75

ADVANCE REGRESSION ANALYSIS 6301

PREDICTING TERM DEPOSIT SUBSCRIPTIONS

GROUP-6

CHARITHA NALLAKA

GEETHIKA SANNALA

HARSHITHA ATLURI

DENNIS MC LEAORD ANNAVARAPU

1. INTRODUCTION

STATEMENT OF PROBLEM:

Banks continually seek innovative methods to engage potential customers and promote their products. A Portuguese bank conducted a phone-based marketing campaign to promote term deposits. This analysis aims to identify key factors influencing the success of the campaign and predict customer subscription to term deposits, using a dataset that includes customer demographics, socio-economic status, and details of marketing interactions.

KEY RESEARCH QUESTIONS:

- **Customer Characteristics:** What key demographic and socio-economic traits define customers who subscribe to the bank's term deposit products?
- **Campaign Effectiveness:** What factors influence the success of the marketing campaign, and how can resource allocation be optimized?
- **Time-Related Analysis:** Are there specific times of the year, month, or day with higher success rates for term deposit subscriptions?

2. DATA DESCRIPTION:

The dataset includes 45,153 rows and 17 columns, capturing various attributes related to customer demographics, socio-economic status, and marketing interactions. The categorical variables are whether the customer subscribed to the term deposit (y), job type, marital status, education level, credit default status, housing loan status, personal loan status, type of communication contact, last contact month, and outcome of the previous campaign. The numerical variables are age, average yearly balance in euros, last contact day, duration of the last contact in seconds, number of contacts during this campaign, days since last contact from a previous campaign, and number of contacts before this campaign.

3. **DATA PREPROCESSING**: Our goal in preprocessing was to prepare the dataset for analysis and modelling, ensuring consistency and accuracy. Although there were no missing values or duplicates, we addressed issues such as the presence of unknown values.

3.1 Dropped Columns:

- **poutcome:** Dropped due to the high proportion of "unknown" values, making it unreliable.
- **previous:** Dropped as it was redundant, similar to the campaign variable.

3.2 Imputation Strategy:

- Replaced unknowns with the mode within specific groups in job, education and contact to preserve data integrity and contextual relevance.

3.3 Handling Outliers:

- **Age and Balance:** Retained outliers to maintain the integrity of the data.
- **pdays:** Left unchanged to preserve its relationship with the target variable.
- **Campaign and Duration:** Capped outliers to prevent extreme values from distorting the model.

3.4 Handling Imbalance data: From Fig. 1 The target variable, y, showed class imbalance. To address class imbalance in the target variable, we used random oversampling, duplicating instances of the minority class. This technique ensures the model learns from both classes effectively, improving metrics and reducing bias for more balanced, reliable predictions.

4 **EXPLARATORY DATA ANALYSIS**

4.1 Impact of demographic factors on subscription status: The analysis (Fig. 4) shows that the majority of subscribers are between 25 and 50 years old and there are more subscriptions success in old age people, suggesting the service is more popular among younger individuals and old people are least likely to reject. As seen in Fig. 3, marital status influences subscription behaviour, with married individuals having the highest number of subscriptions, followed by single individuals. Fig. 2 indicates that individuals with secondary and tertiary education levels are more likely to subscribe, implying that education may be a factor in subscription decisions.

4.2 **Impact of socio-economic factors:** The analysis (Fig. 5) shows that individuals who did not default on the service are more likely to subscribe, suggesting that defaulting may affect subscription behavior. As shown in Fig. 6, individuals without loans are more likely to subscribe, indicating that financial stability could play a role. Fig. 7reveals that individuals in the "management" job category have the highest subscription rates, with those in "retired" and "admin." categories also subscribed at relatively high rates. Fig. 8, related to housing loans, suggests that individuals with housing loans are less likely to subscribe, possibly due to financial constraints.

4.3 **Impact of contact methods on outcome**: The fig 9 shows the cellular contact is the most effective method, reaching more individuals and securing more subscribers. Telephone contact is less effective, and the "unknown" method is the least effective, highlighting the need for improved targeting and contact data.

4.4 **Identifying key periods for high subscription success:** The (Fig. 12) shows that the subscription success rate fluctuates throughout the month, with peaks on day 1, day 10, and day 30, suggesting that factors like paydays or specific marketing strategies might influence success. Fig. 11 reveals that December and March have the highest success rates, while July and May have the lowest, indicating that seasonal factors or targeted campaigns may play a role. Fig. 10 shows that longer contact durations result in higher success rates, with "Very Long" contacts being the most successful, suggesting that more time spent with potential customers improves conversion.

5 **CATEGORICAL ENCODING:** We used one-hot encoding to preprocess categorical variables, converting them into binary matrices to avoid assuming any ordinal relationships. This method improves model performance by allowing algorithms to interpret the data accurately.

It also facilitates interpretation by representing the presence or absence of categories, simplifying analysis.

6 MODEL BUILDING

6.1 Model 1: Baseline Model - Generalized Linear Model (GLM) with All Features: We started by building a baseline model using a Generalized Linear Model (GLM) with all available features, which served as a reference point to evaluate more complex models and understand the initial predictive power of the data.

6.2 Model 2: Backward Elimination with Stepwise AIC: We refined the model using backward elimination with stepwise AIC, which iteratively removes the least significant features based on AIC to improve model performance and interpretability.

6.3 Model 3: Complex Model with Interaction Terms: To enhance the model, we incorporated interaction terms to capture the combined effects of multiple variables on the response. Based on domain knowledge and exploratory analysis, we identified interactions where one variable's effect depends on another. Interaction terms included were **Age and Job** (age * job), **Marital Status and Education** (marital * education), **Housing Loan and Personal Loan** (housing * loan), and **Contact Method and Month** (contact * month), to account for how these factors influence campaign responses differently depending on their combinations.

7 MODEL COMPARISON: From table 1 Model 1, our baseline, achieved an accuracy of 81.33%, with similar metrics across precision, recall, and F1-score. However, its AIC and BIC values suggested room for improvement. Model 2, with backward elimination, showed identical performance but slightly better fit. Model 3, which included interaction terms, outperformed both, with the highest accuracy (81.69%), precision (81.51%), recall (81.99%), and lower AIC and BIC, making it the most reliable and accurate model.

8 GOODNESS OF FIT:

8.1 Deviance Test: The deviance test shows that the interaction model is significantly better than the model 2, with a 571.8 deviance reduction and a highly significant p-value (< 2.2e-16).

This suggests that adding interaction terms improves the model's fit and its ability to explain the relationship between predictors and the outcome.

8.2 Likelihood Ratio Test: From fig.22 the Likelihood Ratio Test shows that the interaction model fits significantly better than the baseline model, with a large Chi-square statistic and highly significant p-value, indicating that interaction terms improve the model's explanatory power.

8.3 AIC and BIC Values: Comparing the AIC and BIC values, Model 1 has AIC of 67596 and BIC of 67939.27, Model 2 has a slightly better AIC of 67595 and BIC of 67929.12, while Model 3 has the lowest AIC of 67075 and BIC of 67650.81. These results indicate that Model 3 provides the best balance between fit and complexity, making it the most optimal model.

9 KEY FINDINGS FROM ODDS RATIO: Key good predictors include months like March, May, September, and December, call duration, job-age interactions, marital status with education, and certain contact methods. Bad predictors are self-employed (0.56), those with housing or loans (0.35, 0.41), retirees (0.01), housemaids, and telephone contact in February and January. Age (0.997), previous attempts (1.14), and pdays (1.00) are neutral predictors with minimal impact.

10 CONCLUSIONS: The analysis identifies key factors influencing term deposit subscriptions, including age, marital status, education, job type, and financial stability. Cellular contact and longer interactions are more effective, with higher success rates in December and March. The best-performing model, with an accuracy of 81.69%, incorporated interaction terms to improve predictive power. Recommendations include targeting younger, educated, married individuals without loans, focusing on cellular communication, training agents for engaging calls, and optimizing campaigns during peak months. Regular model updates are essential for adapting to customer behavior and market trends, ensuring continuous improvement in engagement and conversion.

11 DISCUSSIONS

Implications for the Population:

The analysis shows key factors like age, marital status, education, and job type influence subscription rates, suggesting targeted marketing should focus on individuals around 30 years old, married, educated, and in management or technical roles. Marketing should peak in March, May, September, and December.

Improvements for Future Studies:

Expanding the sample to include a broader range of demographics (e.g., retirees, self-employed) and regions would improve generalizability. A larger dataset and adding geographic location could further refine predictions.

Potential Biases:

Sample selection bias may exist due to underrepresentation of certain groups. Temporal factors could also bias the effectiveness of marketing campaigns.

Pros and Cons:

Pros: High accuracy and precision, better insights from interaction terms.

Cons: Relies on historical data, possible overfitting with a small sample.

Improvements with More Time:

Cross-validation and advanced algorithms like ensemble methods could improve model generalizability and predictive power.

Other Applications:

This approach could be used in other marketing areas (e.g., loan offers) or industries such as retail and telecommunications, focusing on customer segmentation.

In conclusion, refining the model with broader data and further validation would improve accuracy and applicability.

12 REFERENCES

https://archive.ics.uci.edu/dataset/222/bank+marketing

APPENDIX

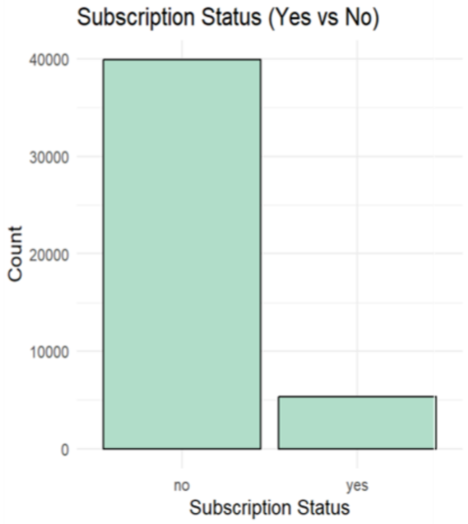

Figure 1IMBALANCE DATA

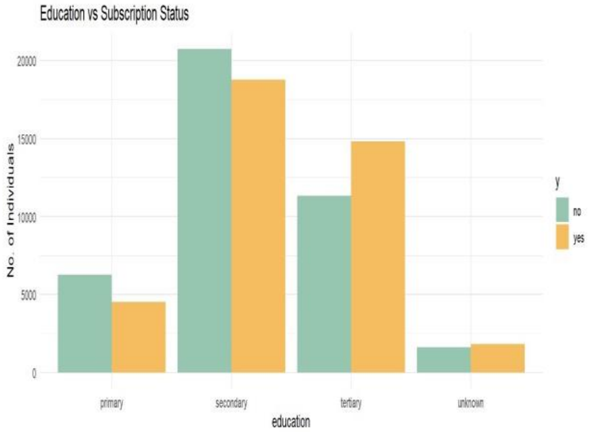

Figure 2 EDUCATION VS SUBSCRIPTION STATUS

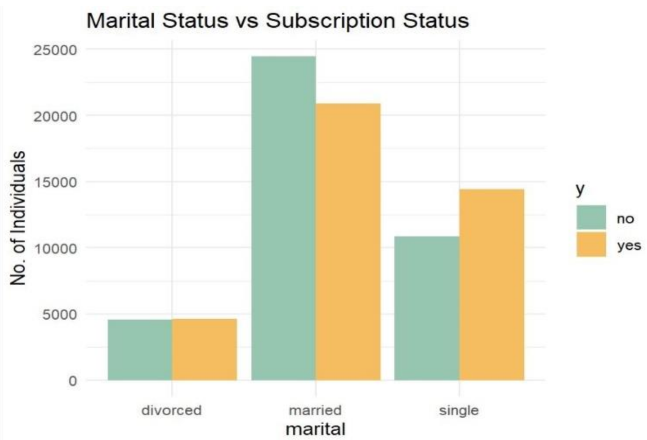

Figure 3 Marital Status vs Subscription Status

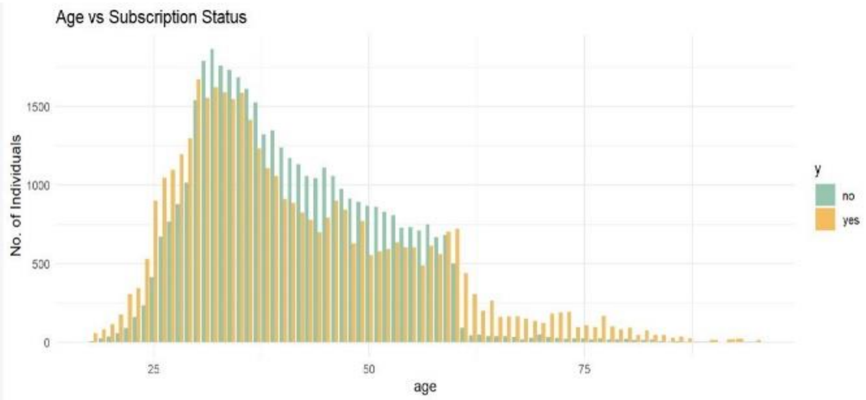

Figure 4 Age vs Subscription Status

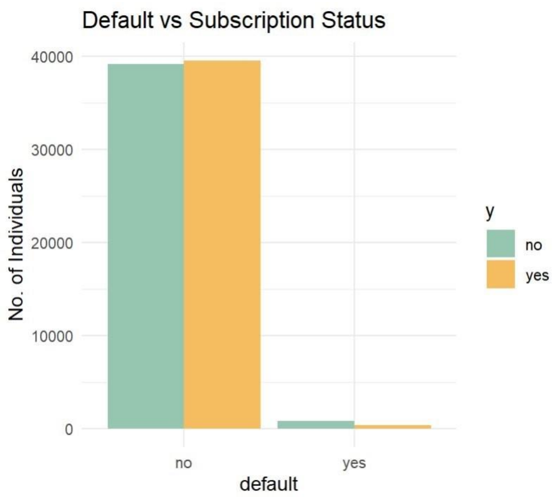

Figure 5 Default vs Subscription Status

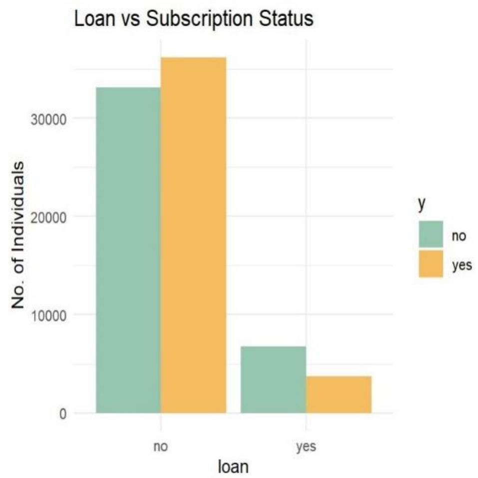

Figure 6 Loan vs Subscription Status

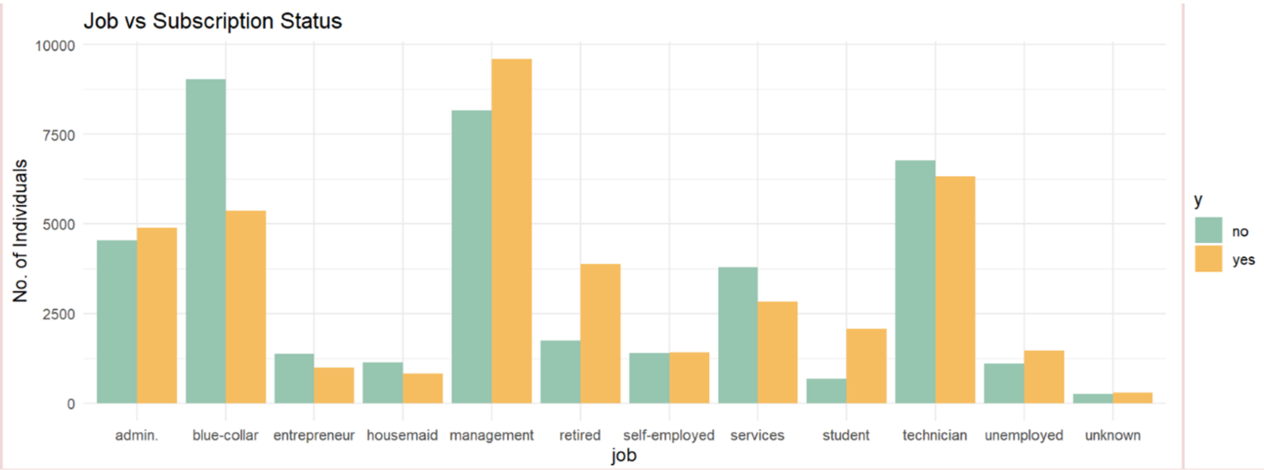

Figure 7 Job vs Subscription Status

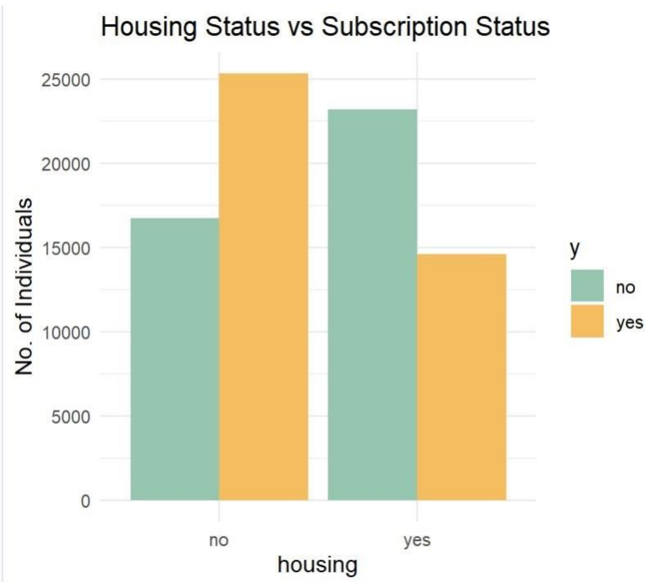

Figure 8 Housing vs Subscription Status

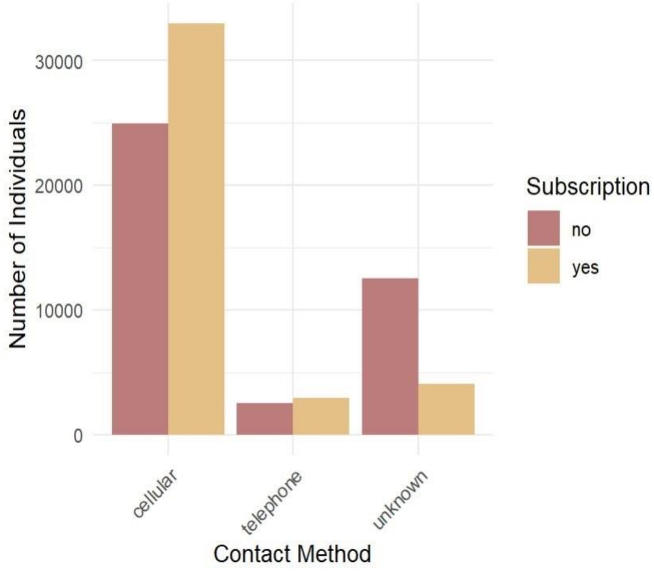

Figure 9 Contact Method vs Subscription

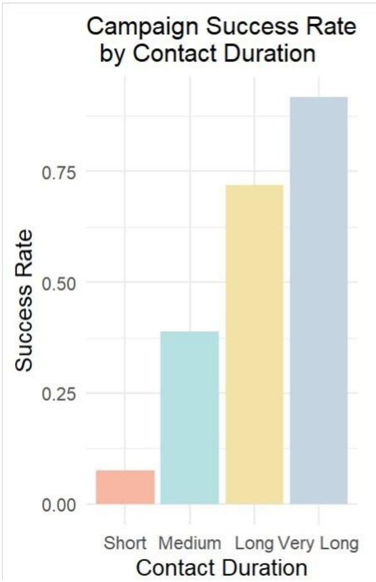

Figure 10 Contact Duration vs Subscription

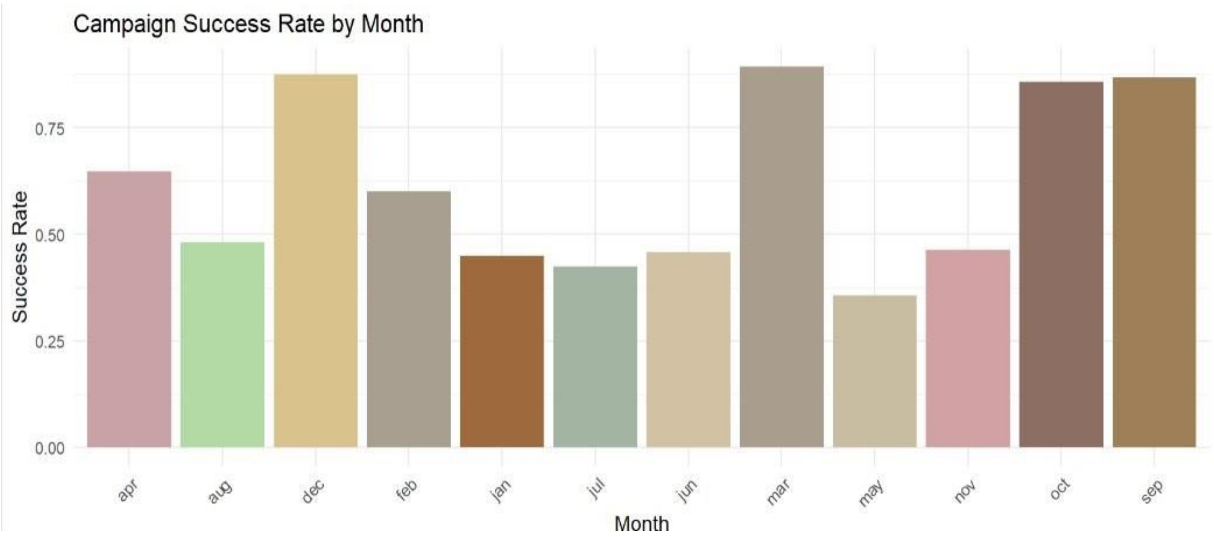

Figure11 month vs Subscription Status

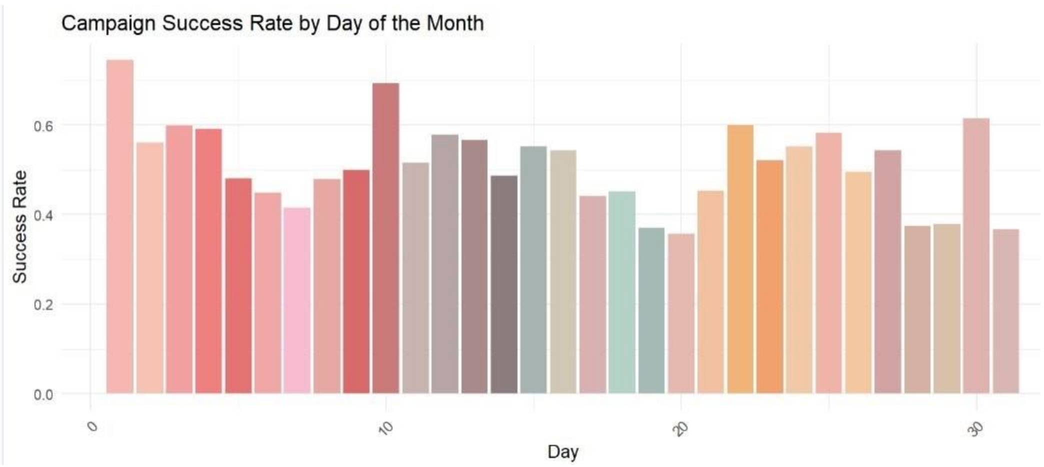

Figure12 month vs Subscription Status

```
> # Print the summary of the initial model
> summary(initial_model)

Call:
glm(formula = y ~ ., family = binomial, data = data)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -1.441e+00  8.547e-02 -16.856  < 2e-16 ***
age                 1.250e-03  1.150e-03   1.087 0.277079    
jobblue-collar     -5.245e-01  3.733e-02 -14.052  < 2e-16 ***
jobentrepreneur    -5.925e-01  6.459e-02  -9.173  < 2e-16 ***
jobhousemaid       -5.769e-01  6.973e-02  -8.273  < 2e-16 ***
jobmanagement      -3.309e-01  3.935e-02  -8.410  < 2e-16 ***
jobretired          2.244e-01  5.387e-02   4.166 3.10e-05 ***
jobself-employed   -5.447e-01  6.001e-02  -9.077  < 2e-16 ***
jobservices        -3.910e-01  4.343e-02  -9.003  < 2e-16 ***
jobstudent          5.485e-01  6.362e-02   8.622  < 2e-16 ***
jobtechnician      -3.426e-01  3.677e-02  -9.317  < 2e-16 ***
jobunemployed      -2.265e-01  6.124e-02  -3.699 0.000216 ***
maritalmarried     -8.890e-02  3.142e-02  -2.830 0.004658 ** 
maritalsingle       2.452e-01  3.607e-02   6.800 1.05e-11 ***
educationsecondary  2.883e-01  3.284e-02   8.778  < 2e-16 ***
educationtertiary   6.475e-01  3.966e-02  16.329  < 2e-16 ***
defaultyes         -3.618e-01  8.218e-02  -4.403 1.07e-05 ***
balance             2.425e-05  3.185e-06   7.613 2.69e-14 ***
housingyes         -9.681e-01  2.273e-02 -42.586  < 2e-16 ***
loanyes            -5.736e-01  3.013e-02 -19.037  < 2e-16 ***
contacttelephone    9.068e-02  4.095e-02   2.214 0.026813 *  
day                -3.420e-03  1.277e-03  -2.677 0.007418 ** 
monthaug           -7.918e-01  4.193e-02 -18.884  < 2e-16 ***
monthdec            7.729e-01  1.164e-01   6.638 3.18e-11 ***
monthfeb           -1.486e-01  4.834e-02  -3.075 0.002108 ** 
monthjan           -1.151e+00  6.258e-02 -18.389  < 2e-16 ***
monthjul           -1.059e+00  4.226e-02 -25.056  < 2e-16 ***
monthjun           -7.147e-01  4.400e-02 -16.241  < 2e-16 ***
monthmar            1.854e+00  8.213e-02  22.574  < 2e-16 ***
monthmay           -1.322e+00  3.852e-02 -34.312  < 2e-16 ***
monthnov           -8.341e-01  4.529e-02 -18.416  < 2e-16 ***
monthoct            1.242e+00  6.795e-02  18.277  < 2e-16 ***
monthsep            1.117e+00  7.537e-02  14.822  < 2e-16 ***
duration            7.939e-03  5.862e-05 135.421  < 2e-16 ***
campaign           -1.070e-01  5.010e-03 -21.356  < 2e-16 ***
pdays               2.381e-03  1.090e-04  21.849  < 2e-16 ***
previous            1.306e-01  5.984e-03  21.824  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 110687  on 79843  degrees of freedom
Residual deviance:  67522  on 79807  degrees of freedom
AIC: 67596

Number of Fisher Scoring iterations: 6
```

Figure13 model-1 summary

```
[1] "Confusion Matrix:"
         Actual
Predicted     0     1
        0 32406  7394
        1  7516 32528
[1] "Model Accuracy:  81.33 %"
[1] "Precision:  81.23 %"
[1] "Recall (Sensitivity):  81.48 %"
[1] "F1 Score:  81.35 %"
Setting levels: control = 0, case = 1
Setting direction: controls < cases
[1] "AUC:  0.889210525451237"
> #BIC value for initial model
> bic_inital <- BIC(initial_model)
> bic_inital
[1] 67939.27
```

Figure14 model-1 evaluation metrics

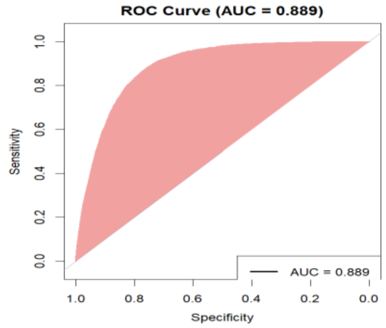

Figure15 model-1 ROC curve

```
> summary(final_model)

Call:
glm(formula = y ~ job + marital + education + default + balance + 
    housing + loan + contact + day + month + duration + campaign + 
    pdays + previous, family = binomial, data = data)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -1.381e+00  6.574e-02 -21.012  < 2e-16 ***
jobblue-collar     -5.252e-01  3.732e-02 -14.071  < 2e-16 ***
jobentrepreneur    -5.898e-01  6.454e-02  -9.138  < 2e-16 ***
jobhousemaid       -5.711e-01  6.950e-02  -8.218  < 2e-16 ***
jobmanagement      -3.287e-01  3.929e-02  -8.365  < 2e-16 ***
jobretired          2.494e-01  4.872e-02   5.119 3.07e-07 ***
jobself-employed   -5.435e-01  6.001e-02  -9.058  < 2e-16 ***
jobservices        -3.924e-01  4.342e-02  -9.037  < 2e-16 ***
jobstudent          5.371e-01  6.274e-02   8.560  < 2e-16 ***
jobtechnician      -3.420e-01  3.677e-02  -9.302  < 2e-16 ***
jobunemployed      -2.256e-01  6.123e-02  -3.685 0.000229 ***
maritalmarried     -9.137e-02  3.133e-02  -2.916 0.003546 ** 
maritalsingle       2.320e-01  3.396e-02   6.832 8.35e-12 ***
educationsecondary  2.849e-01  3.269e-02   8.715  < 2e-16 ***
educationtertiary   6.420e-01  3.932e-02  16.327  < 2e-16 ***
defaultyes         -3.625e-01  8.217e-02  -4.412 1.02e-05 ***
balance             2.449e-05  3.177e-06   7.708 1.28e-14 ***
housingyes         -9.710e-01  2.258e-02 -42.994  < 2e-16 ***
loanyes            -5.747e-01  3.011e-02 -19.082  < 2e-16 ***
contacttelephone    9.698e-02  4.054e-02   2.392 0.016739 *  
day                -3.426e-03  1.277e-03  -2.682 0.007312 ** 
monthaug           -7.907e-01  4.192e-02 -18.863  < 2e-16 ***
monthdec            7.729e-01  1.164e-01   6.639 3.17e-11 ***
monthfeb           -1.483e-01  4.834e-02  -3.069 0.002150 ** 
monthjan           -1.150e+00  6.257e-02 -18.379  < 2e-16 ***
monthjul           -1.060e+00  4.225e-02 -25.082  < 2e-16 ***
monthjun           -7.147e-01  4.400e-02 -16.241  < 2e-16 ***
monthmar            1.856e+00  8.213e-02  22.595  < 2e-16 ***
monthmay           -1.322e+00  3.851e-02 -34.333  < 2e-16 ***
monthnov           -8.332e-01  4.528e-02 -18.400  < 2e-16 ***
monthoct            1.244e+00  6.792e-02  18.314  < 2e-16 ***
monthsep            1.118e+00  7.535e-02  14.835  < 2e-16 ***
duration            7.938e-03  5.862e-05 135.422  < 2e-16 ***
campaign           -1.070e-01  5.011e-03 -21.358  < 2e-16 ***
pdays               2.383e-03  1.089e-04  21.879  < 2e-16 ***
previous            1.307e-01  5.984e-03  21.835  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 110687  on 79843  degrees of freedom
Residual deviance:  67523  on 79808  degrees of freedom
AIC: 67595

Number of Fisher Scoring iterations: 6
```

Figure16 model-2 Summary

```
+ }
[1] "Confusion Matrix:"
         Actual
Predicted     0     1
        0 32404  7378
        1  7518 32544
[1] "Model Accuracy:  81.34 %"
[1] "Precision:  81.23 %"
[1] "Recall (Sensitivity):  81.52 %"
[1] "F1 Score:  81.38 %"
Setting levels: control = 0, case = 1
Setting direction: controls < cases
[1] "AUC:  0.889199903440786"
>
>
> #BIC value for backward elimination model
> bic_final <- BIC(final_model)
> bic_final
[1] 67929.12
```

Figure17 model-2 evaluation metrics

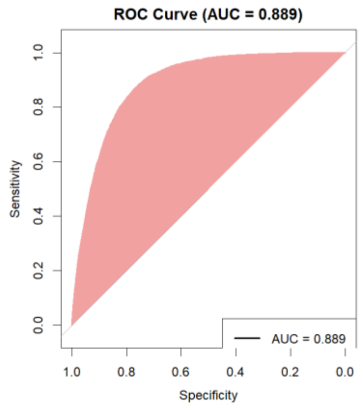

Figure18 model-2 ROC curve

```
> summary(interaction_model)

Call:
glm(formula = interaction_formula, family = binomial, data = data)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -1.029e+00  1.496e-01  -6.880 5.99e-12 ***
age                           -6.024e-04  2.940e-03  -0.205 0.837634    
jobblue-collar                -1.851e-01  1.509e-01  -1.226 0.220054    
jobentrepreneur               -8.475e-01  2.887e-01  -2.936 0.003326 ** 
jobhousemaid                  -7.892e-01  2.877e-01  -2.744 0.006079 ** 
jobmanagement                 -5.251e-01  1.461e-01  -3.595 0.000324 ***
jobretired                    -4.389e+00  3.048e-01 -14.400  < 2e-16 ***
jobself-employed              -5.909e-01  2.285e-01  -2.587 0.009693 ** 
jobservices                    4.727e-01  1.836e-01   2.575 0.010014 *  
jobstudent                     2.523e+00  3.136e-01   8.044 8.66e-16 ***
jobtechnician                 -3.246e-01  1.542e-01  -2.105 0.035332 *  
jobunemployed                 -1.645e-01  2.493e-01  -0.660 0.509388    
maritalmarried                -4.861e-01  8.209e-02  -5.921 3.19e-09 ***
maritalsingle                 -6.459e-02  1.073e-01  -0.602 0.547140    
educationsecondary            -5.323e-02  8.649e-02  -0.616 0.538222    
educationtertiary              2.448e-01  9.468e-02   2.586 0.009720 ** 
defaultyes                    -3.786e-01  8.267e-02  -4.579 4.66e-06 ***
housingyes                    -1.027e+00  2.421e-02 -42.421  < 2e-16 ***
loanyes                       -8.148e-01  4.505e-02 -18.086  < 2e-16 ***
contacttelephone              -9.178e-03  1.316e-01  -0.070 0.944383    
monthaug                      -7.634e-01  4.346e-02 -17.566  < 2e-16 ***
monthdec                       8.553e-01  1.275e-01   6.710 1.94e-11 ***
monthfeb                      -7.655e-02  5.047e-02  -1.517 0.129288    
monthjan                      -1.104e+00  6.520e-02 -16.938  < 2e-16 ***
monthjul                      -1.015e+00  4.435e-02 -22.896  < 2e-16 ***
monthjun                      -6.773e-01  4.534e-02 -14.939  < 2e-16 ***
monthmar                       2.005e+00  8.748e-02  22.923  < 2e-16 ***
monthmay                      -1.301e+00  3.970e-02 -32.768  < 2e-16 ***
monthnov                      -8.258e-01  4.738e-02 -17.430  < 2e-16 ***
age:jobblue-collar                  -8.353e-03  3.742e-03  -2.232 0.025595 *  
age:jobentrepreneur                  6.403e-03  6.763e-03   0.947 0.343738    
age:jobhousemaid                     4.798e-03  6.148e-03   0.780 0.435123    
age:jobmanagement                    4.976e-03  3.576e-03   1.391 0.164119    
age:jobretired                       7.293e-02  5.280e-03  13.810  < 2e-16 ***
age:jobself-employed                 1.278e-03  5.614e-03   0.228 0.819951    
age:jobservices                     -2.283e-02  4.667e-03  -4.892 9.98e-07 ***
age:jobstudent                      -7.536e-02  1.126e-02  -6.694 2.17e-11 ***
age:jobtechnician                   -3.435e-04  3.836e-03  -0.090 0.928643    
age:jobunemployed                   -1.497e-03  5.970e-03  -0.251 0.801998    
maritalmarried:educationsecondary    4.503e-01  9.251e-02   4.867 1.13e-06 ***
maritalsingle:educationsecondary     2.695e-01  1.157e-01   2.330 0.019793 *  
maritalmarried:educationtertiary     4.923e-01  9.956e-02   4.944 7.65e-07 ***
maritalsingle:educationtertiary      4.180e-01  1.220e-01   3.426 0.000614 ***
housingyes:loanyes                   4.554e-01  5.987e-02   7.607 2.81e-14 ***
contacttelephone:monthaug            2.946e-01  1.792e-01   1.644 0.100203    
contacttelephone:monthdec           -6.240e-01  3.668e-01  -1.701 0.088880 .  
contacttelephone:monthfeb           -5.551e-01  1.798e-01  -3.087 0.002024 ** 
contacttelephone:monthjan           -9.344e-01  2.556e-01  -3.655 0.000257 ***
contacttelephone:monthjul            7.703e-02  1.580e-01   0.488 0.625838    
contacttelephone:monthjun            6.982e-01  2.242e-01   3.114 0.001846 ** 
contacttelephone:monthmar           -1.755e+00  2.764e-01  -6.348 2.18e-10 ***
contacttelephone:monthmay            6.026e-02  1.746e-01   0.345 0.730044    
contacttelephone:monthnov            3.925e-01  1.707e-01   2.300 0.021457 *  
contacttelephone:monthoct            4.424e-01  2.049e-01   2.160 0.030808 *  
contacttelephone:monthsep           -4.835e-01  2.597e-01  -1.862 0.062636 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 110687  on 79843  degrees of freedom
Residual deviance:  66951  on 79782  degrees of freedom
AIC: 67075

Number of Fisher Scoring iterations: 6
```

Figure19 model-3 Summary

```
+ }
[1] "Confusion Matrix:"
         Actual
Predicted     0     1
        0 32496  7191
        1  7426 32731
[1] "Model Accuracy:  81.69 %"
[1] "Precision:  81.51 %"
[1] "Recall (Sensitivity):  81.99 %"
[1] "F1 Score:  81.75 %"
Setting levels: control = 0, case = 1
Setting direction: controls < cases
[1] "AUC:  0.891202128881543"
>
>
> #BIC value for interaction model
> bic_complex <- BIC(interaction_model)
> bic_complex
[1] 67650.81
>
```

Figure 20 model-3 evaluation metrics

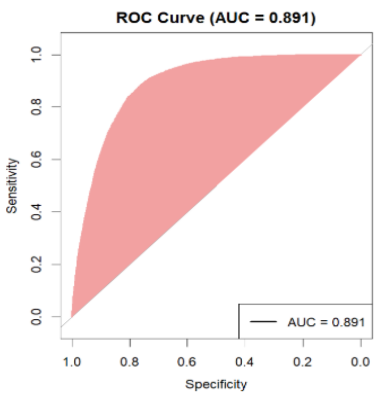

Figure 21 model-3 ROC curve

METRICS	MODEL 1	MODEL 2	MODEL 3
ACCURACY	81.33%	81.34%	81.69%
PRECISION	81.23%	81.23%	81.51%
RECALL	81.48%	81.52%	81.99%
F1- SCORE	81.35%	81.38%	81.75%
AIC	67596	67595	67075
BIC	67939.27	67929.12	67650.81

Table 1 Model Comparison

```
>
> #likelihood ratio test
> # Perform the likelihood ratio test
> lr_test <- anova(initial_model, interaction_model, test = "Chisq")
> print(lr_test)
Analysis of Deviance Table

Model 1: y ~ age + job + marital + education + default + balance + housing +
    loan + contact + day + month + duration + campaign + pdays +
    previous
Model 2: y ~ age + job + marital + education + default + housing + loan +
    contact + month + day + duration + campaign + pdays + previous +
    (age * job) + (marital * education) + (housing * loan) +
    (contact * month)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     79807      67522
2     79782      66951 25   570.66 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Figure 22 Likelihood Ratio Test

INDIVIDUAL CONTRIBUTION AND TIME ALLOCATION

HARSHITHA ATLURI– Data Preprocessing & EDA

Working on data preprocessing was both challenging and rewarding. The process of cleaning the dataset, handling missing values, and removing unnecessary columns required careful attention to detail, but it was satisfying to see the data become ready for analysis. The exploratory data analysis (EDA) helped me gain valuable insights into customer characteristics that influence term deposit subscriptions. By visualizing trends like age, marital status, and education, I better understood how these factors contribute to subscription decisions. I also learned the importance of data integrity and consistency throughout the preprocessing phase. This process not only strengthened my data cleaning skills but also enhanced my ability to identify key patterns and relationships in the data that could impact the model's predictions.

CHARITHA NALLAKA – Model Building & Evaluation

Building and refining the models was an exciting and enlightening experience. Starting with the Generalized Linear Model (GLM) as a baseline gave us a strong starting point. Using backward elimination with AIC to improve the model was a valuable learning experience, as it helped me understand how to simplify models while maintaining accuracy. Adding interaction terms to capture combined effects was particularly insightful, as it deepened my understanding of how different factors, like age and job type, influence subscription rates together. Evaluating the models using metrics like accuracy, precision, recall, and F1-score was essential for comparing their performance and determining the most effective approach. I also learned how to balance model complexity and interpretability, which is crucial for real-world applications. This role significantly improved my skills in model building, feature selection, and evaluation.

GEETHIKA SANNALA – Goodness of Fit & Conclusion

Analysing model performance through tests like deviance and likelihood ratio provided me with valuable insights into how well the models fit the data. I gained a deeper understanding of the importance of interaction terms in improving the model's explanatory power. Comparing AIC and BIC values helped me learn how to assess the trade-off between model fit and complexity, ensuring the most optimal model. Summarizing our findings and drawing actionable conclusions was particularly fulfilling because it allowed me to connect the technical aspects of the project to practical recommendations for the marketing campaign. The process helped me develop my ability to critically analyze model results, assess statistical tests, and communicate findings clearly. I also learned how to prioritize key insights to make the results meaningful for decision-makers.

DENNIS MC LEAORD ANNAVARAPU– PPT & Report

Creating the final presentation and report was a rewarding experience that helped me develop my communication skills. I focused on presenting our findings in a clear, structured way, making sure to highlight the key points of the analysis. Designing the PowerPoint to be visually appealing was important for engaging the audience, and ensuring the report was thorough allowed me to provide detailed insights into the methodology and results. This process reinforced the importance of translating complex analyses into accessible formats for both technical and non-technical audiences. Additionally, I learned the significance of teamwork in pulling together all the components of the project, ensuring that our final deliverables were cohesive and aligned. This role improved my ability to distil complex information into a concise and understandable format and emphasized the importance of collaboration throughout the project.