Final Project

Caleb Morse
MIS450
07/29/2024

1. Summary Statistics

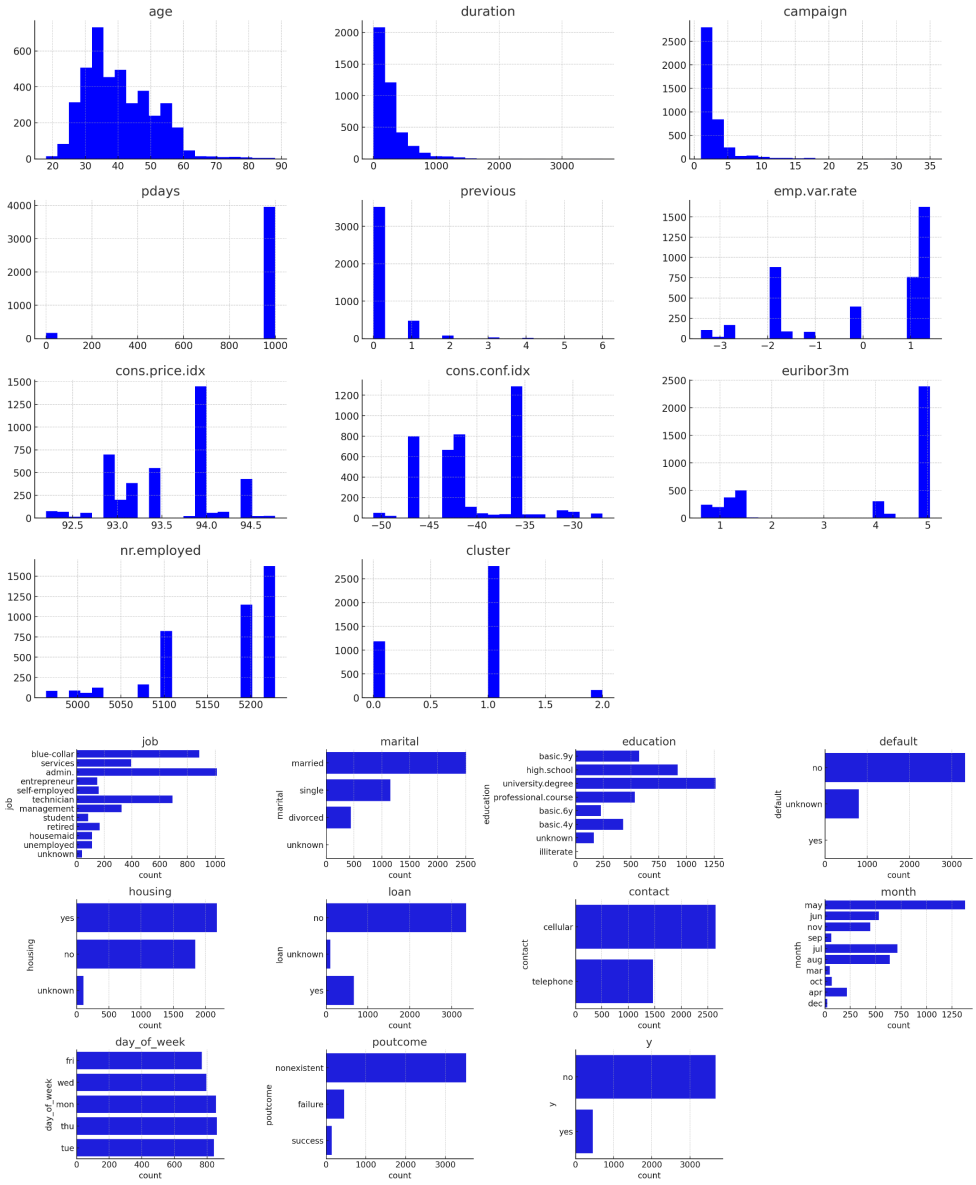
Numerical Columns:

- age:
    - Mean: 40.11 years
    - Std: 10.31 years
    - Min: 18 years
    - Max: 88 years
- duration:
    - Mean: 256.79 seconds
    - Std: 254.70 seconds
    - Min: 0 seconds
    - Max: 3643 seconds
- campaign:
    - Mean: 2.54 contacts
    - Std: 2.57 contacts
    - Min: 1 contact
    - Max: 35 contacts
- pdays:
    - Mean: 960.42 days
    - Std: 191.92 days
    - Min: 0 days
    - Max: 999 days
- previous:
    - Mean: 0.19 contacts
    - Std: 0.54 contacts
    - Min: 0 contacts
    - Max: 6 contacts
- emp.var.rate:
    - Mean: 0.085
    - Std: 1.56
    - Min: -3.4
    - Max: 1.4
- cons.price.idx:
    - Mean: 93.58
    - Std: 0.58
    - Min: 92.20
    - Max: 94.77
- cons.conf.idx:
    - Mean: -40.50
    - Std: 4.59
    - Min: -50.80
    - Max: -26.90

- euribor3m:
    - Mean: 3.62
    - Std: 1.73
    - Min: 0.63
    - Max: 5.04
- nr.employed:
    - Mean: 5166.48
    - Std: 73.67
    - Min: 4963.60
    - Max: 5228.10

Categorical Columns:

- job:
    - Most common: "admin." (1012 instances)
- marital:
    - Most common: "married" (2509 instances)
- education:
    - Most common: "university.degree" (1264 instances)
- default:
    - Most common: "no" (3315 instances)
- housing:
    - Most common: "yes" (2175 instances)
- loan:
    - Most common: "no" (3349 instances)
- contact:
    - Most common: "cellular" (2652 instances)
- month:
    - Most common: "may" (1378 instances)
- day_of_week:
    - Most common: "thu" (860 instances)
- poutcome:
    - Most common: "nonexistent" (3523 instances)
- y:
    - Most common: "no" (3668 instances)



2. The dataset provides valuable insights into the bank's marketing campaign and its clientele. Among the numerical variables, the age distribution shows that the majority of clients are middle-aged, with a concentration between 30 and 50 years. This suggests that the bank's marketing efforts are targeting or reaching a more mature demographic, likely those who are financially stable and have a greater need for term deposits.

Regarding the economic indicators, the employment variation rate ranges from -3.4 to 1.4 with a mean around 0.085, reflecting the employment variation rate during the campaign period. The consumer price index varies slightly, suggesting minor fluctuations in the cost of living, while the consumer confidence index (cons.conf.idx) ranges from -50.8 to -26.9, indicating general pessimism among consumers. The Euribor 3-month rate (euribor3m) ranges from 0.63 to 5.04, reflecting varying borrowing costs in the market. The number of employees (nr.employed) varies slightly, suggesting stable employment levels during the campaign period.

In conclusion, the dataset did not have any missing values, indicating it is complete and ready for analysis. The box plots revealed several outliers across various numerical variables, particularly in age, duration, campaign, and economic indicators. These outliers could represent genuine variability in the data or could be influential in predicting the target variable. It is essential to consider these outliers in further analysis and model building to ensure accurate predictions.

4. To examine if there are any associations among the variables in the dataset, I used two primary approaches: correlation analysis for numerical variables and chi-square tests for categorical variables.

The results of the correlation analysis revealed several significant associations. The employment variation rate is highly correlated with the Euribor 3-month rate with a correlation coefficient of 0.97, suggesting that as the employment variation rate increases, the Euribor rate also tends to increase. The employment variation rate is also highly correlated with the number of employees with a correlation coefficient of 0.90, indicating that changes in employment rates are closely related to the number of employees. There is a negative correlation between the number of days since the client was last contacted and the number of previous contacts with a correlation coefficient of -0.59, indicating that clients who were contacted more frequently in the past are less likely to have a high value. Additionally, the Euribor rate is strongly correlated with the number of employees, with a correlation coefficient of 0.94, suggesting a relationship between market interest rates and employment levels.

In conclusion, the correlation analysis revealed strong associations, particularly between economic indicators and employment levels. The chi-square tests showed significant associations between the target variable and various categorical variables, including job type, marital status, education, and financial factors such as default, housing loan, and personal loan. These associations provide valuable insights into the factors that influence the likelihood of clients subscribing to a term deposit and can inform the development of predictive models.

5. To analyze the quantitative variables using a clustering technique, I used K-Means clustering. The results of the clustering analysis provided distinct groupings. Cluster 0 includes younger clients with fewer contact attempts and higher pdays values, indicating that they were not contacted previously. Call durations in this cluster are relatively shorter, and there are few previous contacts. Economic indicators for this cluster generally show moderate values. Cluster 1 consists of older clients with longer call durations, suggesting more engaged conversations. This cluster has more contact attempts during the campaign, lower pdays values indicating recent contacts, and more previous contacts. The economic indicators for this cluster generally show higher values, indicating a better economic context. Cluster 2 represents a moderate group with clients of a moderate age range. The economic indicators for this cluster show mixed values, reflecting variability in the economic context.

The K-Means clustering analysis revealed distinct groups within the dataset based on quantitative variables. These insights can help tailor marketing strategies to different client groups, improving the effectiveness of future campaigns. Cluster 0's younger clients may

require different engagement strategies compared to the older, more engaged clients in Cluster 1, while the mixed characteristics of Cluster 2 suggest a need for a varied approach.

6. To build a model that predicts whether a client will subscribe to a term deposit, I used the Logistic Regression classification technique. It provides probabilities for class membership, which is useful for understanding the likelihood of a client subscribing.

The Logistic Regression model performed well, achieving an overall accuracy of 90%. While there is room for improvement, especially in predicting the minority class (clients who subscribed), the model provides a solid foundation for understanding the factors that influence a client's decision to subscribe to a term deposit. Further improvements could include using more advanced classification techniques, feature engineering, and addressing class imbalance to enhance the model's performance.

7. **Model Evaluation**

The Logistic Regression model used to predict whether a client will subscribe to a term deposit achieved an overall accuracy of 90%. However, accuracy alone does not provide a complete picture of the model's performance, especially when dealing with imbalanced datasets.

Classification Report

The classification report provides a detailed breakdown of the model's performance for each class:

- **Precision**:
  - Class 0 (did not subscribe): 0.94
  - Class 1 (subscribed): 0.55
- **Recall**:
  - Class 0: 0.96
  - Class 1: 0.46
- **F1-Score**:
  - Class 0: 0.95
  - Class 1: 0.50
- **Support**:
  - Class 0: 1105 instances
  - Class 1: 131 instances

The precision for class 0 is high, indicating that when the model predicts a client will not subscribe, it is correct 94% of the time. However, the precision for class 1 is lower at 55%, meaning that when the model predicts a client will subscribe, it is correct only 55% of the time. The recall for class 0 is 96%, indicating that the model successfully identifies 96% of clients who do not subscribe. For class 1, the recall is 46%, indicating that the model correctly identifies only

46% of clients who subscribe. The F1-score, which balances precision and recall, is 0.95 for class 0 and 0.50 for class 1.

Model Fit and Improvement

The model fits well overall, particularly in predicting clients who do not subscribe. However, its performance in predicting clients who do subscribe is less satisfactory, as indicated by the lower precision, recall, and F1-score for class 1. This is likely due to the class imbalance in the dataset, where the number of clients who did not subscribe far exceeds the number of clients who did.

Improving the Model

Several approaches can be taken to improve the model:

1. **Address Class Imbalance**:
    - **Oversampling**: Increase the number of instances in the minority class (clients who subscribed) using techniques such as SMOTE (Synthetic Minority Over-sampling Technique).
    - **Undersampling**: Reduce the number of instances in the majority class (clients who did not subscribe).
    - **Class Weights**: Adjust the class weights in the Logistic Regression model to give more importance to the minority class.
2. **Feature Engineering**:
    - Create new features or transform existing ones to better capture the underlying patterns in the data. For example, interaction terms between features or non-linear transformations.
3. **Advanced Classification Techniques**:
    - Use more advanced algorithms such as Random Forest, Gradient Boosting, or Neural Networks, which might capture complex patterns better than Logistic Regression.
4. **Cross-Validation**:
    - Use cross-validation to ensure the model's robustness and generalizability to unseen data.

Conclusion

The Logistic Regression model provides a solid foundation for predicting whether a client will subscribe to a term deposit, achieving an overall accuracy of 90%. However, the model's performance in predicting the minority class (clients who subscribed) can be improved. By addressing class imbalance, performing feature engineering, exploring advanced classification techniques, and tuning parameters, the model's predictive power can be enhanced, leading to more accurate and reliable predictions.

References:

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms* (2nd ed.). Wiley.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. Retrieved from https://www.springer.com/gp/book/9781461471370

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. doi:10.1109/TKDE.2008.239. Retrieved from https://ieeexplore.ieee.org/document/4678653

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. Retrieved from https://www.springer.com/gp/book/9780387848570