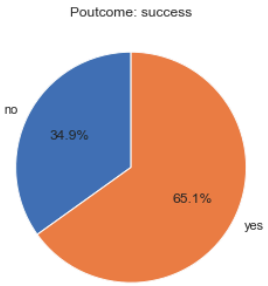






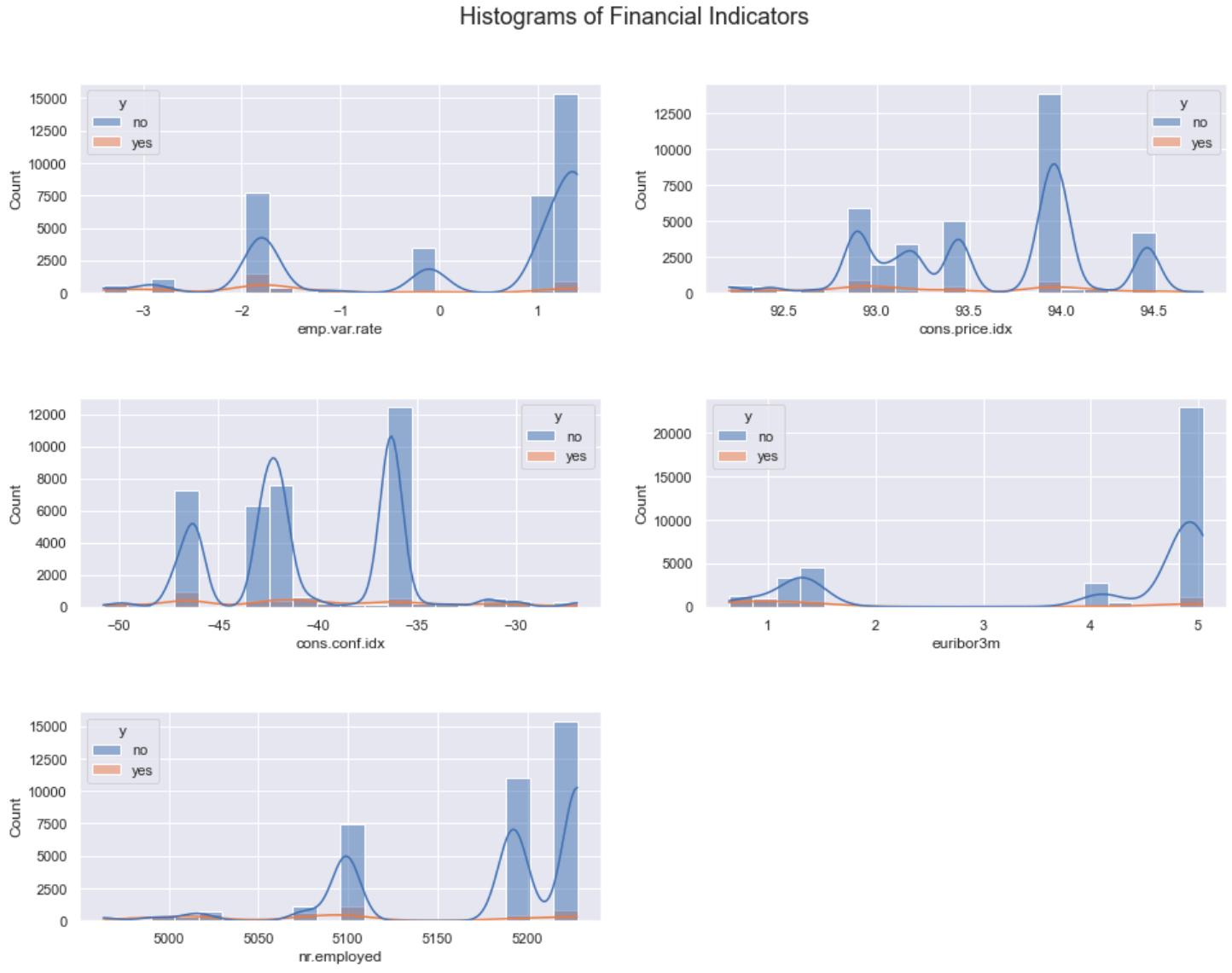






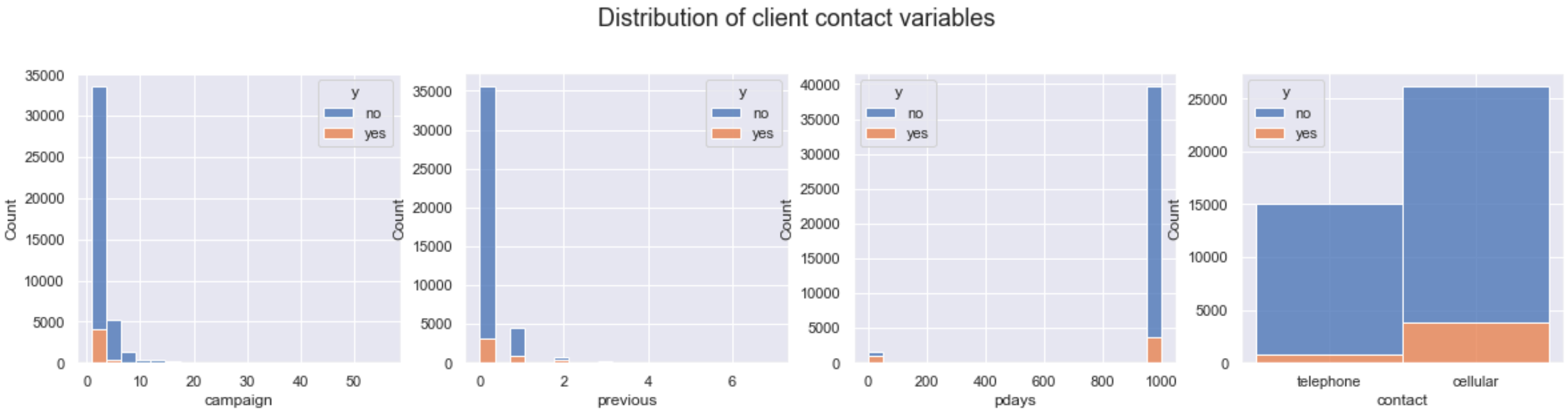
- We can look at the break-up of responses within each sub-category of our financial indicators.
- For ‘housing’ and ‘loan’, the response to the campaign doesn’t seem to be affected by the clients standing. The percentages are about the same for ‘yes’, ‘no’ and ‘unknown’.
- For ‘default’, there is a difference in response between ‘yes’ and ‘unknown’. However, since we do not know the unknown values, not much can be said at this point.
- For ‘poutcome’, it seems that clients who responded positively to the previous campaign are far more likely to respond positively to the current one. This is an important finding that can help the bank focus advertising efforts.

Financial Indicators - Economy



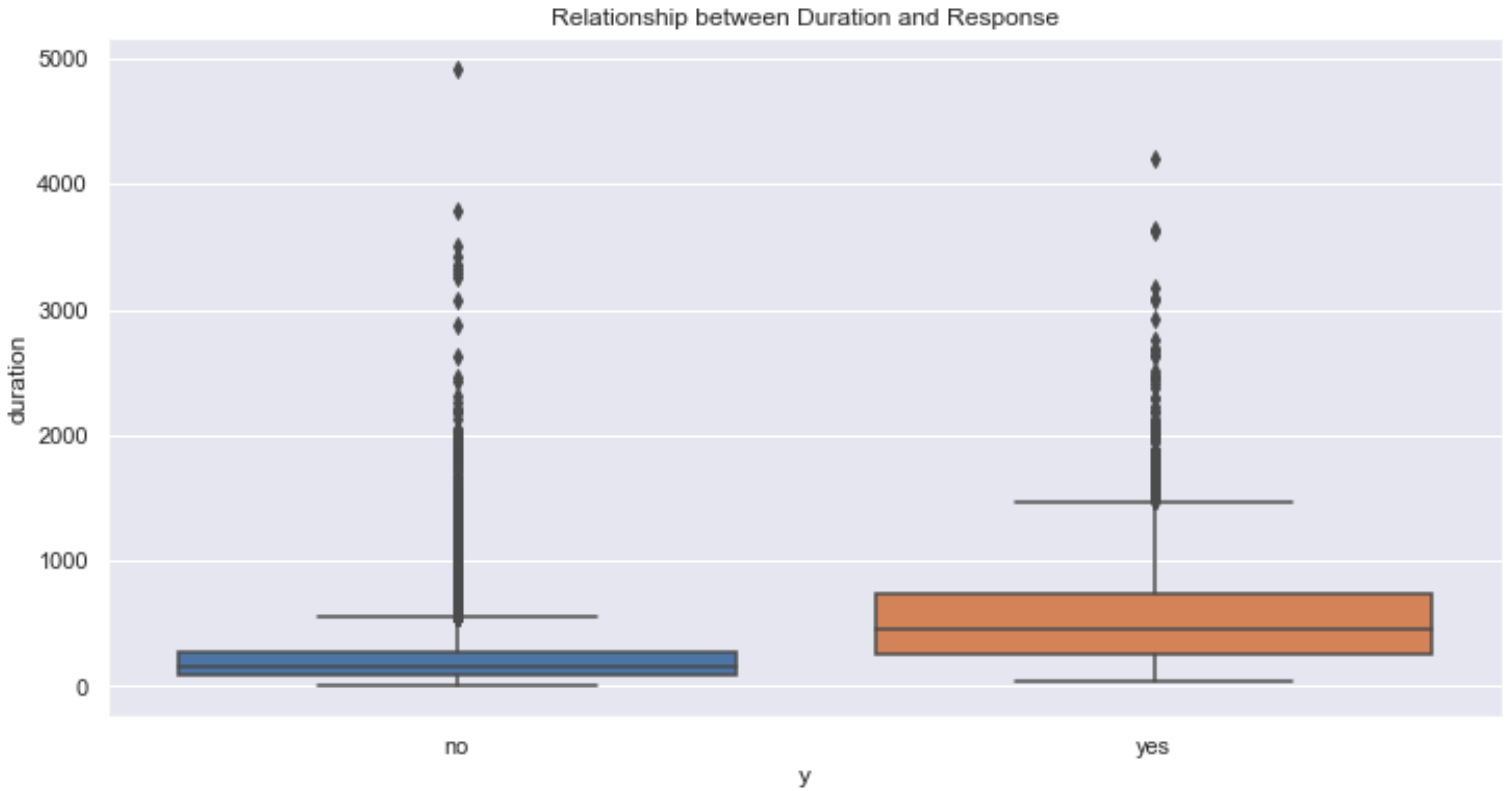
- **emp.var.rate**: A higher employment variation rate seems to correspond with a higher concentration of 'no' outcomes, while lower rates show a more balanced distribution between 'yes' and 'no' outcomes, indicating that clients may be more inclined to subscribe when the employment rate is stable or improving.
- **cons.price.idx**: Clients with a 'yes' are slightly more concentrated around lower CPI values, while clients with a 'no' have a relatively uniform distribution, suggesting that price level stability or a lower inflation rate may influence clients' decisions to subscribe to a term deposit.
- **cons.conf.idx**: Clients with a 'yes' are more likely to be associated with lower CCI values, while clients with a 'no' are more concentrated around higher values, implying that clients may prefer term deposits during times of pessimistic economic outlooks.
- **euribor3m**: A lower Euribor 3-month rate is associated with a higher number of 'yes' outcomes, while higher rates correspond with a higher concentration of 'no' outcomes, indicating that clients may be more inclined to subscribe to a term deposit when interest rates are lower.
- **nr.employed**: The distribution of 'yes' outcomes is skewed towards lower levels of employment, while 'no' outcomes are more concentrated at higher levels, suggesting that clients may be more likely to invest in term deposits when the number of employed people is lower, possibly due to concerns about job stability and financial security.

Client Contact



- **Campaign**: Most clients were contacted fewer than 10 times during the campaign, and clients with fewer contacts seem more likely to subscribe to a term deposit. The success rate appears to decrease as the number of contacts increases.
- **Previous**: The majority of clients had not been contacted in a previous marketing campaign (zero previous contacts). Clients with a higher number of previous contacts seem to have a slightly higher success rate, though the overall numbers are small.
- **Pdays**: A large proportion of clients were not contacted before (999 indicates no previous contact). For those who were contacted previously, the success rate appears to be higher when the number of days since the last contact is lower.
- **Contact**: The 'cellular' communication type has a higher overall success rate for term deposit subscriptions compared to the 'telephone' communication type. The 'cellular' type also has a higher number of clients in the dataset.

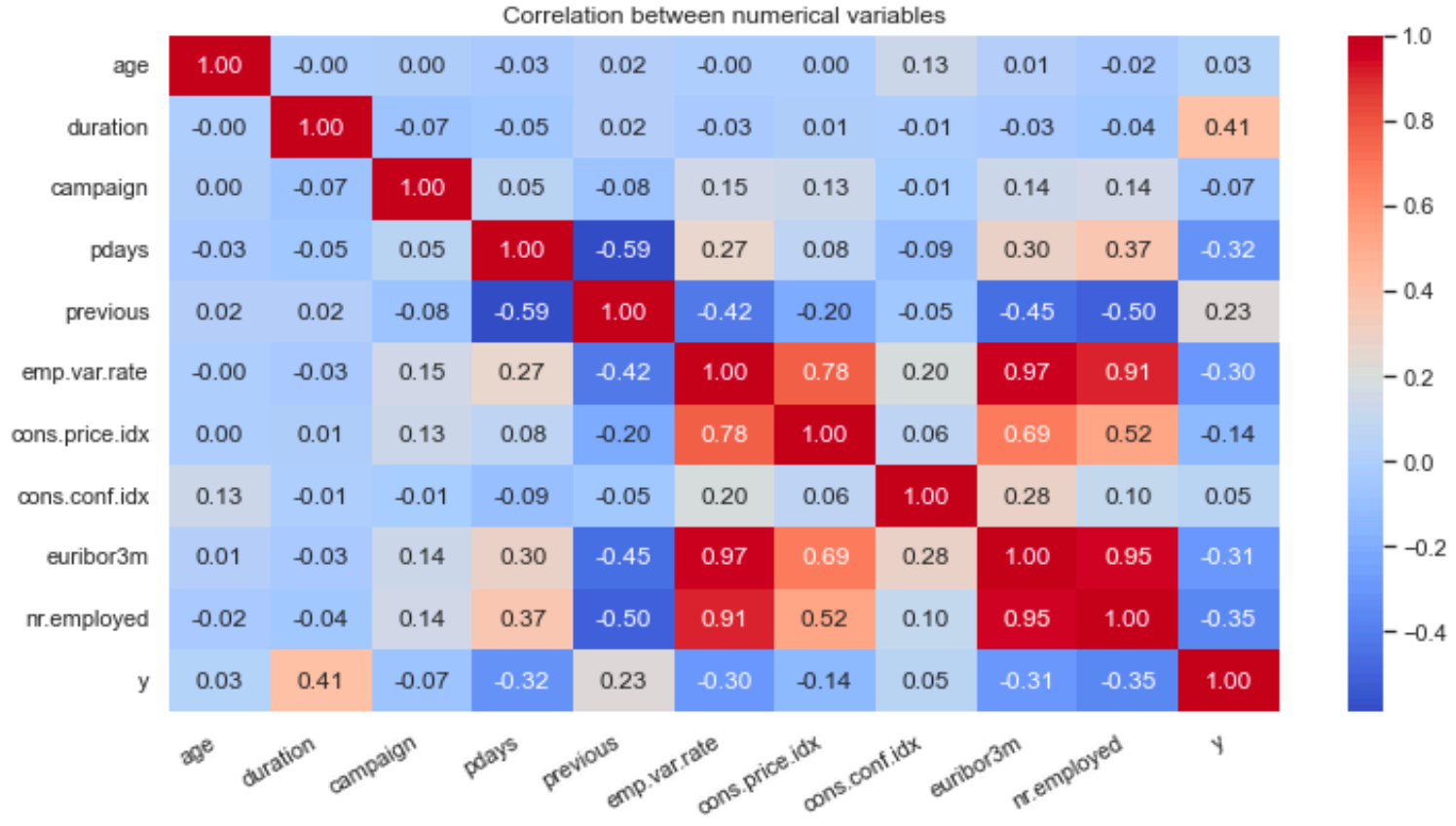
Contact Duration



- The 'yes' response group has a higher median duration compared to the 'no' response group, indicating longer calls are generally more successful in securing a subscription.
- The range of duration for both 'yes' and 'no' groups is quite large, suggesting there is variability in call durations.
- Both groups have several outliers, especially in the 'no' response group, indicating there are some exceptionally long calls that did not result in a subscription.
- The box plot suggests that a longer call duration may be positively correlated with a higher likelihood of securing a subscription, but other factors are likely involved as well.

The Duration feature is a bit tricky because it is only known once the outcome (y) is also known. In a real-world scenario, we would not know the duration of a call beforehand. Including this feature in predictive modelling would lead to data **leakage**, as it introduces information from the future that would not be available at the time of prediction. However, we can draw the conclusion that there is a positive correlation between call duration and positive outcome and this insight can be used to inform the bank's client outreach protocols. We can also add it as a feature for benchmarking model performance.

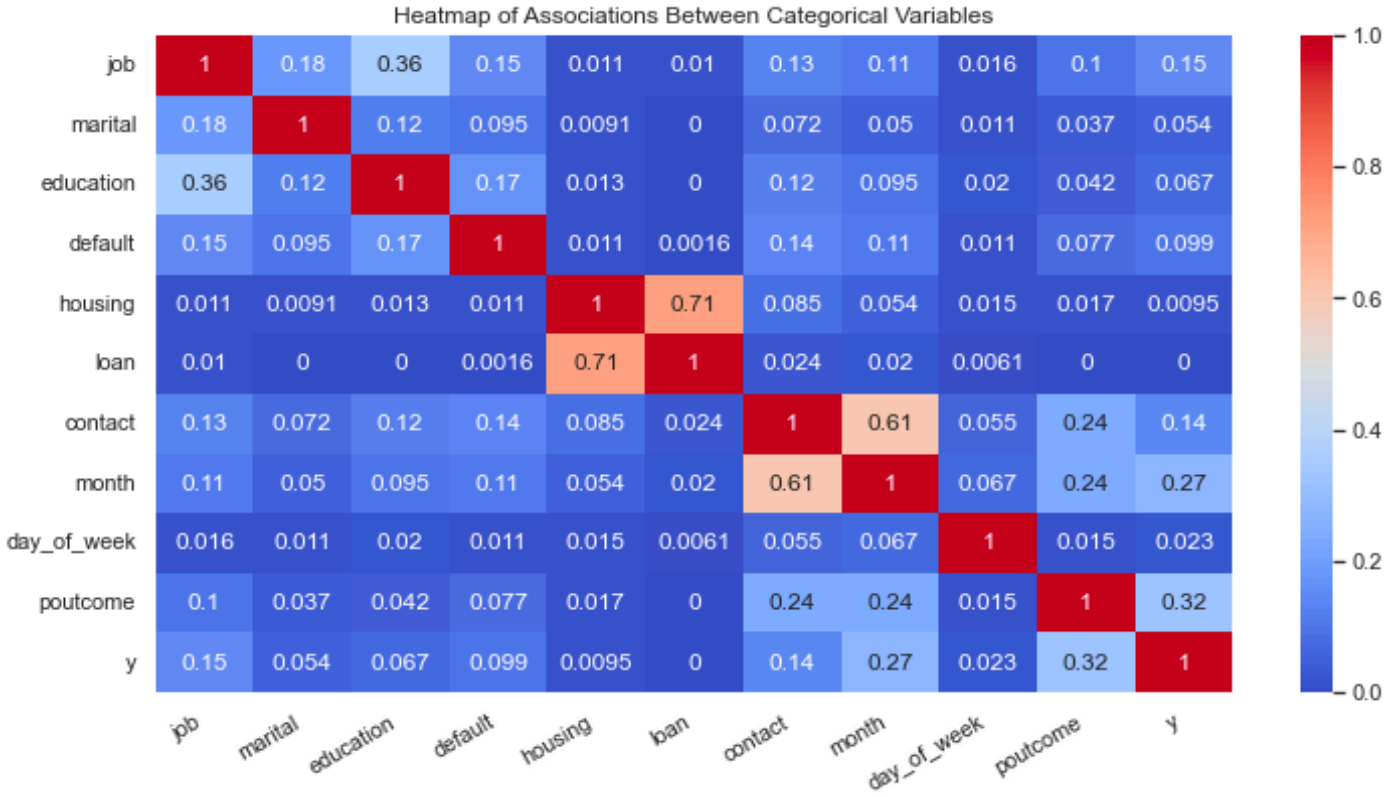
Numerical Variables



- emp.var.rate has a strong positive correlation with euribor3m (euribor 3-month rate) and nr.employed.
- cons.price.idx has a moderate positive correlation with emp.var.rate and euribor3m.
- cons.conf.idx has a weak negative correlation with most of the other numerical variables, except a moderate negative correlation with euribor3m.
- pdays has a weak negative correlation with the target variable y (response to the marketing campaign).
- previous (number of contacts performed before this campaign) has a weak positive correlation with the target variable y.
- Most of the other numerical variables have weak correlations with the target variable y, with nr.employed and euribor3m having the strongest negative correlations among them.

Overall, the heatmap shows that some macroeconomic indicators (employment variation rate, consumer price index, euribor 3-month rate, and the number of employees) have relatively strong correlations among themselves. However, these numerical variables have weak correlations with the target variable, indicating that they might not be strong predictors of the response to the marketing campaign.

Categorical Variables



- job and education have a moderate association (around 0.36). This indicates that people with certain job types tend to have specific education levels.
- housing and loan have a very strong association (0.71), meaning that it is likely a client with a housing loan will also have a personal loan and vice-versa.
- There is also a very high association between month and contact. It cannot be said whether this is due to some intrinsic cause or simple coincidence. It might simply have to do with the fact that most contact were made during particular (summer) months.
- The target variable y has weak associations with all categorical variables, except poutcome and month. It is easy to see why outcome from previous campaign would be correlated to outcome from current campaign. As for the association with month, once again this can be caused by the fact that the majority of contacts were made during the summer months.
- Most other pairs of categorical variables have weak associations, implying that the majority of categorical features in this dataset are not strongly related to each other.

Cramer's V statistic is a measure of association between two categorical variables. It is based on the chi-square statistic and takes values between 0 and 1. A value of 0 indicates no association between the variables, whereas a value of 1 implies a perfect association or dependency between them.

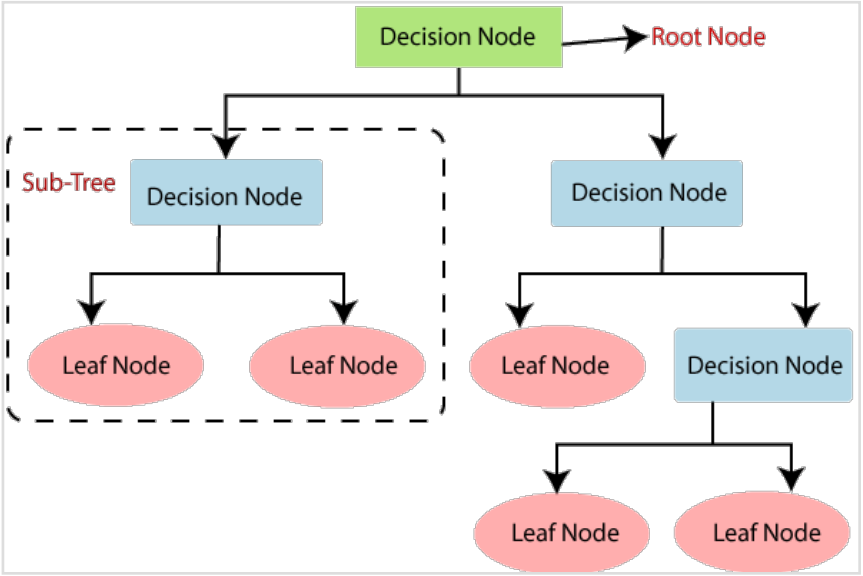
Recommendations

1. **Focus on high-potential customer segments:** The EDA showed that certain customer groups (e.g., specific age groups, job categories, or education levels) have a higher likelihood of subscribing to the product. Target marketing efforts towards these segments to improve conversion rates.
2. **Optimise contact strategies:** The analysis revealed that the success rate varies depending on the communication channel used (cellular or telephone) and the month of contact. Optimise the contact strategy by using the most effective channels and timing to reach customers.
3. **Leverage the insights from previous campaigns:** There is a relationship between previous campaigns, previous outcomes, and the number of contacts made with the customer. Use these insights to refine the approach for future campaigns and avoid over-contacting customers who are less likely to convert.
4. **Monitor economic indicators:** The EDA showed that some economic variables, such as the employment variation rate, consumer price index, and consumer confidence index, are correlated with the target variable. Monitor these indicators to identify potential opportunities or challenges in the market and adjust the marketing strategy accordingly.
5. **Consider the impact of duration on modelling:** Since the call duration is highly correlated with the target variable, it's essential to be cautious when using this feature for predictive modelling. It's recommended not to include this variable in the model, as the duration of a call is not known before the call is made.
6. **Address data quality issues:** The EDA identified some missing or inconsistent data in certain columns. Consider using data imputation, outlier treatment, or other data cleaning techniques to improve the quality of the dataset and ensure more reliable insights.
7. **Further investigate low correlation variables:** Some categorical variables have a low association with the target variable, as measured by Cramer's V. Consider additional feature engineering or selection techniques to better capture the relationship between these variables and the target, or remove them from the analysis if they prove to be irrelevant.

Modelling Suggestions

Classification algorithms can be broadly categorised into several main classes. Here are some of the most common ones:

1. Linear Models: Logistic Regression, Linear Support Vector Machines (SVM)
2. Decision Trees: Decision Tree Classifier, Random Forest Classifier
3. Bayesian Classifiers: Naive Bayes Classifier, Gaussian Naive Bayes
4. Neural Networks: Multi-layer Perceptron (MLP) Classifier, Deep Learning (Convolutional NN, Recurrent NN, etc.)
5. Instance-Based Methods: K-Nearest Neighbours (KNN) Classifier
6. Support Vector Machines (Non-linear): Kernel Support Vector Machines (SVM), Radial Basis Function (RBF) SVM
7. Ensemble Methods: Bagging (Bootstrap Aggregating), Boosting (AdaBoost, Gradient Boosting, etc.)



Decision Tree Architecture

It is recommended to train and compare all models for best outcome. However, from past experience, Decision Tree models have proven to be highly accurate and cost effective. They are quick to train, compared to Non-Linear SVM's, Neural Networks, KNN and Ensemble methods. In terms of accuracy they beat the Linear Models and most others. Another promising approach is the Naive Bayes Classifier. Although it is incredibly quick to train it is not very accurate. AdaBoost is another method that has shown great promise in classification problems such as this. Lastly, it is important to remember that the data is heavily imbalanced, (88.7% no), so we might need to employ resampling techniques such as SMOTE so ensure that the model is able to learn well and generalise.

Decision Tree Classifiers

Decision tree models offer several advantages in classification tasks. Some of the key benefits include:

1. **Interpretability:** Decision trees are easy to understand and interpret, as they mimic human decision-making processes. The tree structure visually represents a series of decisions based on feature values, making it simple to explain the model's logic to non-experts.
2. **Minimal data preprocessing:** Decision trees can handle both numerical and categorical variables and don't require extensive data preprocessing, such as feature scaling or normalisation. They are also less sensitive to outliers compared to some other models.
3. **Handling missing data:** Decision trees can handle missing data more effectively than some other models, as they can make splits based on the available data and create surrogate splits to manage the missing values. This strength of DT will come in handy since our data as ‘unknown’ values in 6 columns.
4. **Nonlinear relationships:** Decision trees can capture nonlinear relationships between features and the target variable, which might not be easily captured by linear models.
5. **Feature selection:** Decision trees perform automatic feature selection as part of the model building process. Features that contribute more to the target variable prediction will appear higher up in the tree, while less important features may not be included at all.
6. **Parallelisable:** Decision tree algorithms, particularly ensemble methods like random forests, can be easily parallelised, leading to faster training times on multi-core machines or distributed systems.

However, decision tree models have some limitations as well, such as their propensity to overfit and their sensitivity to small changes in the training data. Ensemble methods like random forests or gradient boosting machines can help address some of these issues by combining multiple trees to improve generalisation and reduce overfitting.

Thank You