Week 11: EDA Presentation and proposed modeling technique

Project: Bank Marketing (Campaign)

Data Glacier Virtual Internship

July 16 2024

Group Name: BRAVO

Name	Email (registered with Data Glacier)	Country	College/Company	Specialization
Jackson Taylor	jacksonian.r.taylor@gmail.com	United States	Santa Clara University	Data Science
Balamurugan Purushothaman	balamurugan2001viruda@gmail.com	United Kingdom	University of Liverpool	Data Science
Nazrin Thanikattil Rafeeque	101nazrin@gmail.com	United Kingdom	University of Hertfordshire	Data Science
Gunjan Varyani	gunjanvaryani916@gmail.com	United States	University of the Cumberlands	Data Science

Problem description:

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which helps them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

The steps to solving this task include outlining the project, the initial data understanding and strategies to solve data problems, data cleansing and transformation, exploratory data analysis code, exploratory data analysis presentation and model recommendation, model selection and building, and presenting the final solution and code.

This presentation is the exploratory data analysis presentation and model recommendation stage.

Content Summary

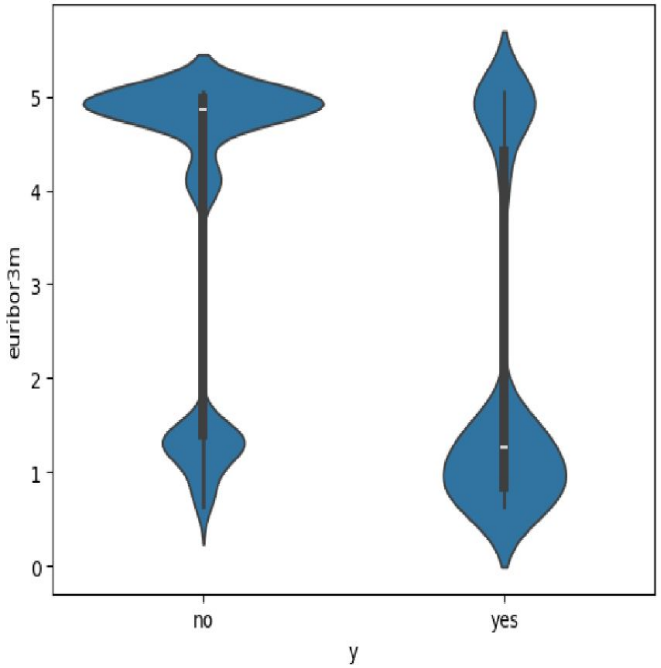
EDA insight slides:

These slide are dedicated to explaining visualizations and trends and transforming them into actionable insights. This is to help ABC bank make effective decisions to maximise their focus on users who are more likely to subscribe to the term deposit product.

Model Recommendation slides:

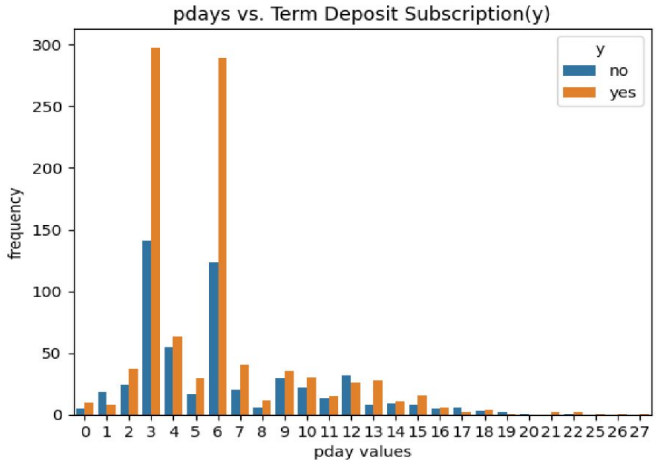
The final model recommendations are the models we recommend at this stage which can determine which users will buy the product in an automated and robust way.

Part 1: euribor3m vs. Term Deposit Subscription(y)



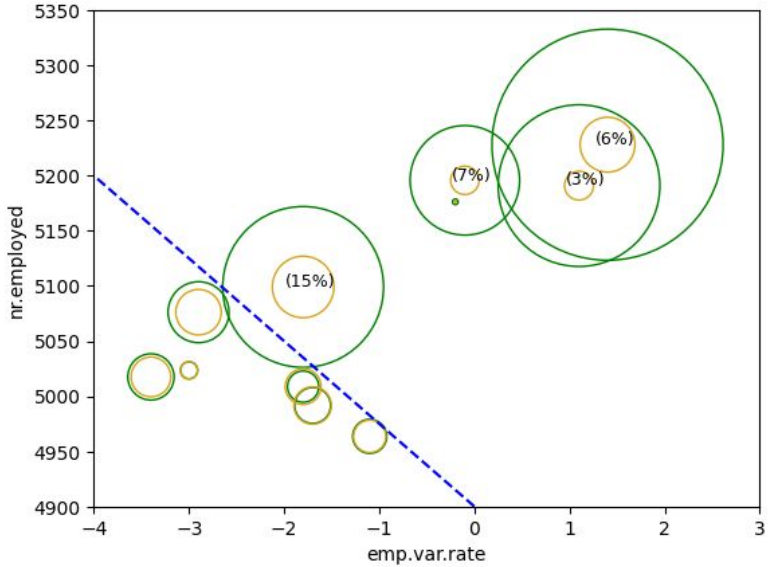
- This is the violin plot between numerical variable euribor3m (Euro Interbank Offered Rate for a three-month maturity with a daily indicator) vs y (Term Deposit Subscription).
- This graph shows that lower values of euribor3m are associated with increased subscriptions and subscription rates.
- ABC bank should make sure they are prepared to launch more aggressive campaigns when they see a downward trend in euribor3m.

Part 2: pdays vs. Term Deposit Subscription(y)



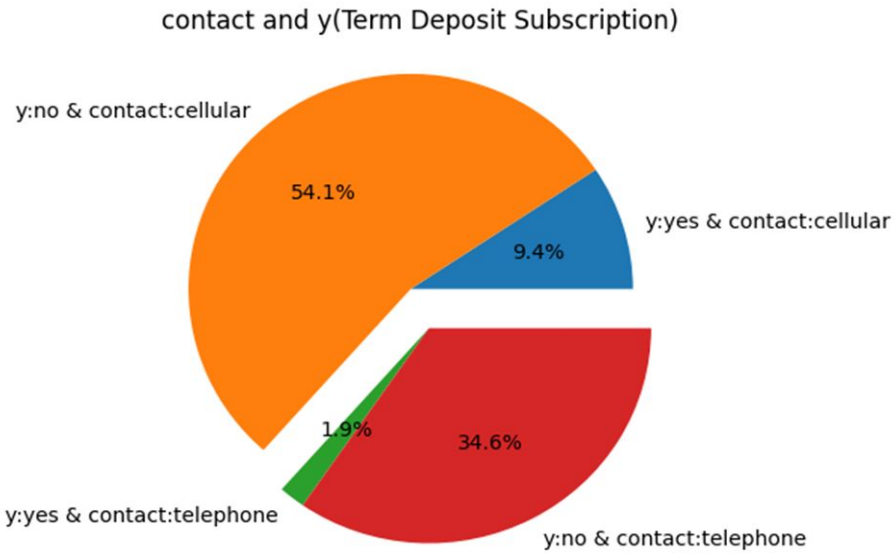
- Observing when pdays is not equal to 999 reveals critical information about the sub-distribution
- At pday values 13, 6, and 3, these levels have the highest ratio of subscribers to all clients.
- It seems likely that the high purchase percentage for pday value of 13 is due to variance and small sample size.
- ABC Bank may have a follow-up meetings on pdays 6 and 3.
- The recommendation to ABC Bank is to continue focusing on those levels (3 and 6) to contact clients. With additional data it should also be confirmed if the high purchasing success in pday = 13 is an anomaly or not. Likewise, the pdays with small samples should be re-sampled.

Part 3: emp.var.rate vs. nr.employed



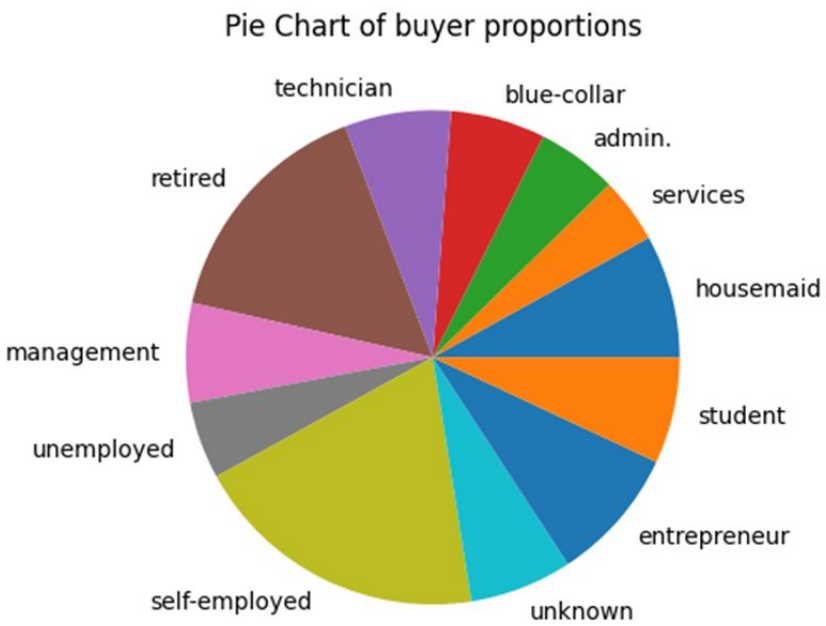
- This is a scatter plot of emp.var.rate and nr.employed where each point is the center of a pair of bubbles.
- For each point, the size of the green bubbles represents the number of clients who did not buy the product and the size of the gold bubbles represents the number of clients who bought the product.
- Some points have percentages which indicate the percentage of how much the gold bubble is of the green bubble: (clients who bought/clients who didn't)*100
- The graph displays a discernable boundary between combinations of emp.var.rate and nr.employed. On one side, all but one of the points have a low purchase percentage and a large number of observations. On the other side, the points have more equal percentage of subscribers to non-subscribers and less observations.
- The intuition from this visualization is when emp.var.rate and nr.employed are low, the purchase percentage is higher. It may be a good idea to maximize campaign efforts under these circumstances.

Part 4: Pie chart of contact and y (Term Deposit Subscription)



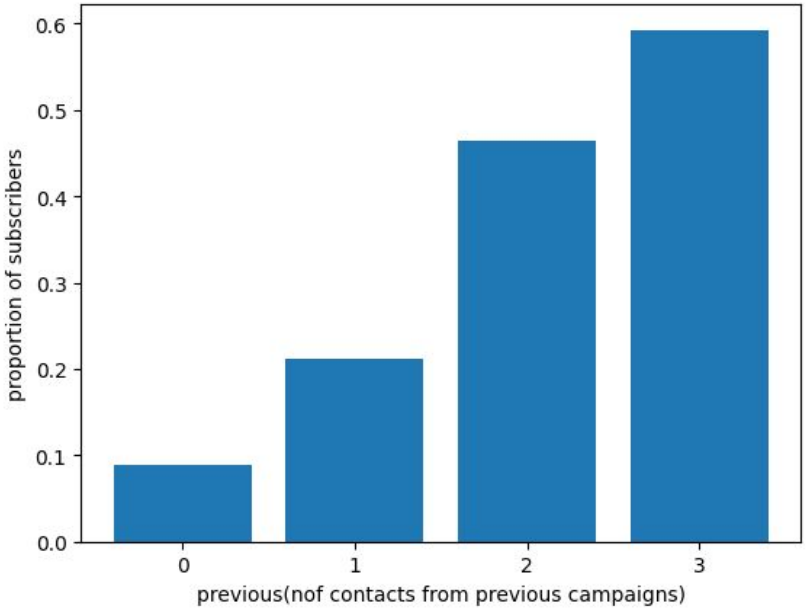
- The pie chart shows that it is clearly better to market to those with cellphones because they have a stronger rate of purchasing the product.
- This could be the result of the accessibility that a phone offers. It makes it more convenient for clients to engage in banking services.
- Targeting those with cellphones can be done through selective advertisement, especially if ABC bank has a mobile app that clients need to use to make transactions, check balance etc.

Part 5: Pie Chart of Buyer Proportions:



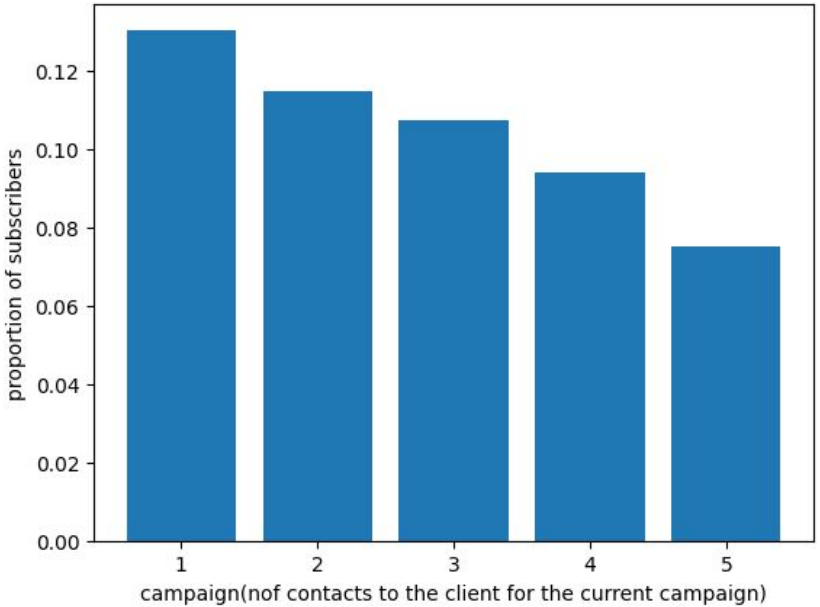
- The value of each slice of the pie is clients who subscribed in a job category divided by total clients in that job category.
- Putting this to use in a business scenario would be trying to target the groups that represent substantial pieces of the pie.
- The standout sections appear to be clients who are retired and clients who are self-employed.
- Marketing to certain jobs can be an easier task than many of the marketing tasks that could be construed from this dataset. Data is everywhere which could contain people's job title or heavily allude to it.
- Once a population with jobs that are part of the key titles has been identified, it is time to target them. Targeted advertising can happen on google, job sites, and social media. There are also strategies like email marketing and propositions.

Part 6: Proportions of subscribers vs. previous



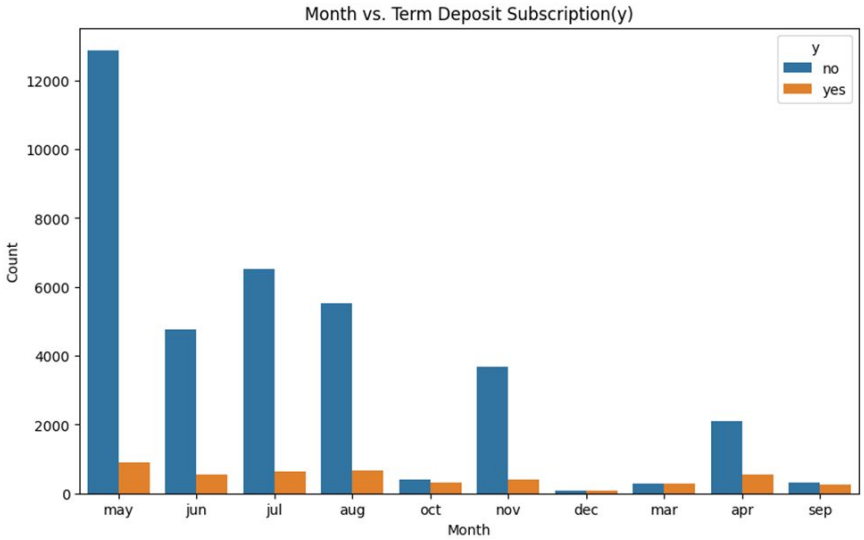
- This is a bar plot where each bar is the proportion of subscribers for the corresponding number of contacts in previous campaigns.
- The more contacts performed in previous campaigns for this client the more likely they will buy the product. Contacts in previous campaigns often indicate that they have previously bought the product which means they are more likely to repeat their decision of buying again in the current campaign.
- ABC bank should make sure to focus on clients who have been contacted more during previous campaigns.

Part 7: Proportions of subscribers vs. campaign



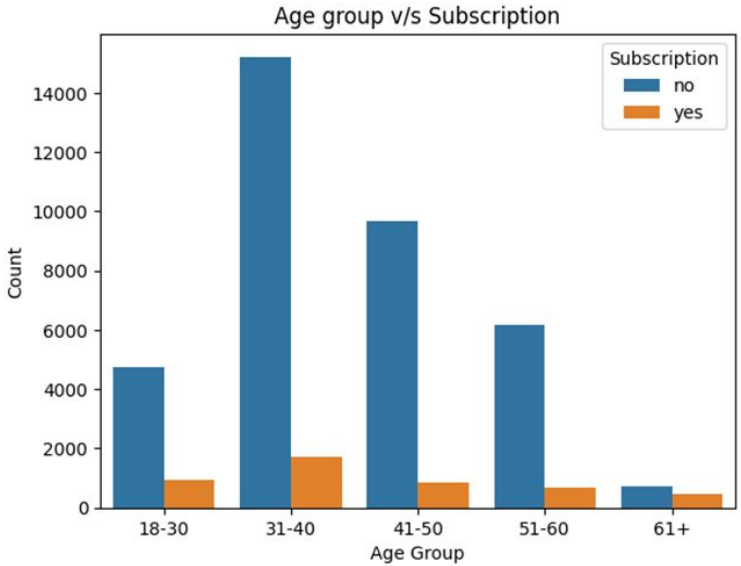
- This is a bar plot where each bar is the proportion of subscribers for the corresponding number of contacts in the current campaign.
- The first contact leads to the most success compared to any other contact level. The second contact and thereafter have less individual success because the client is less likely to subscribe in a later contact when they already know about the product from a previous contact.
- On the other hand, it is possible that the client did not make a decision yet but has expressed interest requiring further conversation.
- Calling again and again still leads to more customers buying the product, but we recommend focusing on contacting those who have not been contacted for the current campaign and those that have expressed interest that require more than one contact.

Part 8: Month vs. Term Deposit Subscription(y)



- The month of May has the highest number of contacts but lowest subscription rate.
- Reconsider the campaign timing.
- Try to shift the campaign to June, July, August and other months which have higher purchase rates.
- Analyze other months for the Campaign.

Part 9: Age vs. Term Deposit Subscription(y)



Primary Insights:

- **Highest Subscribers:** Ages 31-40
- **Lowest Subscribers:** Ages 61+

Statistical Findings:

Chi-square Test:

- **Value:** 664.315
- **Significance:** $p < 0.05$
- Indicates subscription rates vary significantly between age groups.

Part 9 (continued)

Detailed Analysis:

Positive Residuals:

- **Ages 18-30 and 61+** Higher-than-expected subscription rate

Neutral Residuals:

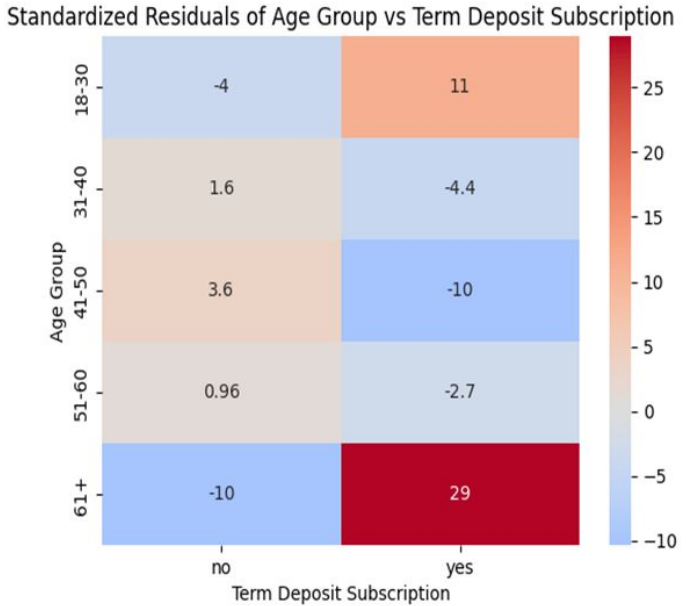
- **Ages 31-40 and 51-60:** Subscription rate similar to expectations

Negative Residuals:

- **Ages 41-50:** Lower-than-expected subscription rate

Strategic Recommendations:

1. **Continue** existing marketing for ages 31-40 and 51-60.
2. **Focus** on financial awareness and inviting offerings for the ages 18-30
3. **Encourage** safety and stability among those aged 61.
4. **Investigate** and address barriers for ages 41-50.



Recommended Models

Random Forest (bagging ensemble method):

- Can handle class imbalance by adjusting class weights or using balanced class sampling. The target variable for our dataset (y) is highly imbalanced.
- Provides feature importance. In our EDA, there are a handful of features that seem to be less important. This was found with the tests done of many categorical variables. The random forest can allow us to remove features post build based on feature importance. If a feature is just producing noise, then the model can be rebuilt with those features removed.
- Less prone to overfitting and noise due to the averaging of multiple trees. There is a possibility of overfitting when a certain number of iterations are run for other classification methods.
- Can handle nonlinear data without need for feature transformation. This is critical because there are only a handful of linear relationships in our data. Random forests can accommodate that.
- Can work well with datasets that have a mix of numerical and categorical features with less preprocessing like scaling or one hot encoding. Our dataset is mixed with numeric and categorical features.

SVM (Support Vector Machine) (discriminative classifier):

- Can work well with both linear and nonlinear relationships. This is good because there are only a handful of linear relationships in our data. SVM can use different kernel functions to make complex decision boundaries.
- Works well with high dimensional space. Data is high dimensional when there is high value to combining many features rather than just concerning the individual value of features. This is the case with our dataset.
- Robust to outliers. Our data has many outliers, and we can test with the removal or without the removal of those outliers with SVM. SVM is resilient to this because it has regularization parameters.
- Less overfitting. Our dataset seems to have a solid number of irrelevant or less relevant features that could produce a lot of noise with other models. SVM is resilient to this because it has regularization parameters.
- Most effective in small to medium sized datasets. Our dataset and number of features are considered medium.
- Well suited for binary classification tasks. That is our model type. This should improve performance.

GBM (Gradient Boosting Machine) (ensemble model):

- Our dataset has significant class imbalance within the target variable. This can be addressed by adjusting learning rates, using class weights, and employing boosting algorithms like AdaBoost or XGBoost.
- GBM calculates feature significance scores, which aid in identifying essential predictors and reducing noise from less significant information. This could improve model accuracy.
- This algorithm works by building sequential trees, where each tree corrects errors of the previous ones. This enables it to capture complex nonlinear relationships present in our dataset.
- The parameter tuning (nof trees, learning rate, depth) help to deal with different datatypes and structures more effectively. Our dataset has a combination of both numerical and categorical values. Hence, it would be ideal.
- However, careful parameter tuning is required to achieve a computationally inexpensive model.

Logistic Regression (linear model)

- LR is a more clear and straightforward algorithm. The coefficients of this model that illustrate the impact of each feature on the probability of subscribing to term deposit, aid in understanding and transparency.
- LR is inherently designed for binary outcomes due to the logistic function's capability to model probabilities between two classes. This feature matches well with our requirements so it could be an appropriate choice.
- LR is computationally efficient, ideal for managing and processing the large number of records in the dataset, enabling quick iterations.
- LR can handle non-linearity in the predictor variables effect on the independent variables with the addition of interaction terms and applying transformations.
- LR is more robust to outliers than linear regression. This means tests can be done with and without the removal of outliers.

Note: None of these models assume normality of the predictor variables which is a requirement since none of the predictor variables in our dataset are normal.

Thank You