**Spark – Mini Project #2**



**Date:** 29 November 2023

**Team Name:**

MTTDATA

**Team Members:**

Brock Bennett

Isabella Lieberman

John Ramirez

Nathan Zlomke

Table of Contents

[Executive Summary 3](#_Toc152081867)

[Data Preparation 7](#_Toc152081868)

[Predictive Modeling 7](#_Toc152081869)

[Feature Importance 9](#_Toc152081870)

[Clustering and Segmentation 11](#_Toc152081871)

[Insights & Recommendations 11](#_Toc152081872)

# **Executive Summary**

MTTDATA (Meet the Team DATA) has been tasked to uncover insights for XYZ Bank that affect the classification outcome that identifies clients who will subscribe yes or no for a term deposit. **XYZ Bank has conducted direct marketing campaigns in hopes of securing new term deposit clients. The purpose of making classifications is to predict customer response for XYZ Bank’s term deposits. One of XYZ Bank’s objectives through this project is to better understand the customer behavior to enable better financial forecasting and decision-making for their leadership team.**

**Problem Statement**

XYZ Bank wants MTTDATA to conduct Exploratory Data Analysis (EDA) to identify relationships and trends within the data whether that is through correlations, bivariate analysis of target versus input variables, facts, univariate patterns, or missing data. XYZ also asks MTTDATA to develop and save a predictive model for XYZ Bank to roll out for future use. For MTTDATA to meet this objective is by exploring different techniques and sharing the findings about the approach and benefits of the champion model. Achieving exploratory data analysis, various machine learning approaches, and K-means clustering for customer segmentation is paramount for MTTDATA to complete for succession. To note, XYZ Bank’s preferred method of getting the task triumphantly completed is by utilizing Apache Spark (Spark) as the primary tool used for analysis, so showing great prowess in Spark is another vital part of the succession for MTTDATA.

**Background**

**The dataset provided by XYZ Bank is from May 2008 to November 2010. It is worth noting that “the housing market crash of 2008 was a catastrophic event in the history of the United States housing market, leading to a severe economic recession that impacted millions of Americans. The crash was primarily caused by a combination of factors, including the subprime mortgage crisis, increased levels of debt, and a lack of regulation in the financial sector (Santarelli, 2023).” This means that the data will reveal the customer's behaviors as it relates to the impact of the 2008 Housing Crisis at that time and the aftereffects. It is crucial for MTTDATA to consider the surrounding national news in mind, as it can heavily influence the insights and recommendations generated from the data.**

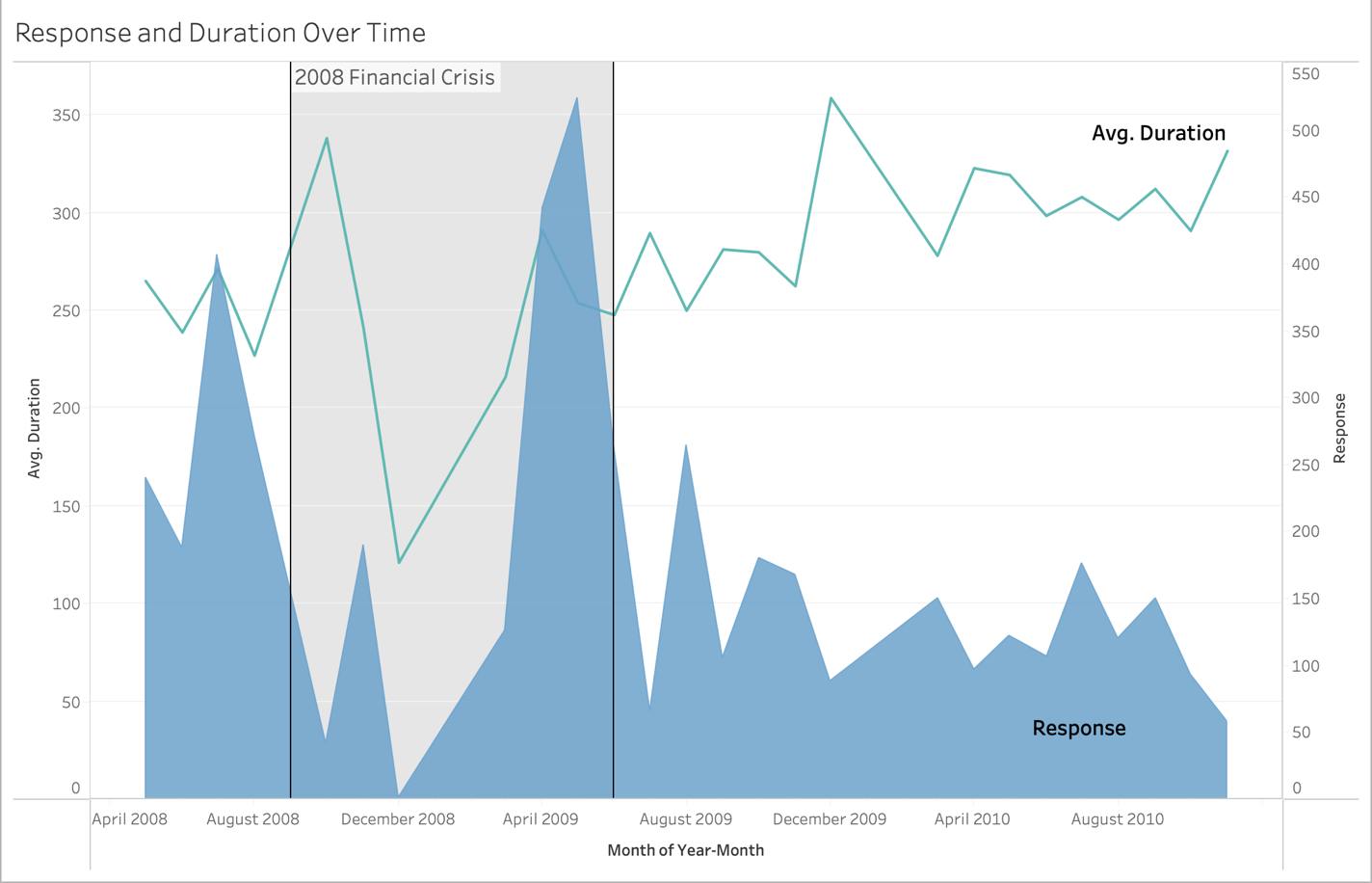
**Exploratory Data Analysis**

Spark in Jupyter Notebooks was the primary analytical tool as requested by XYZ Bank utilized in this project with GitHub version control and README to provide documentation. Upon inspection, it does not appear that there are significant data cleansing requirements that require missing value imputations, nor does it appear that erroneous data is present in the dataset provided.

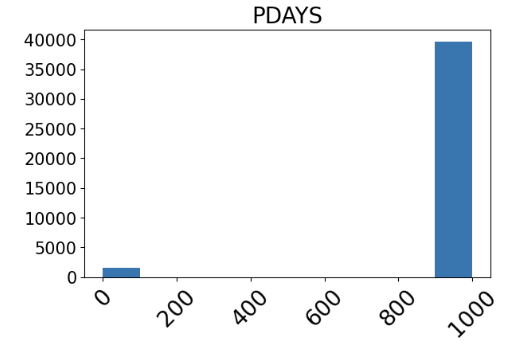
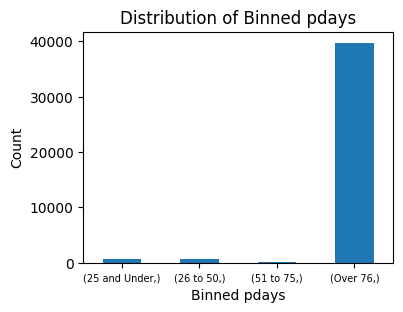
There are 41,188 rows and 21 columns in the original dataset. Looking at the month and period contextually, there are no data provided for January and February 2009 and 2010. For numerical features, summary statistics were gathered that computed mean, min, max, and the like. Categorical variable summary statistics provided mode, distinct count, and category frequencies. These are displayed in the appendix.

Response rate is quite low, with 12.7% of clients purchasing term deposits. In general, response rates for telephone campaign response rates range from 10-20% (Connors, n.d.).

Before descriptive statistics were explored, a time series graph was plotted that displays the response over time. Discussed in the model below, Duration is the most important feature and was also included in the graph to visualize the relationship. The period of the 2008 Financial Crisis is highlighted in grey and demonstrates extreme variability in response rates. There appears to be a similar trend pattern emerging between duration and response rates. When the economy normalizes, response rates are less responsive to phone calls duration and become flatter.

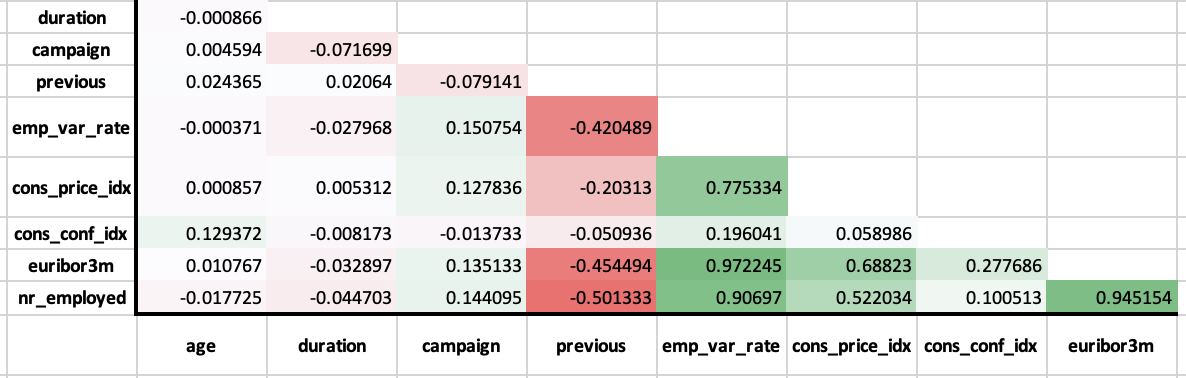


In addition, distributions were plotted for numerical features, with each feature provided in the appendix. From looking at the charts, it is apparent that Pdays is extremely skewed, as displayed below.

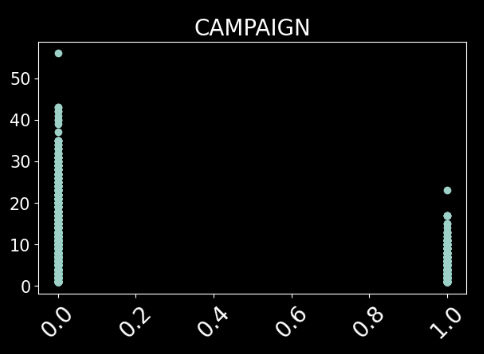
 

Considering this, binning was employed as a feature engineering technique to attempt to remedy this issue. Although most of the values for this feature were ‘999’ or, not previously contacted, binning allowed other relationships to emerge and transformed this feature from numerical to categorical. This can help reduce an outlier’s influence on a model; a decision tree model, for example, could easily partition this category out.

Pearson correlation coefficients were calculated and are displayed in the heatmap below. There are strong positive correlations between euribor3m and nr\_employed as well as most other economic indexes, excluding cons\_conf\_idx. The number of previous contacts before this campaign has a strong negative association with most other columns.



**Scatterplots were plotted for numerical features, as seen in the appendix. Most features had common variability across positive and negative responses, but the campaign had a noticeably different response, displayed below. Positive responses are more concentrated for lower campaign contacts, curtailing around 15-18 contacts.**



**Contingency Tables were plotted for both numerical and categorical features, seen in the appendix. This is one of the most effective tools for understanding relationships with responses. To summarize some of the features:**

* **When employee variation rate is at its lowest, there is extremely poor response**
* **Almost all purchases are when contacts in a campaign are ideally 1 and at most 3**
* **Most positive responses were when there was no previous campaign**
* **High school and university degree were most responsive**
* **About 83% of positive responders have no loan**
* **Vast majority of positive responders were using a cell phone**
* **Nearly all positive responses when more than** **76 days occurred since last contact**

.

# **Data Preparation**

To process data in Spark’s machine learning applications, there are certain data preprocessing steps that need to occur first. Categorical variables that are in string format were converted to integers, using StringIndexer and then OneHotEncoder cast these as a single vector of indexes in a single column per tuple. Subsequently, these features were combined with the numerical features into a single vector, using VectorAssembler. The target variable is indexed, and the predictor features are scaled using StandardScaler.

This project will utilize four different machine learning algorithms, and it was desirable to automate as many processes as possible. Thus, all the preprocessing steps mentioned were fed into a pipeline as stages, with a dataframe output that can be read into various machine learning models.

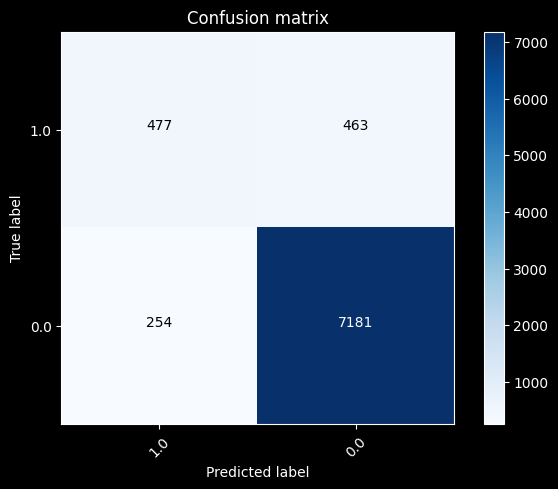
# **Predictive Modeling**

A train/test split of 80/20 was used, and each model utilized the same dataset. Four different models were compared to find the highest performing model. Since there is such a high skew of negative responses to positives, the models tend to predict negative cases more accurately when compared to true positives. To improve predictions, models used the Area Under the Precision-Recall Curve metric instead of Area Under ROC. The former tries to find tradeoffs between precision and recall, while the latter emphasizes positive classifications and is more appropriate for balanced datasets.

The four models consisted of logistic regression, gradient boosting, random forest, and a basic decision tree. Though a neural network model would likely handle complex nonlinear relationships well, it was decided that models with practical interpretability are more valuable for making recommendations.

Confusion matrixes were plotted for each model, as seen in the appendix. A summary table of each model’s metrics is provided below as well as the confusion matrix for the winning model, Gradient Boosting.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Precision | ROC-AUC | Accuracy | F1 Score |
| Logistic Regression | 0.59609 | 0.93548 | 0.909134 | 0.514 |
| Gradient Boosting | 0.64289 | 0.93930 | 0.914388 | 0.573 |
| Random Forest | 0.57202 | 0.91349 | 0.903284 | 0.336 |
| Decision Tree | 0.49864 | 0.79886 | 0.913552 | 0.561 |



As one can observe from the table, Gradient Boosting performs best in every accuracy metric evaluated. In terms of accuracy, the model performs quite well with 91.4% accuracy and an ROC score of .939; however, looking at the confusion matrix reveals that false negatives are nearly equivalent to true positives. Overall precision was 64.3%. When attempting to manage the disparity between accuracy and precision for imbalanced datasets, F1 Score is commonly employed. In this case, the F1 Score is 0.573. Optimal F1 Score is close to 1, so this model would benefit from several adjustments.

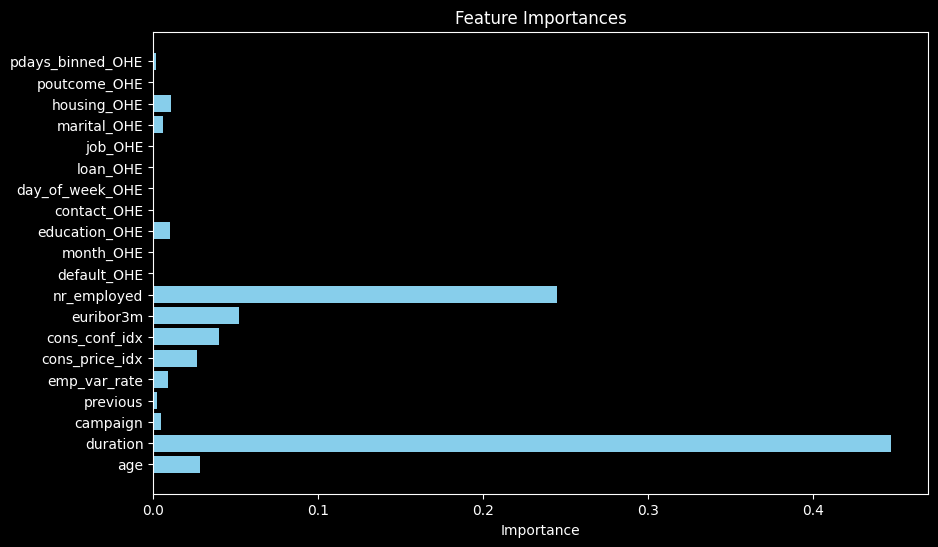
Gradient Boosting is an attractive model is such a complex case, as it can better capture complex nonlinear relationships and learn from weak performing trees. Additionally, though scaling was employed, it does not require scaling and is therefore robust in its ability to approach datasets that might be problematic for mathematical models such as logistic regression.

Suggestions for optimizing the model are as follows:

* Address the class imbalance with resampling techniques
* Investigate false prediction instances
* Further hyperparameter tuning
* Feature engineering approaches that further refine features where distributions are problematic through binning or power transformations

# **Feature Importance**

Feature importance is critical towards understanding which features have more significance in affecting classification outcomes. The feature importance graph is plotted below, and the top 10 features are as follows:



|  |
| --- |
| Top 10 Feature Importance: |
| Duration |
| Nr\_employed |
| Euribor3m |
| Cons\_conf\_idx |
| Age |
| Cons\_price\_idx |
| Education\_OHE |
| Housing\_OHE |
| Emp\_var\_rate |
| Campaign |

It is seen from the plot above that duration is of vital importance, suggesting that customers either need a relational connection with the product or company or they require education about the product and how it benefits them, especially amid varying economic contexts. It is not yet understood which comes first: the customer is on the phone longer, because they are interested in the product, or the longer they are on the phone speaking with a representative, the more interested they become in the product. Conversely, those who are not interested in a term deposit would hang up quickly, or it may be that due to lack of meaningful exchange with the customer, they were not interested.

The next most prominent features were “nr\_employed,” “euribor3m,” and “emp.var.rate.” These are economic indicators that provide context during the campaign. The number of employed variable is a quarterly job growth indicator that provides another index for economic context. The has an importance in affecting the classification outcome, which makes sense, as one desires job security before setting aside money for investments. Euribor3m has to do with the “Euro Interbank Offered Rate which is a benchmark in financial markets and affects rates on financial products; higher interest rates are indicative of a worse economy (CITE).” Employee Variation Rate is a quarterly indicator that shows how much employment has changed. The Employee Variation Rate for the contacted consumers ranges from –3.4 to 1.4.

Like the previous, Consumer Confidence Index and Consumer Price Index also provide economic context. The Consumer Confidence Index deals with consumers’ outlook on the economy, optimism, or pessimism. All negative values in the data tell us that consumers have a negative outlook on the economy and are saving money more than spending. The Consumer Price Index (CPI) is a monthly indicator that measures the average change over time in the prices paid by consumers for goods and services. The CPI was plotted with the response over time, shown in Appendix A. This metric indicates inflation if it is increasing. The graph shows that as interest rates go down during the crisis, CPI stabilizes for a time and then inflation picks up. As interest rates (i.e., euribor3m) decrease, often term deposit rates decrease, and vice versa. This may help explain why after the crisis responses to term loans reduce—inflation increases, and the lower interest rates also result in low term deposit rates.

Several notable features included demographic information about clients, like age, marital status, education, and housing that played minor roles in affecting the predicted outcome.

Remaining features include marketing campaign metrics, “campaign,” “previous,” and “Pdays.” “Campaign” and “previous” are how many contacts were made with the customer during and before the campaign, respectively. Pdays is how many days have elapsed since last contact. These also have a small magnitude in affecting the outcome but are in the top 10 features.

# **Clustering and Segmentation**

**Post model analysis was conducted using K-means clustering for the top 10 features. To determine k, an elbow plot was plotted, shown in Appendix C.**

The elbow appears at k=6, with a silhouette score = 0.615. A k=5 value was selected with silhouette score = 0.6080, as this is regionally close and simpler to analyze. Cluster profile metrics are displayed in the table in Appendix C. Cluster profiles are summarized in the table below.

|  |  |
| --- | --- |
| Cluster 1 | Young, optimistic spenders with brief calls exhibit surprisingly low campaign response during a positive economic period with high Euribor3m and strong job market growth. |
| Cluster 2 | With the second-longest call duration and job growth, Cluster 0 contrasts by having the lowest confidence index and age. Despite similarities, the longer calls and lower confidence lead to a notably better response rate in a good economy, indicating potential pessimistic investment opportunities. |
| Cluster 3 | With medium call duration and the second-lowest job growth, it shows a decent response overall. Notably, it responds well in a poor economy, driven by its unique combination of higher age and specific economic dynamics, presenting potential investment opportunities in challenging conditions. |
| Cluster 4 | Longest calls, high Euribor3m, peak consumer confidence, and age. Best response despite fewer contacts. Resembles Cluster 0 but older and has fewer contacts. In a good economy, this segment prefers longer conversations, hinting at potential investment interest during favorable times. |
| Cluster 5 | Second-lowest call duration, lowest job growth, and Euribor3m. Poor response, particularly during bad economies, signaling a reluctance to engage in financial matters while economic conditions are unfavorable. |

# **Insights & Recommendations**

It is important to note the economic context during this campaign. This telephone campaign’s duration was from May 2008 to November 2010, which preceded and followed after the 2008 economic collapse. Understandably, customer behavior was highly volatile at times during and after the crisis, making predictions more difficult. If a generalized model is the desired outcome, it would be advisable to use data from a less exceptional time to train a model; after all, "the global financial crisis...was the worst since the Great Depression (Norad).”

A strategy to take advantage of from this project would be to adjust marketing strategies depending on the economic context. The customer segmentation analysis provides insights for different approaches in certain scenarios.

Call duration is critical in terms of connecting with the customer relationally as well as educating them about the product. Economic indicators such as job security and interest rates also play crucial roles. Like above, cater strategies to the appropriate economic backdrop, noting cluster profiles. To ensure connections with customers and product education, explore customer engagement strategies and explore what drives long phone calls.

To optimize the model better, investigate false negative predictions to better identify positive responses. The model would also benefit from balancing the dataset by resampling techniques like oversampling. This increases minority class instances by duplicating them to better balance the dataset. Additionally, hyperparameter tuning will improve model performance.

**Duration**

**If XYZ calls someone, the longer they are on phone the more the customer is engaged. The quickest time for someone at XYZ Bank to get a yes was 37 minutes.**

**Number of Employees**

**The data processing was strange with this variable as when we got output, all it was reading was either “Null” or 5191. Further exploration on this variable is needed to get the most from the variable.**

**Euribor3m**

**A higher interest rate proves a worse economy which is on par due to the historical context as the data is set right in the middle of the 2008 Housing Crisis. All of the rates recorded were between 0.5% to 3.5% and 4% to 5%, Customers are most likely divided in those two buckets and further analyzing those buckets will generate great insight.**

**Consumer Confidence Index**

**All of the values were negative which means that consumers are saving money versus spending. It shows that consumers are not confident in the current market at that time.**

**Age**

**People who are older are more likely have reserved funds for a deposit. A lot of older and college aged people, early 20s, had term deposits.**

**Consumer Price Index**

**There was between a 5% and 8% decrease in the value of goods.**

**Education**

**In this variable, there were a lot of unknown meanings of the categorical inputs as one of the confusing parts to decipher was the difference between basic\_4yr, basic\_6yr, and basic\_9yr. No significance among the categorical inputs, so this variable is not relevant towards generating great insight for XYZ Bank.**

**Employment Variance Rate**

**Of those who said yes, have a negative employment variance rate, which makes sense as those who either lost their job or do not have one thought getting a term deposit was a sound financial decision.**

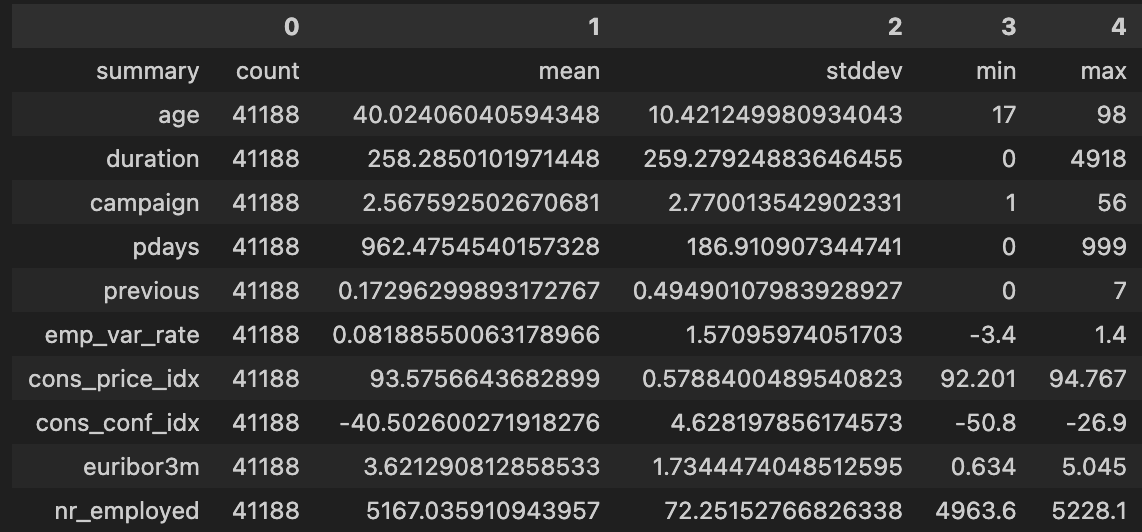
**Campaign**

**Most of the yes category came from one previous call, and decreased as more calls were made. It makes sense because if someone wants something they will say yes soon, as the more calls made to one client is wasted effort. Better optimization for XYZ Bank is to move on to newer clients as they will have a better potential of those saying yes instead of repeatedly calling the same clients.**

**Appendices**

**Appendix A – Exploratory Data Analysis**

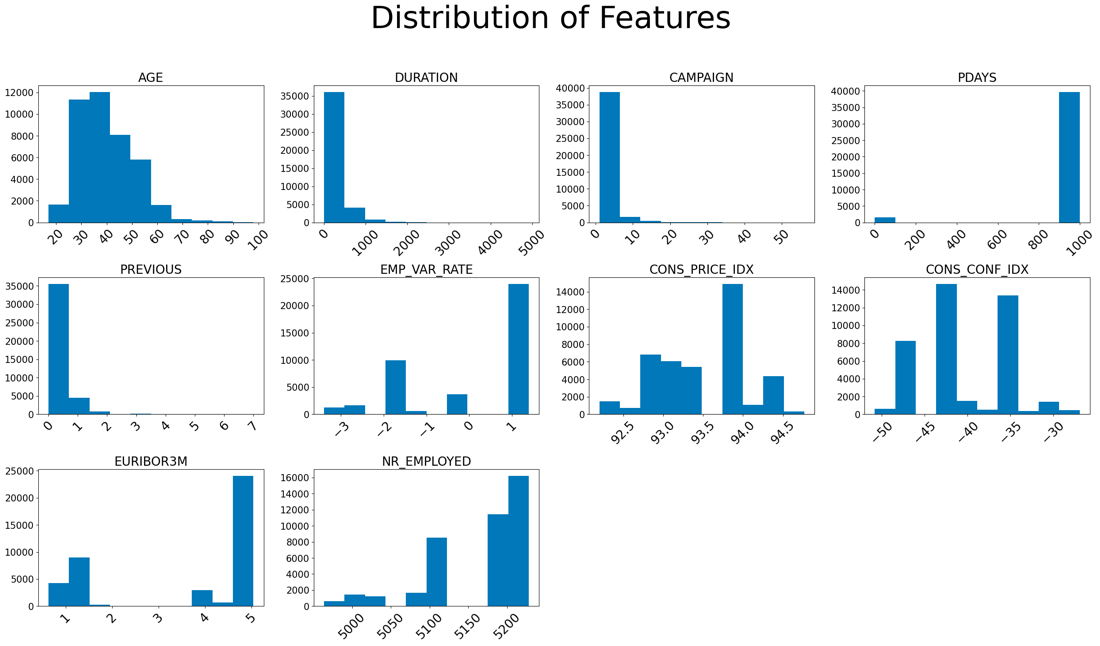
**Numerical summary statistics**



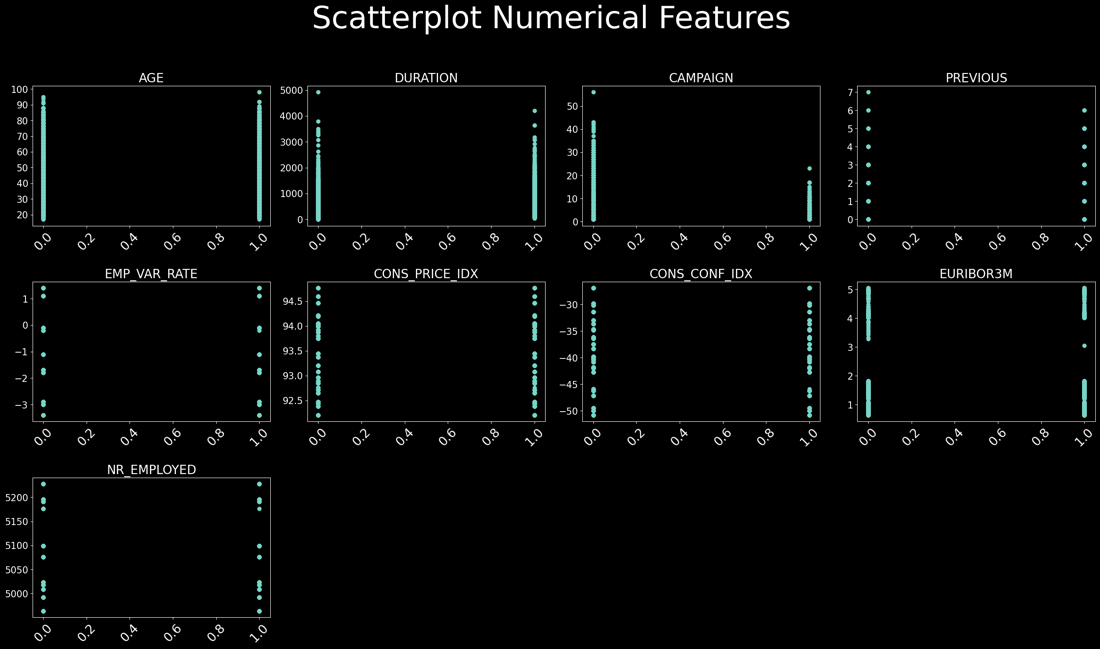
**Categorical summary statistics**



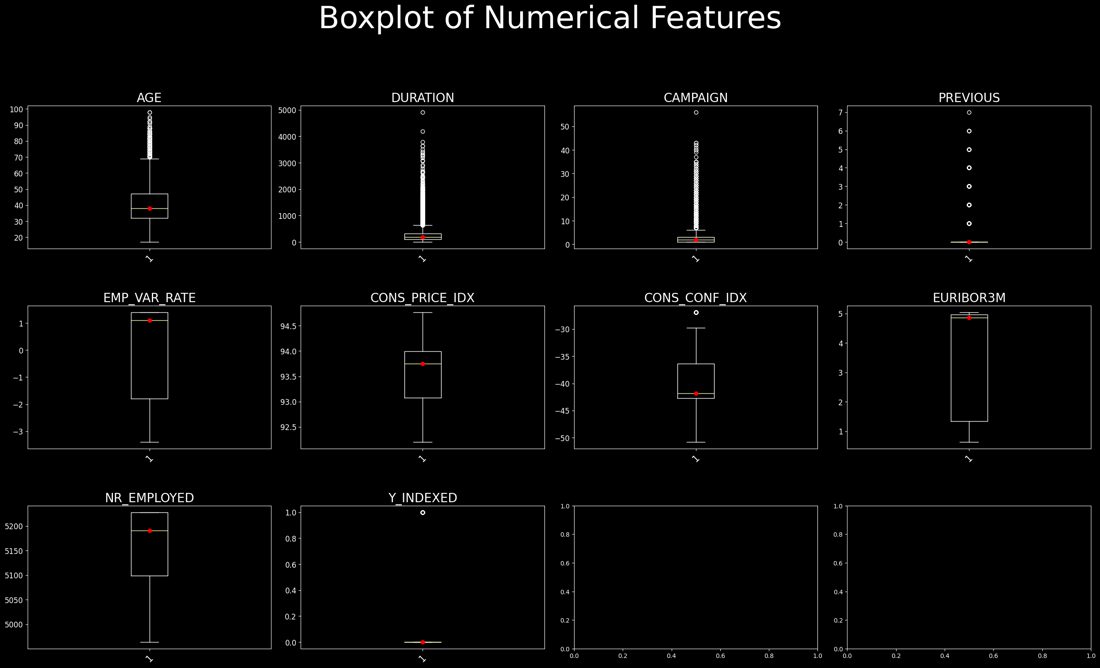
**Distribution of Numerical Features**



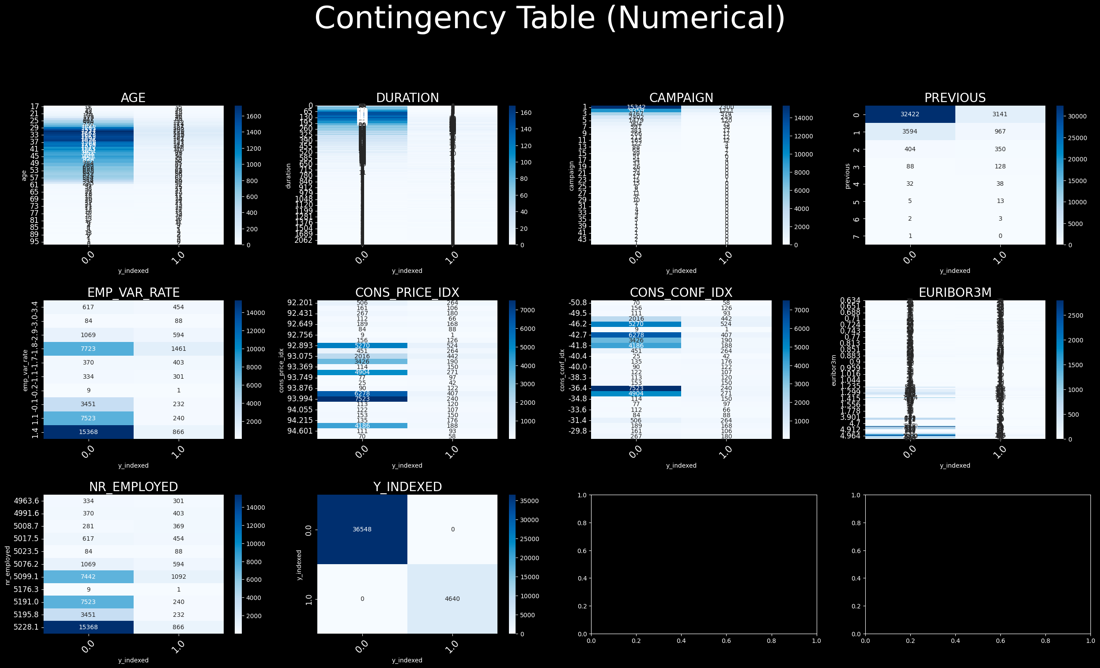
**Scatterplot of Numerical Features**



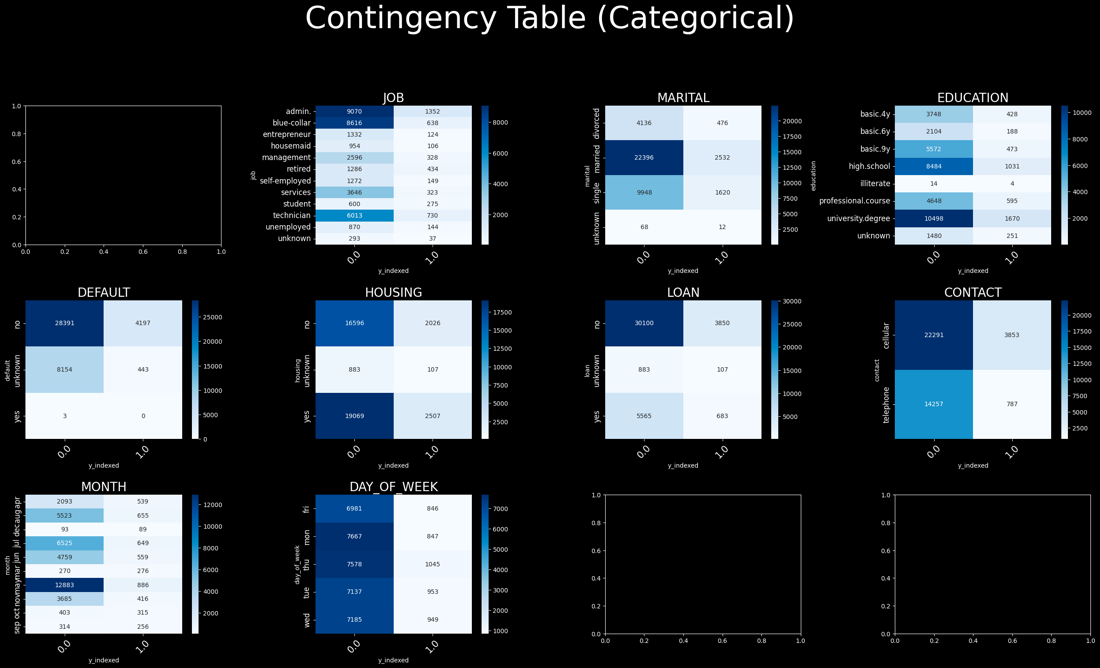
Boxplot for Numerical Features

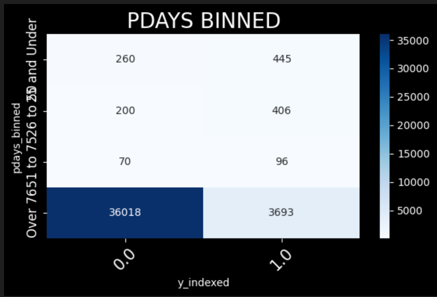


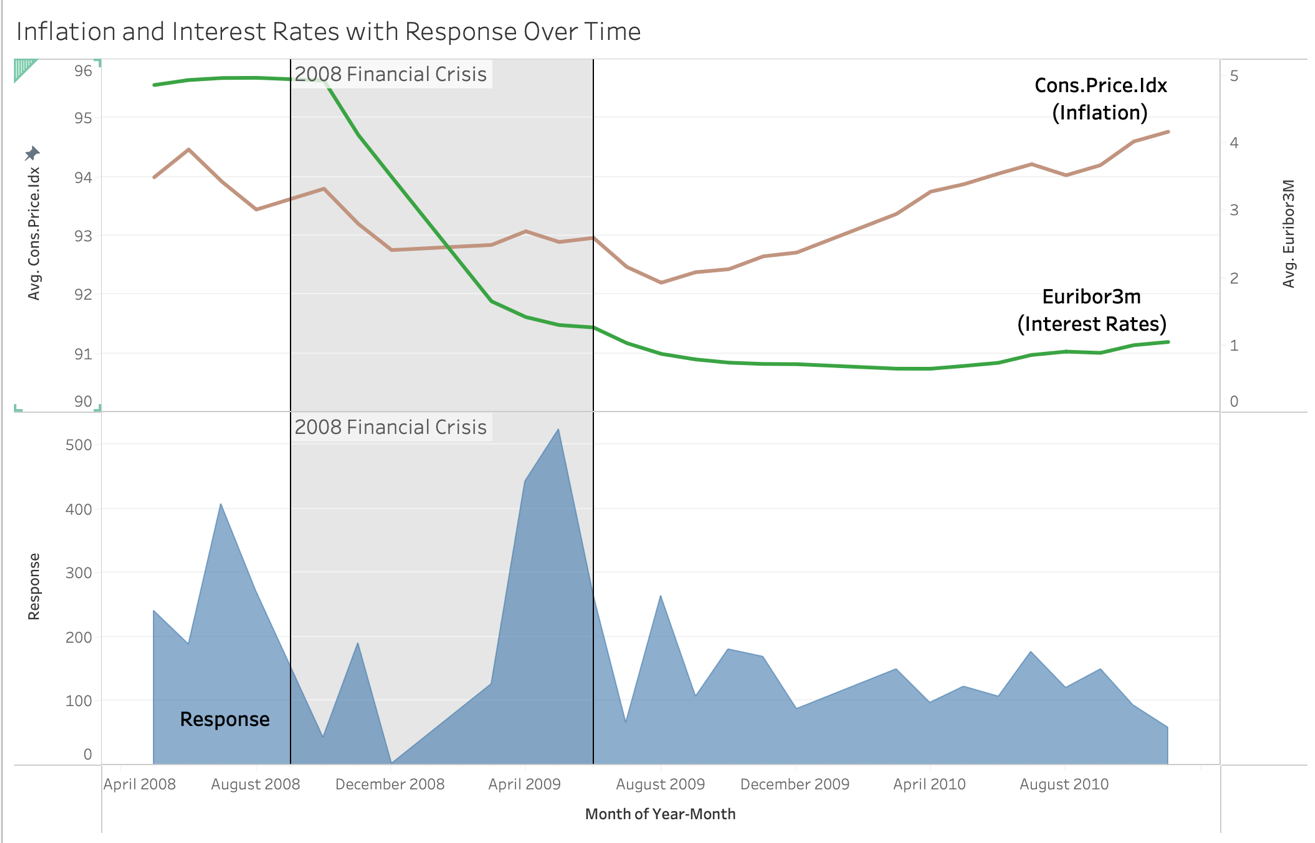
Contingency Table for Numerical Variables



Contingency Table for Categorical Variables

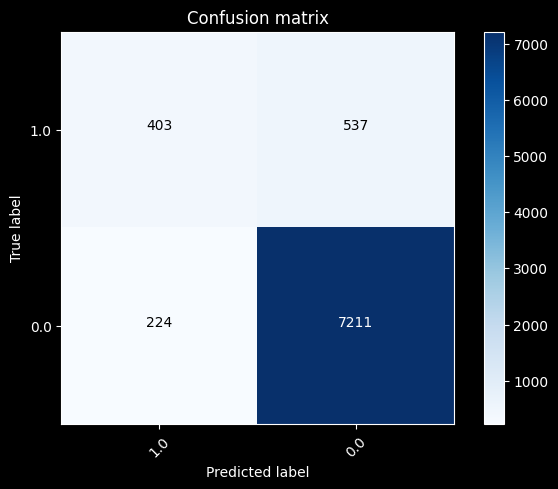




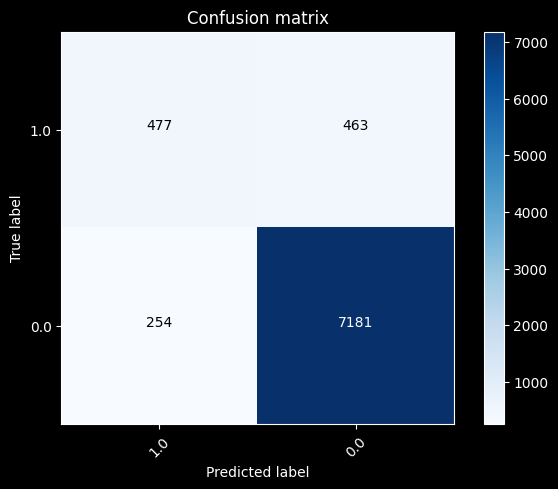


CPI (Inflation) and Euribor3m (Interest Rates) with Response Over Time

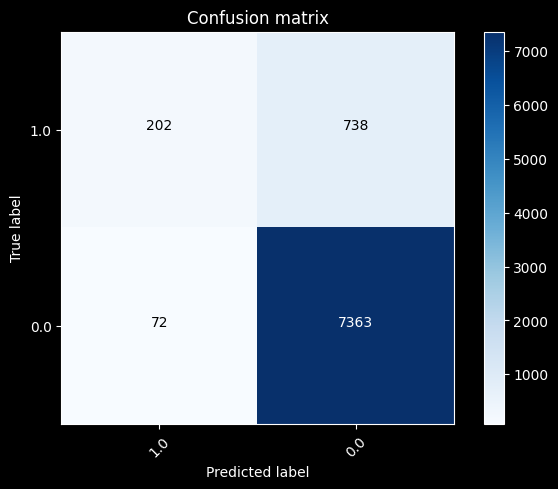
**Appendix B – Machine Learning Models**



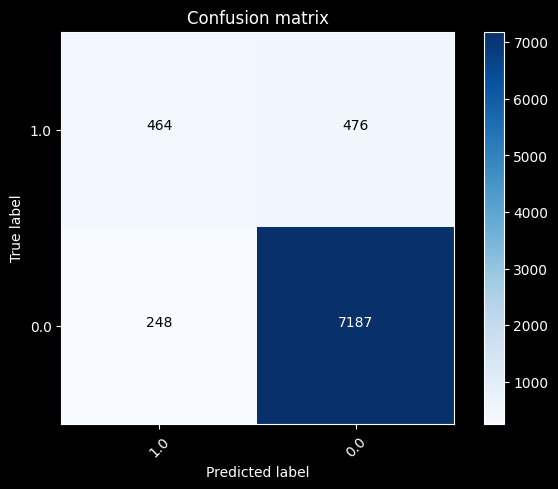
**Logistic Regression**



**Gradient Boosting**



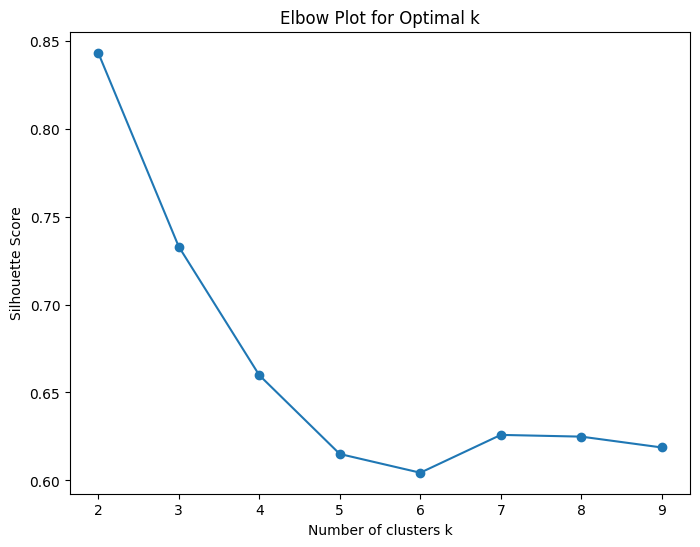
**Random Forest**

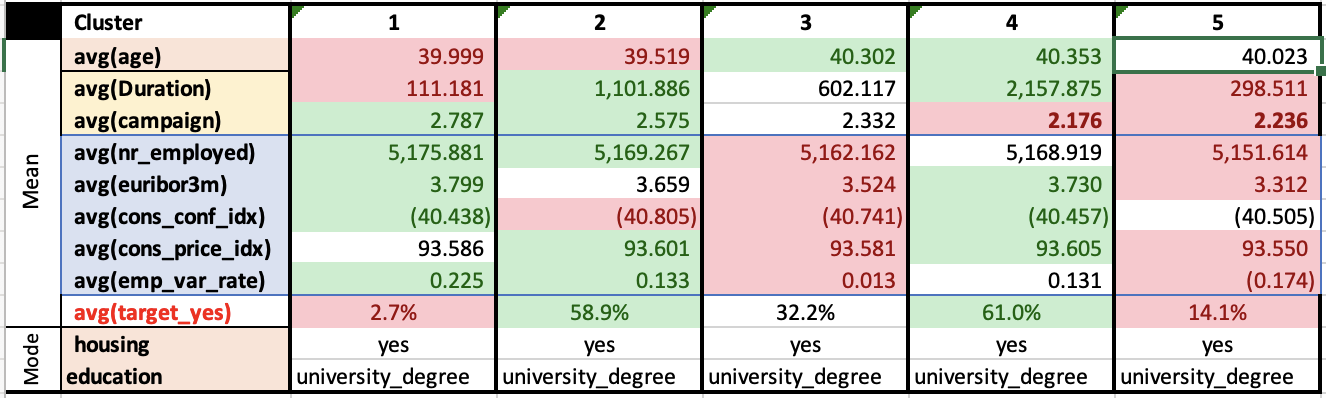


**Decision Tree**

**Appendix C – Clustering/Customer Segmentation**

**Elbow Plot for K-means**





The 5 clusters are presented in the table above. The values in green are the highest two in their respective features, contrasted with red values as the lowest values. Features highlighted in light blue are economic indexes, light orange are demographics, and yellow are campaign metrics.

**References**

Connors, J. (n.d.). *Typical response rates for direct marketing efforts*. Campaign Now. <https://www.campaignnow.com/blog/typical-response-rates-for-direct-marketing-efforts#:~:text=Campaign%20Now%20Response%20Rates&text=Telephone%3A%2010%E2%80%9320%25%20house,%3A%200.5%E2%80%933%25%20house%20file>

Santarelli, M. (2023, October 20). *Housing market crash 2008 explained: Causes & effects*. Norada Real Estate Investments. [https://www.noradarealestate.com/blog/housing-market-crash-2008/](https://www.noradarealestdddate.com/blog/housing-market-crash-2008/)