Predicting success of telemarketing campaigns in term deposit subscriptions

KARTHIK RADHAKRISHNAN

OVERVIEW

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

OBJECTIVE 2 GOAL

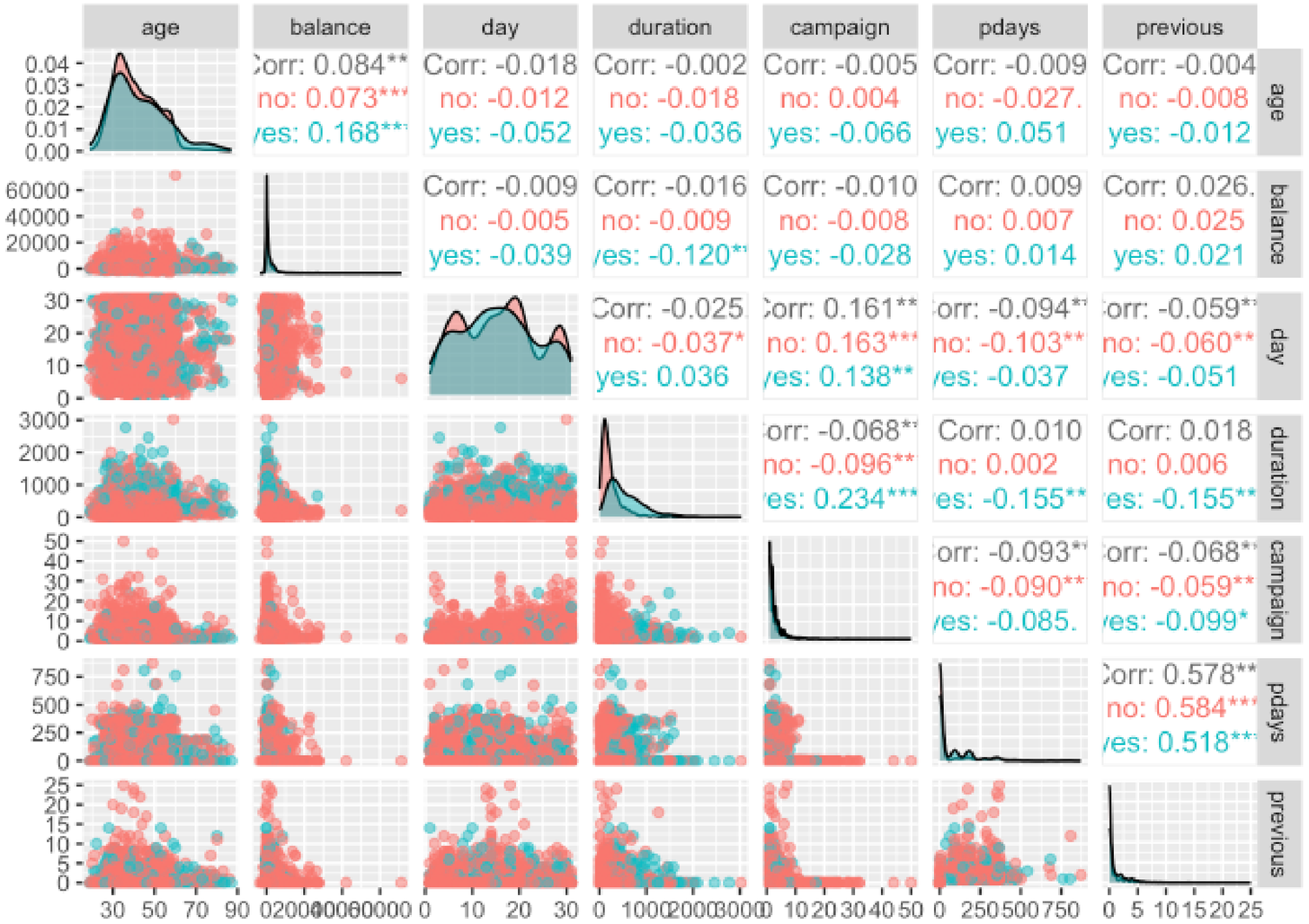
- The goal of this objective is to build a model where prediction performance is prioritized.
- 3 additional classification models should be built and compared based on 6 metrics (Sensitivity, Specificity, Prevalence, PPV, NPV, and AUROC).
- Summarize the overall findings and provide recommendations for what model should be used for making future prediction.

dataset description

The dataset contains 16 input variables related to bank client information, marketing campaign details, and contact history, used to predict term deposit subscription (target variable). These variables are grouped as follows:

Client Demographics and Financial Data	• **Demographic data:** Age, Type of Job , Martial Status, Education • **Financial data:** Default (whether client has credit in default), Balance, Housing loan information, Personal Loan information
Contact Information	• **Contact:** Type of contact communication (categorical: “telephone”, “cellular”, “unknown”). • **Day:** Day of the month when the last contact occurred (numeric). • **Month:** Month of the year when the last contact occurred (categorical: “jan”, “feb”, …, “dec”). • **Duration:** Duration of the last contact in seconds (numeric).
Campaign performance	• **Campaign:** Number of contacts during the current campaign • **Pdays:** Number of days since the client was last contacted from a previous campaign • **Previous:** Number of contacts before the current campaign • **Poutcome:** Outcome of the previous campaign (categorical: “success”, “failure”, “unknown”).

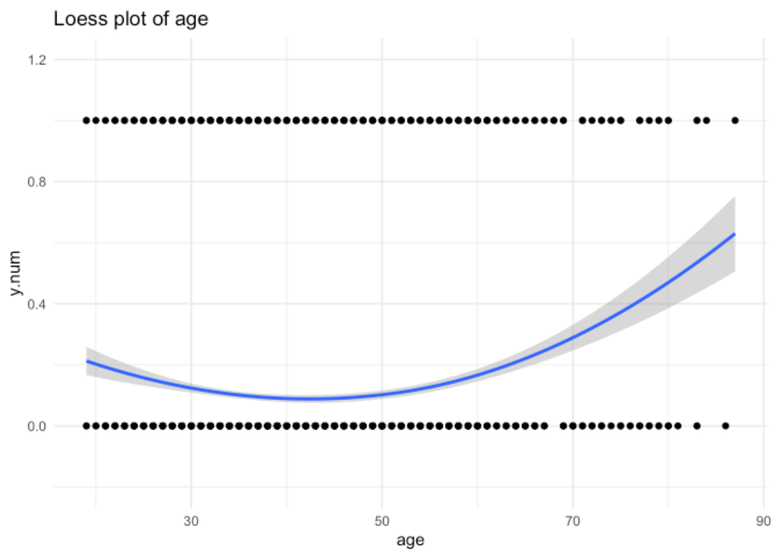
EDA – Class separation analysis

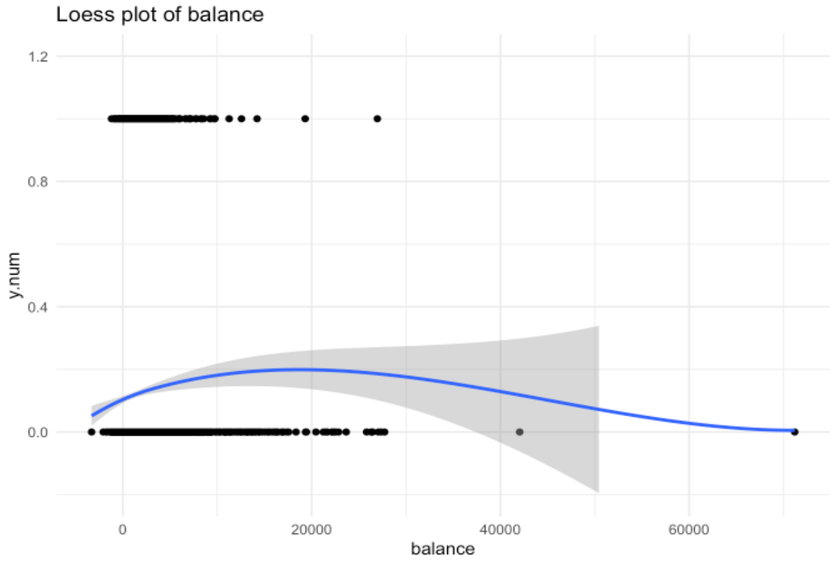


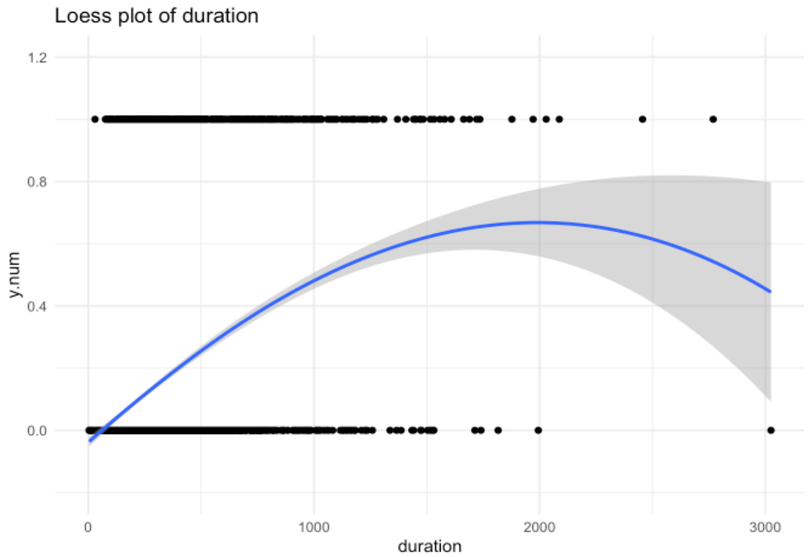
Key Insights

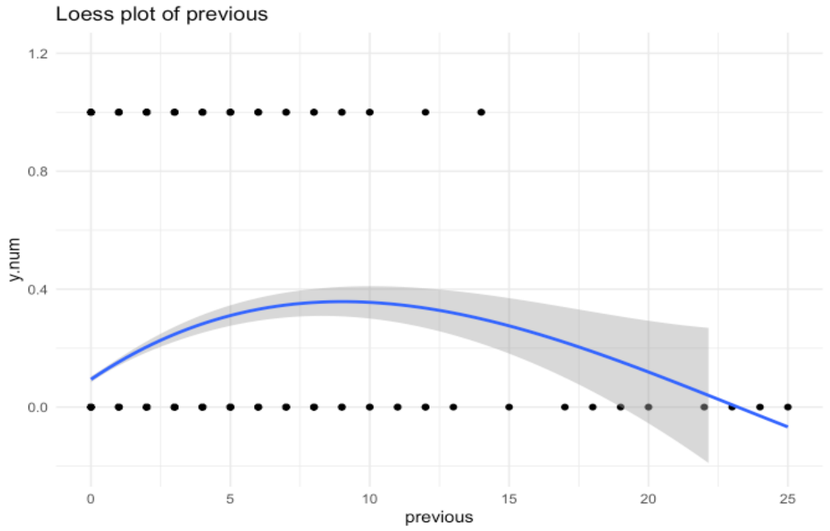
- Variable ‘Duration’ strongly distinguish between yes and no subscriptions, with clear evidence that longer call durations often lead to subscriptions.
- Other variables such as balance, day, campaign, pdays and previous doesn’t show significant separation, indicating these variables may not be a strong differentiator for subscription.
- **Correlation analysis**: pdays and previous are having moderate positive correlation. None of the other numerical variables shows significant correlation with each other.

EDA – Relationship between numerical variables and response





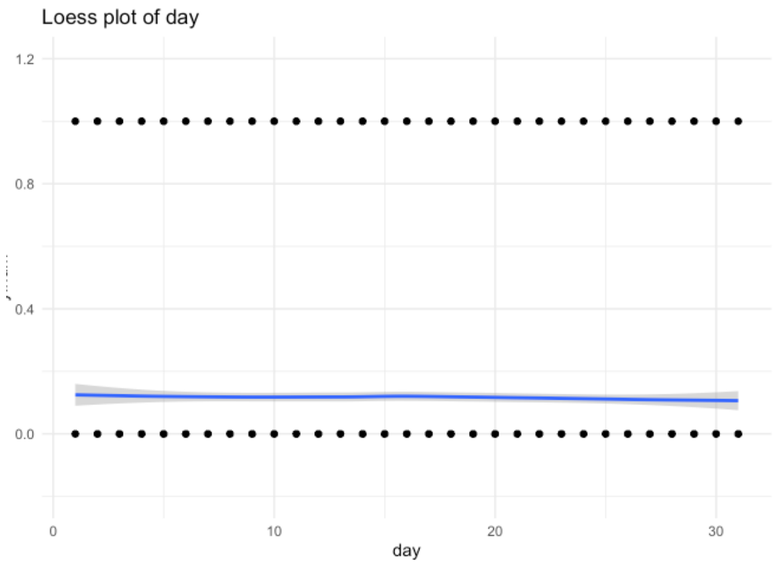


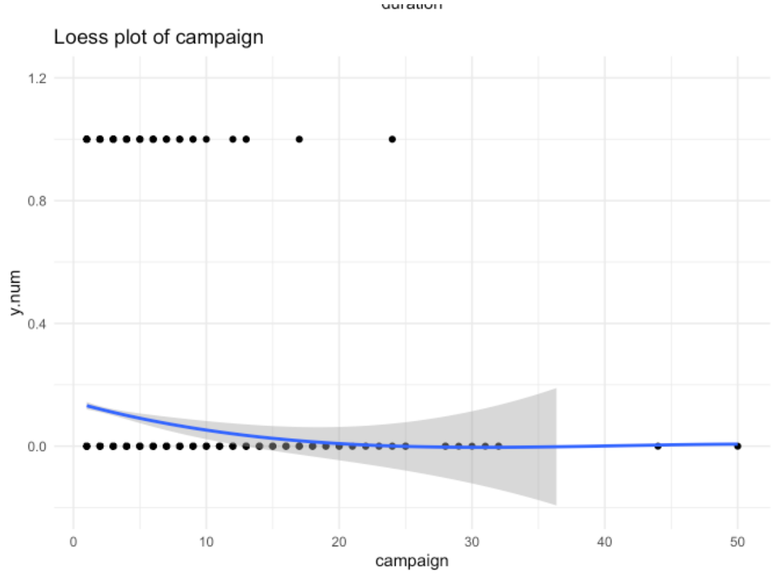


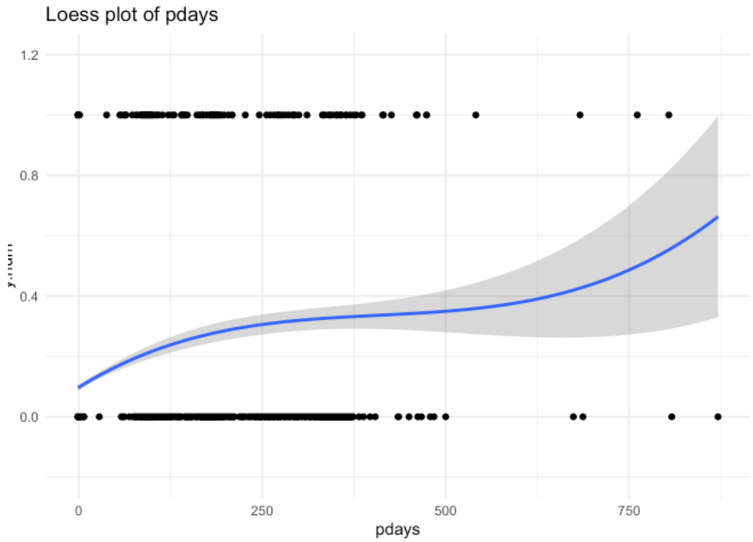
Key Insights

Variables such as Age, Balance, Duration & Previous have non linear relationship with likelihood of subscription.

EDA – Relationship between numerical variables and response



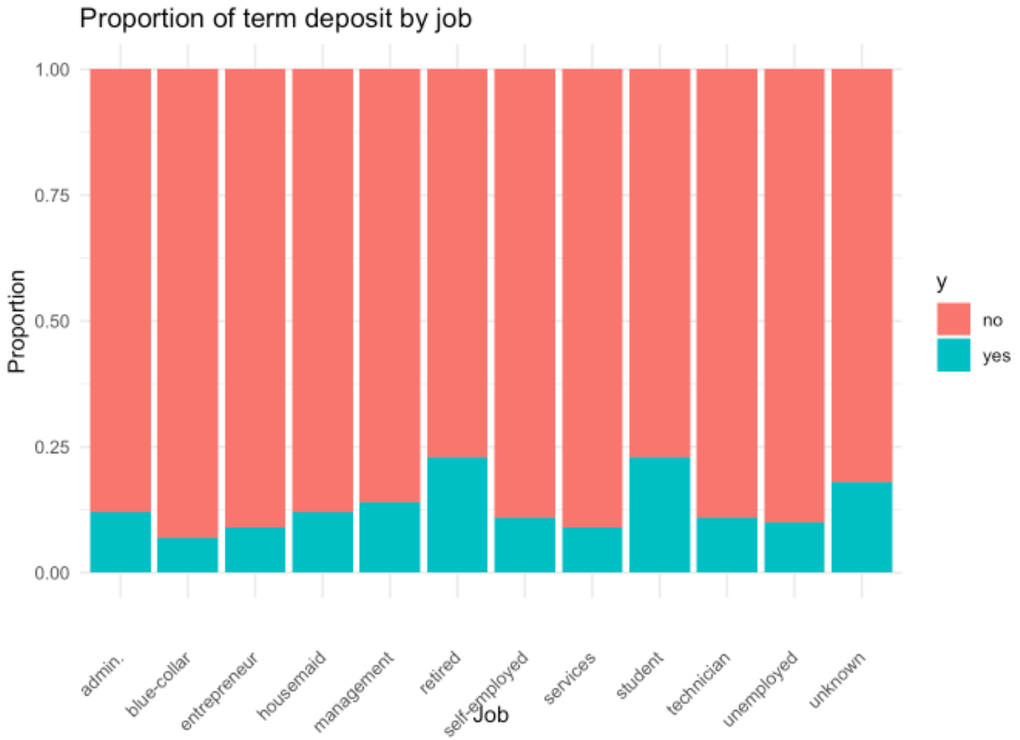


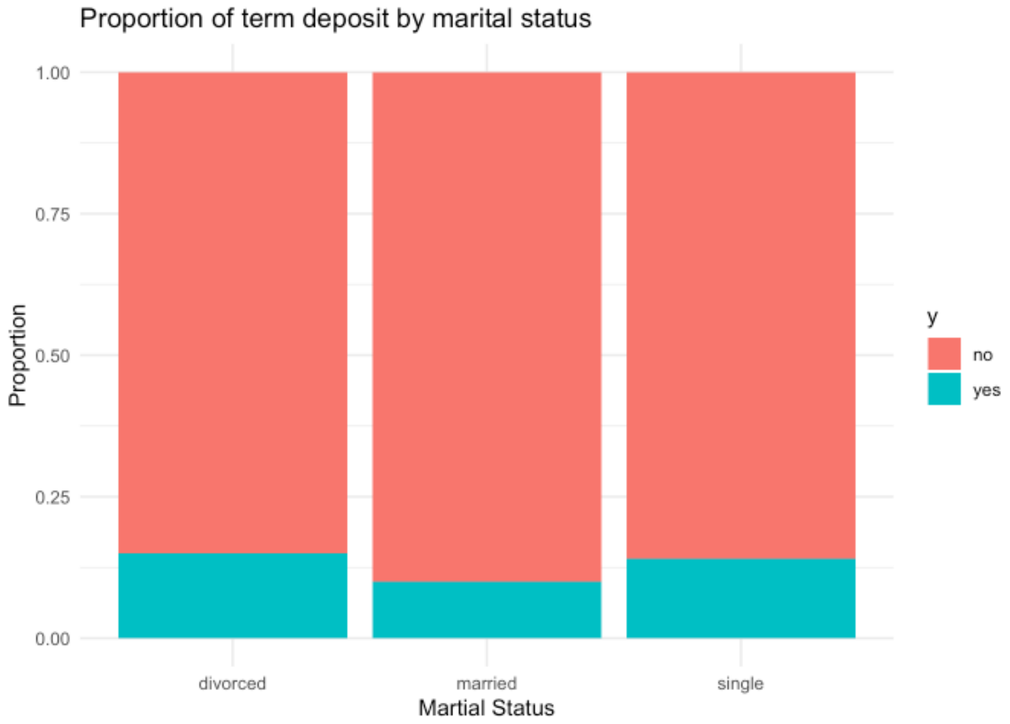


Key Insights

- Day of the month does not have a significant impact on whether a client subscribes to a term deposit.
- Slight negative trend between the number of contacts made during the campaign (campaign) and the likelihood of subscription.
- Positive relationship between pdays (number of days since the client was last contacted in a previous campaign) and likelihood of subscription. The increase in subscription likelihood becomes more pronounced for values of pdays above 400 days. This suggests that clients who were contacted a long time ago (or rarely contacted) may be more receptive to the current campaign.

EDA – impact of categorical variables on the response

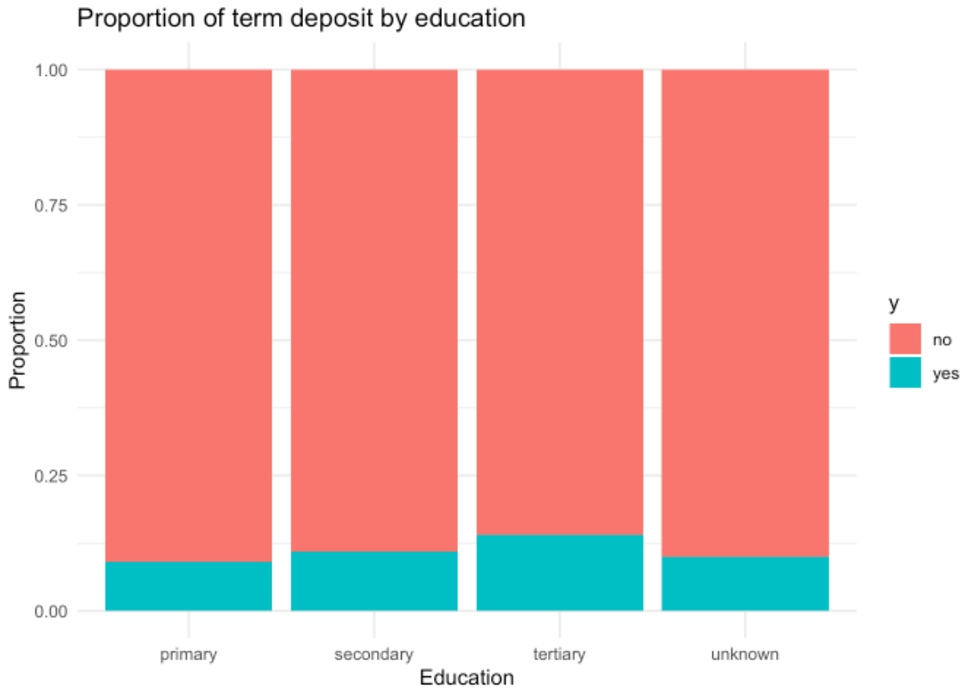




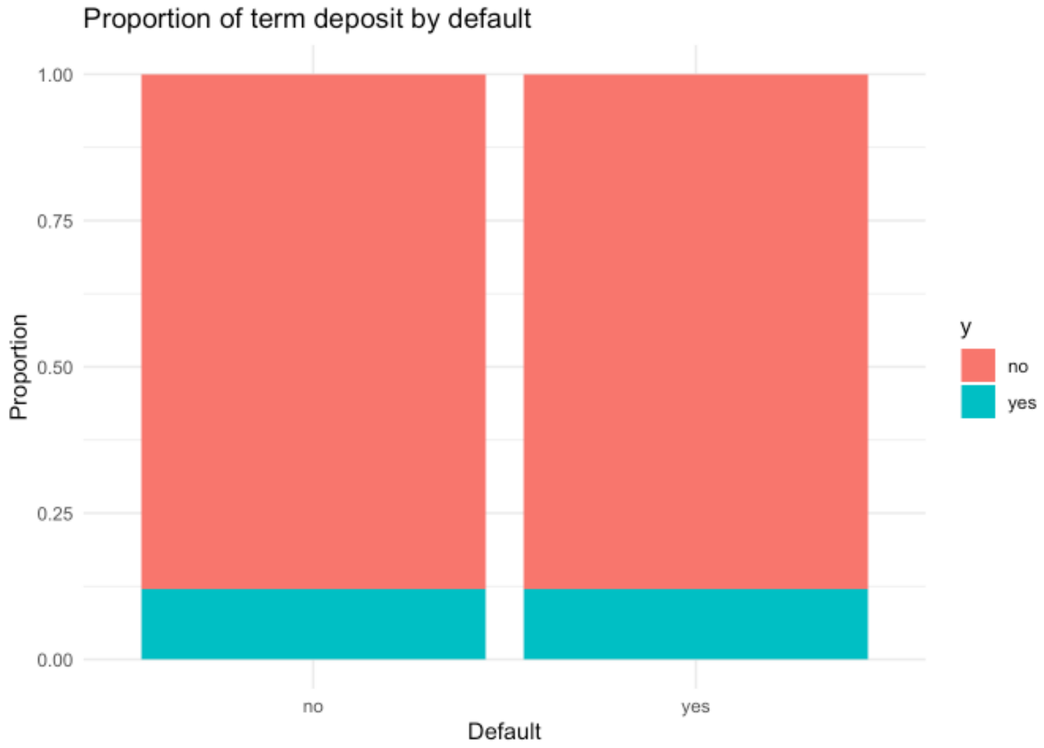
- Job roles such as retired and student show a notably higher proportion of subscriptions compared to other job roles.
- In contrast, roles like blue-collar and housemaid have lower subscription proportions.

- The proportion of clients who subscribed to term deposits is relatively similar across all marital status categories: divorced, married, and single. This indicates that marital status may not be a significant predictor in a model.

EDA – impact of categorical variables on the response

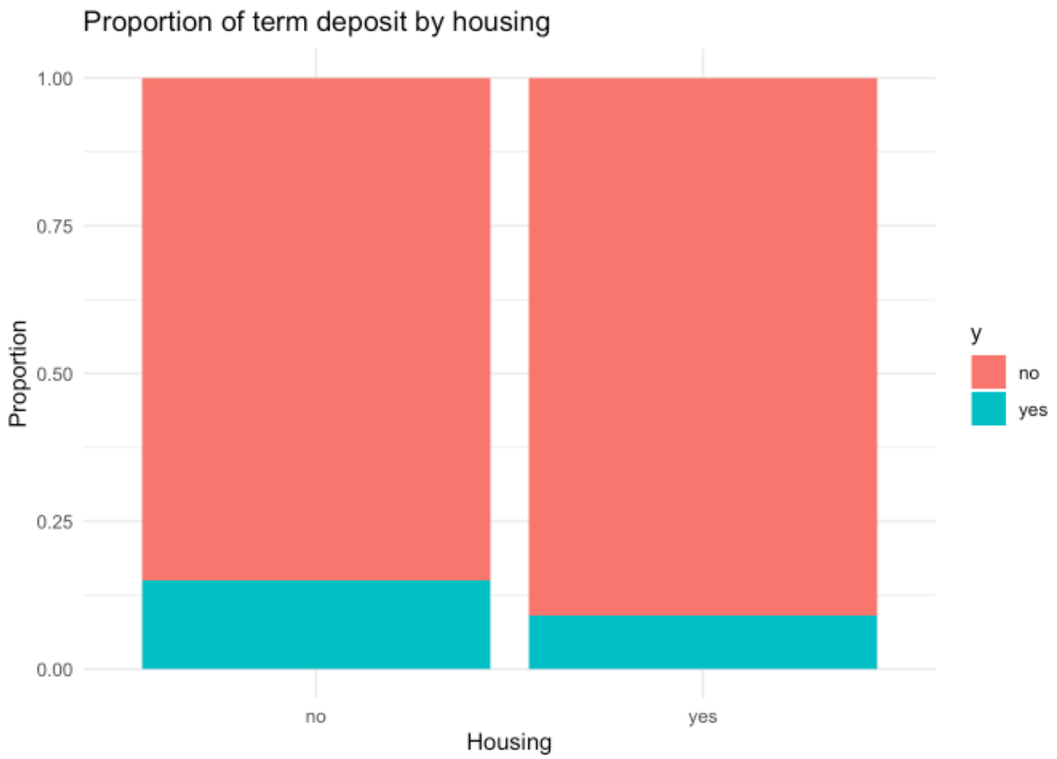


- Tertiary educated clients have higher subscription rates compared to other groups.

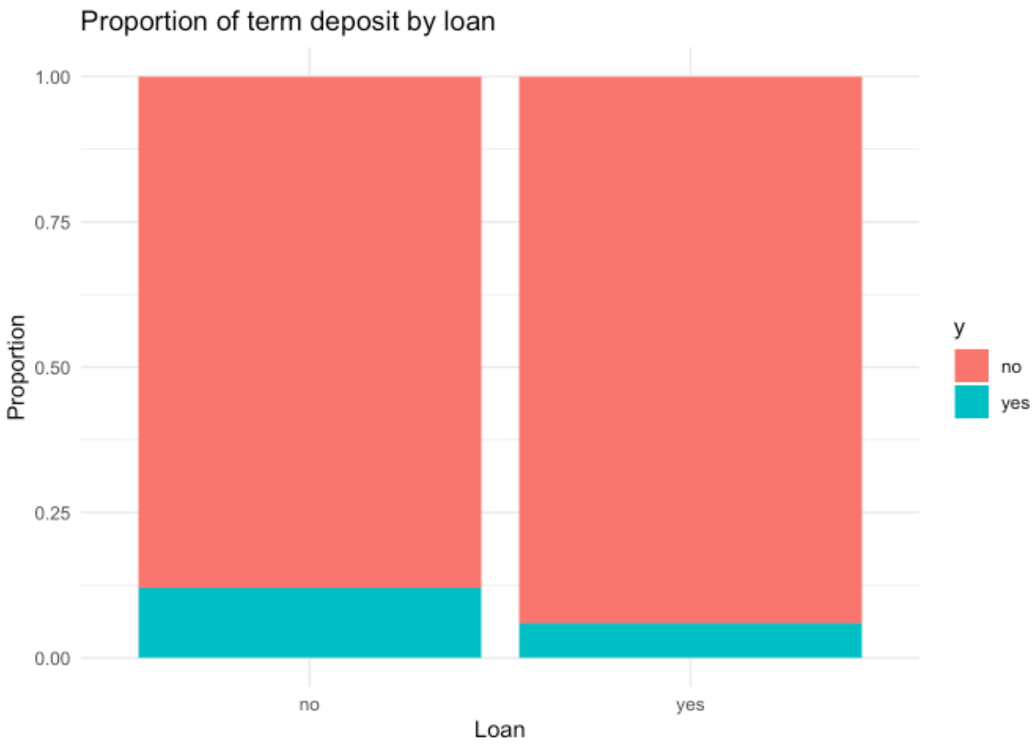


- Presence of a credit default does not have a significant impact on the likelihood of subscription.

EDA – impact of categorical variables on the response

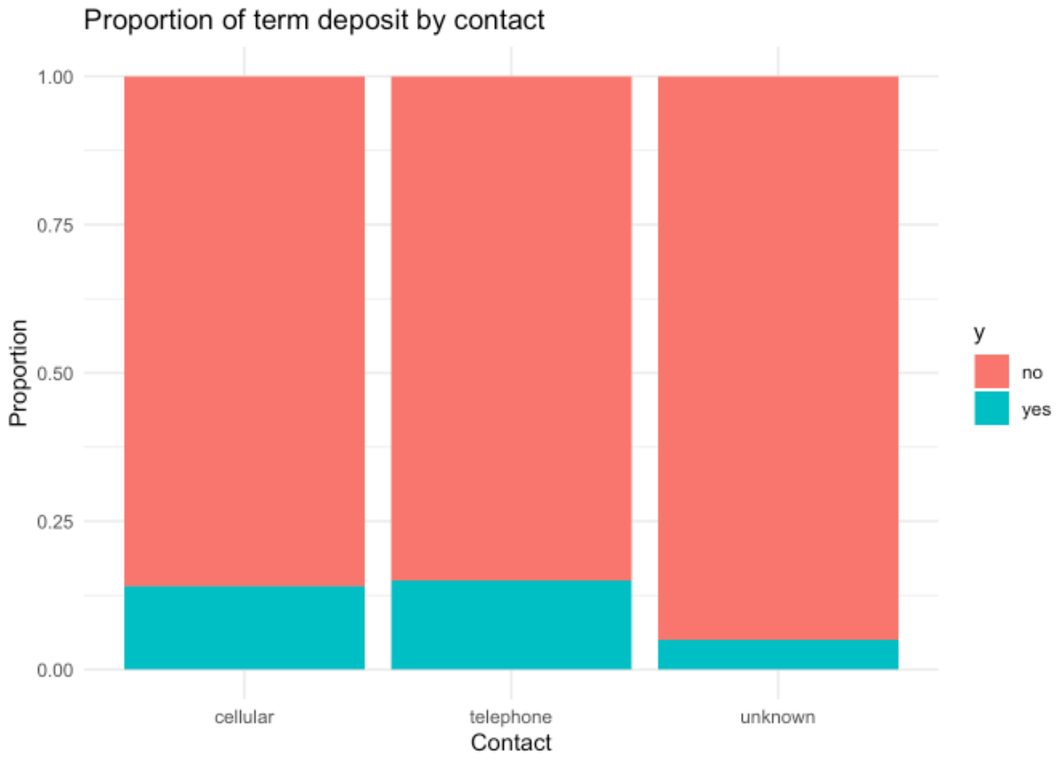


- The proportion of clients who subscribed is slightly higher for those who do not have a housing loan compared to those who have a housing loan.

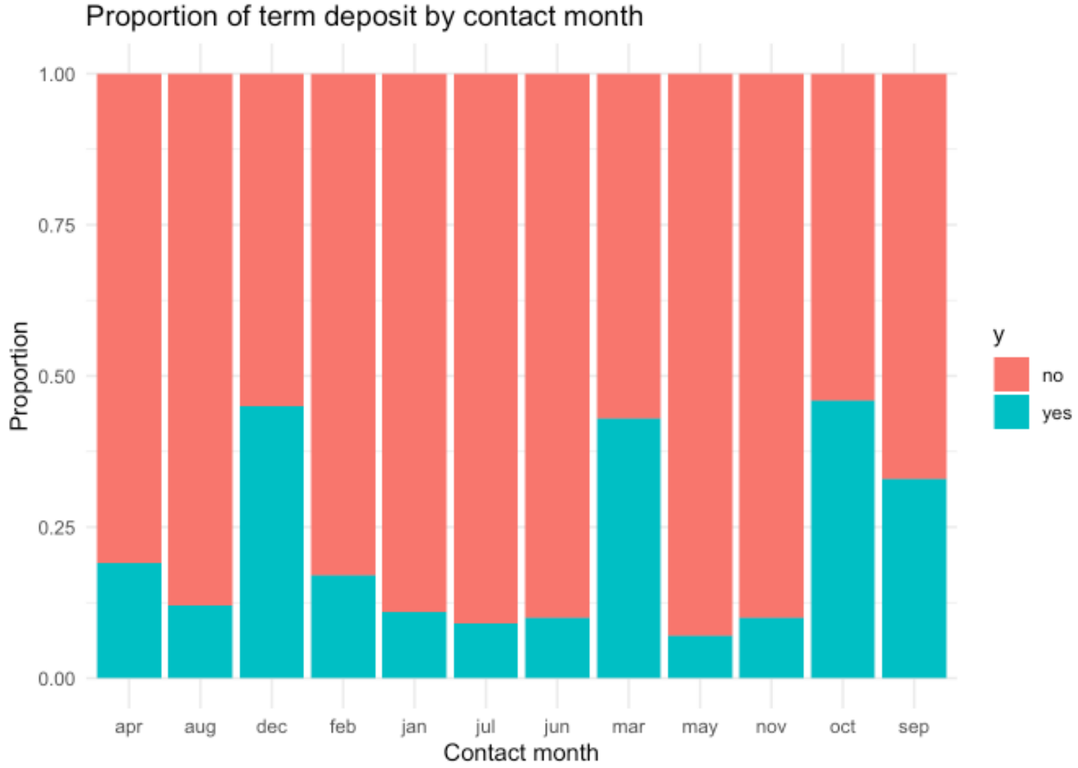


- Having a personal loan seems to correlate with a marginally lower likelihood of subscribing to a term deposit. This could be because clients with loans might have reduced disposable income or prefer to prioritize loan repayments over new investments.

EDA – impact of categorical variables on the response

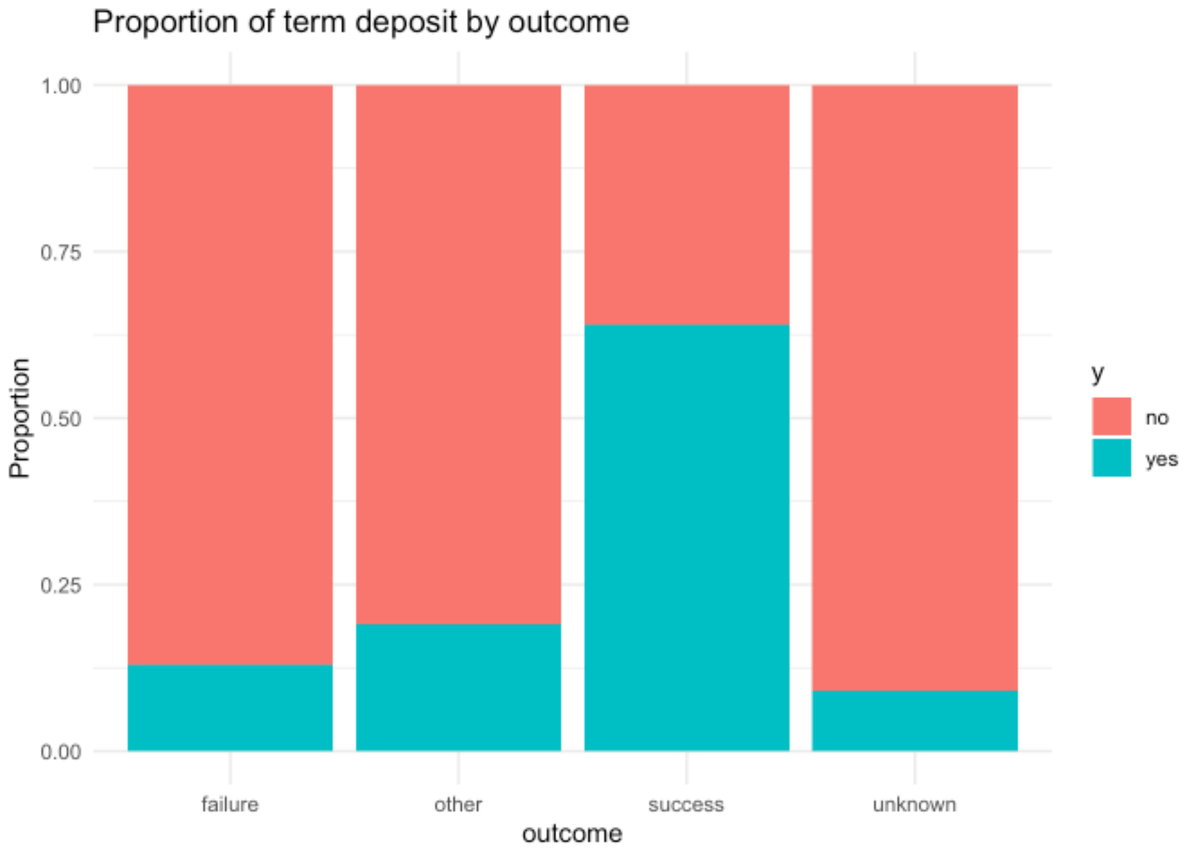


- The proportion of clients who subscribed is higher for contacts made via telephone compared to those made via cellular or where the contact type is unknown.
- The unknown category shows the lowest proportion of subscribers.



- March, September, December, and October show higher subscription rates compared to other months.
- May, July, and June show the lowest subscription rates.

EDA – impact of categorical variables on the response



- Clients with a “success” outcome in the previous campaign have the highest subscription rates in the current campaign. The proportion of subscribers is significantly higher for this group.
- Clients with a “failure” outcome show a much lower subscription rate, indicating that a prior unsuccessful campaign negatively impacts the likelihood of subscribing.

approach followed for building complex logistic regression model

Step 1. Implemented feature selection using lasso regression

- **Variables selected by Lasso regression:** job, martial, housing, loan, contact, month, duration, poutcome

Step 2. Built a simple logistic regression model using the features selected by lasso regression.

- **Simple logistic regression Model:** y~job + marital + housing + loan + contact + month + duration + poutcome

Step 3 . Added complexity by adding polynomial and interaction terms for the statistically significant variables from the simple model.

- Polynomial term were included for the statistically significant numeric variable (Duration) to account for it's non-linear relationship with the response variable.
- Interaction terms are added between statistically significant variables from the simple model
- **Complex model :** y ~ job + marital + housing + loan + contact + month + duration + poutcome + **I(duration^2) + duration : poutcome**

approach followed for building complex logistic regression model

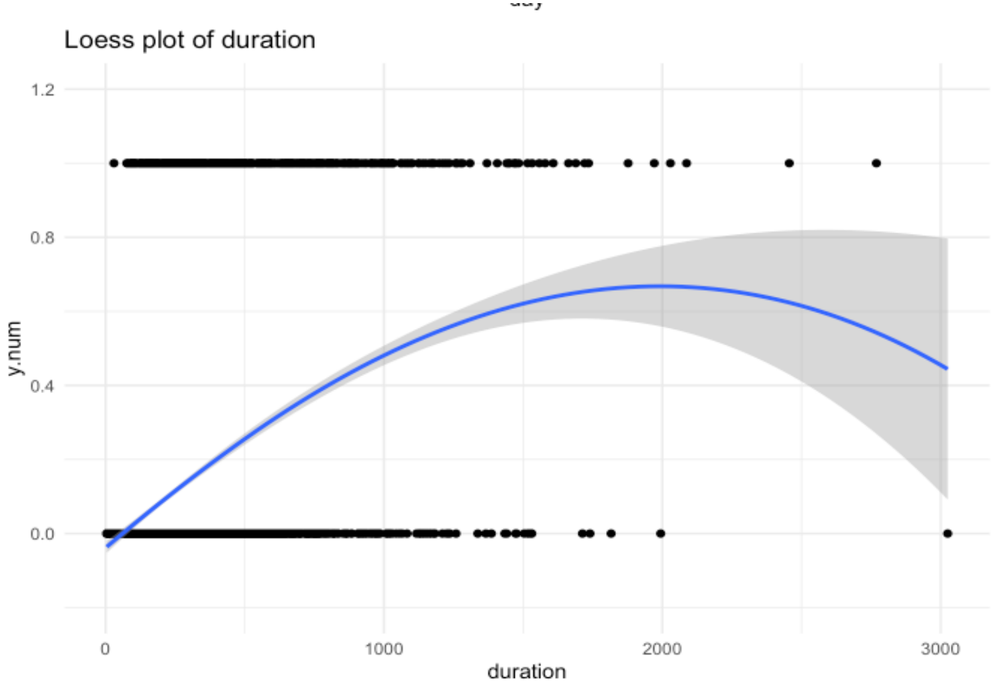
Step 4. Performed deviance test between simple and complex model.

- Deviance test is statistically significant. This indicates that at least one of the coefficients unique to complex model is not 0.

EDA Recap: Duration has non linear relationship with response.

Linear Term (Duration) - Captures the initial increase in response variable with duration.

Quadratic term (Duration ^ 2) - Captures the eventual decline for very high values of duration.



METRICS for model EVALUATION

MOST IMPORTANT METRICS TO PREDIT TERM DEPOSIT SUBSCRIPTIONS

Positive Predictive Value

- Crucial because it measures the proportion of predicted subscribers actually end up subscribing.
- Important for resource allocation - This needs to be high if bank is going to assign sales staff to follow up with predicted subscribers.
- Helps minimize wasted resources on false leads

Sensitivity

- Critical because it measures the proportion of actual subscribers that the model correctly identifies.
- High sensitivity means fewer missed opportunities for successful sales.
- Important for maximizing revenue opportunity - Bank shouldn’t be missing potential customers who would subscribe.

AUROC

- Very important because it shows the model's ability to distinguish between customers likely to subscribe and those who won't across different threshold value.
- Helps in finding the optimal threshold for your classification model based on business needs.
- Provides a single score (AUC) to compare different models.

METRICS for model EVALUATION

LESS CRITICAL METRICS TO PREDIT TERM DEPOSIT SUBSCRIPTIONS

Specificity

- Less important in this context because misidentifying a non-subscriber as a potential subscriber is less costly than missing a potential subscriber.
- However, still might be relevant for managing resource efficiency.

Negative Predictive Value

- Less crucial because the cost of missing a non-subscriber is lower than missing a potential subscriber.
- Still might be useful for understanding model performance.

Prevalence

- While useful for understanding data distribution, it's less important for model comparison
- However, it might be used to consider when setting model thresholds and interpreting other metrics

MODEL COMPARISONS

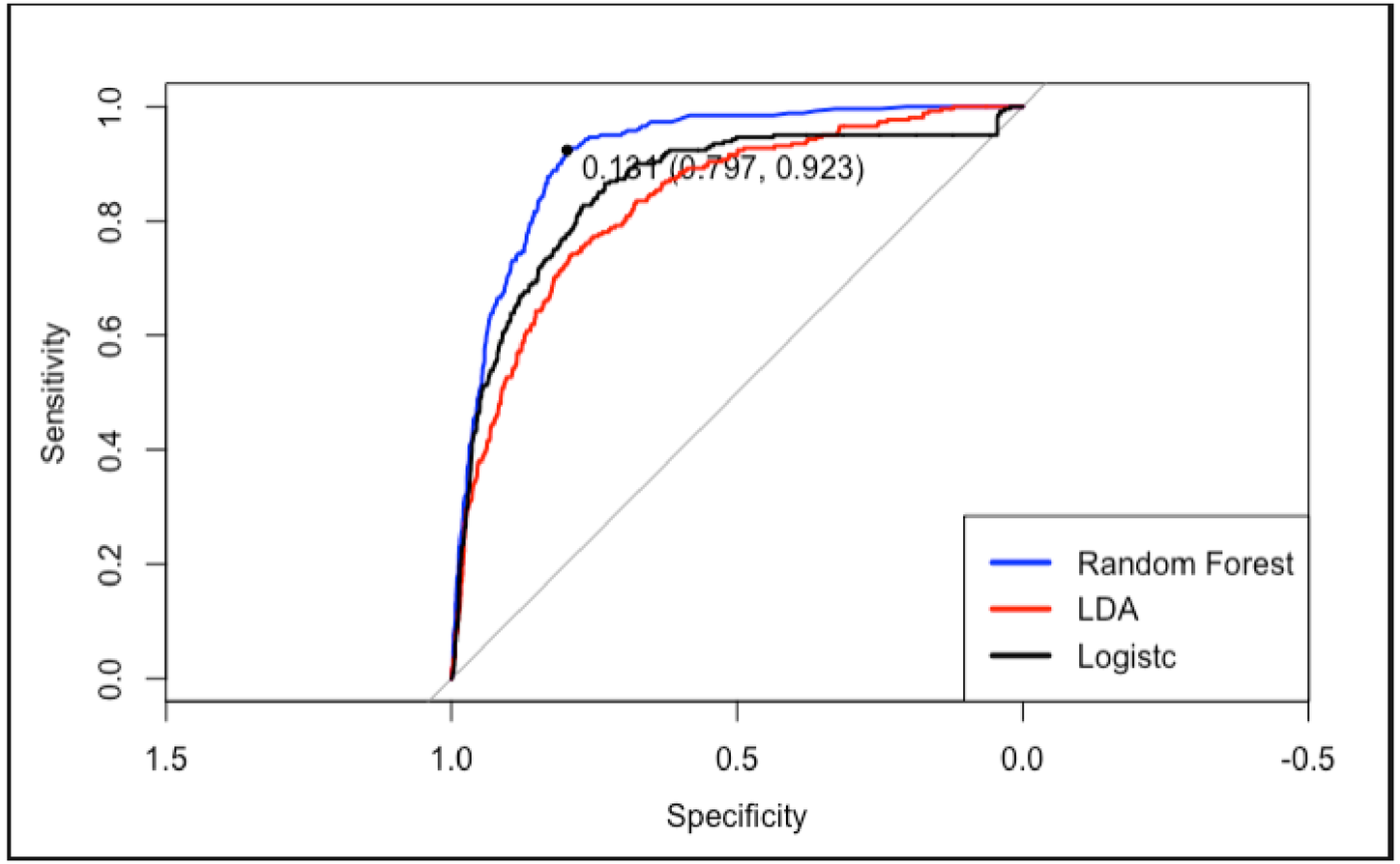
Metric	Complex Logistic Regression	LDA	Random Forest
Sensitivity	0.38846	0.31923	0.36154
Specificity	0.97000	0.96450	0.97300
Positive Predictive Value	0.62733	0.53896	0.63514
Negative Predictive Value	0.92425	0.91595	0.92140
Prevalence	0.11504	0.11504	0.11504
AuROC	0.898	0.832	0.916

Key Insights

Random forest has the balanced metrics with highest PPV & AuROC, balancing the need to capture potential subscribers while avoiding wasted efforts on false positives.

Note: 0.5 is used as threshold in logistic regression and LDA models.

ROC CURVES



Key Insights

Random forest model (blue curve) lies closer to the top-left corner compared to the others, indicating the highest overall discriminatory ability.

Model performance analysis

REASONS FOR BETTER PERFORMANCE OF RANDOM FOREST MODEL

Better Handling of non-linear relationships

- **The relationship between most of the features such as duration and previous and the likelihood of subscription is often non-linear, which Random Forest captures effectively.**
- **Simpler models like Logistic Regression or LDA assume linear relationships, which can limit their performance when these assumptions don't hold.**

Captures Feature Interactions

- **Random Forest naturally captures interactions between features without requiring manual specification of interaction terms.**
- **For example, interactions between month (seasonality) and contact (type of communication) could influence the likelihood of subscription, which Random Forest can detect and leverage, whereas Logistic Regression and LDA would need these interactions explicitly modeled.**

Model performance analysis

REASONS FOR BETTER PERFORMANCE OF RANDOM FOREST MODEL

Higher AUC and Generalization

- **The AUC (Area Under the ROC Curve) of Random Forest is higher because it balances sensitivity and specificity more effectively across different thresholds.**
- **Random Forest generalizes better because it aggregates the predictions of multiple trees, leading to smoother decision boundaries.**

No assumptions about data distribution

- **Logistic Regression and LDA rely on strong assumptions about the data:**
  - **Logistic Regression assumes a linear relationship between features and the log-odds of the outcome.**
  - **LDA assumes normally distributed features with equal covariance matrices for each class.**
- **Random Forest makes no assumptions about the data distribution, making it more versatile in handling real-world datasets that often violate these assumptions.**

Summary Conclusions

Summary of Findings in Objective 2

Complex Logistic Regression

- Highest sensitivity. It captures more actual subscribers compared to other models.
- Lower PPV and AuROC compared to Random Forest, indicating less reliable predictions and weaker overall performance, compared to random forest.

Random Forest

- Best PPV hence it's ability to capture how many of the predicted subscribers actually subscribe, is better than other models.
- Best AuROC . This indicates that this model has the best overall discriminatory power.
- Slighly lesser sensitivity than logistic regression model but almost equal.

LDA

- Lowest sensitivity, PPV and AuROC across the 3 model, making it the weakest option overall.

MOST OPTIMAL MODEL FOR MAKING FUTURE PREDICTIONS : RANDOM FOREST