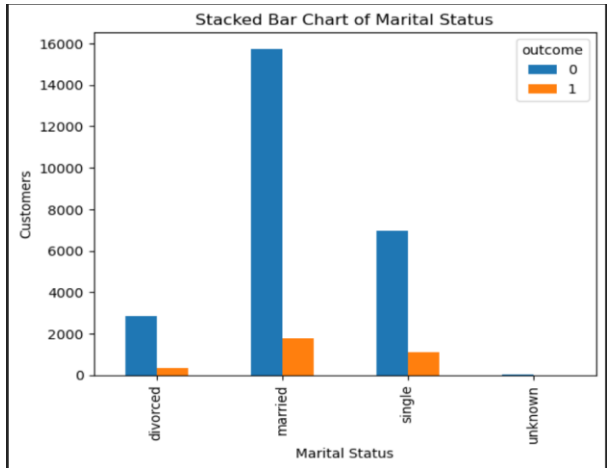
Executive Summary

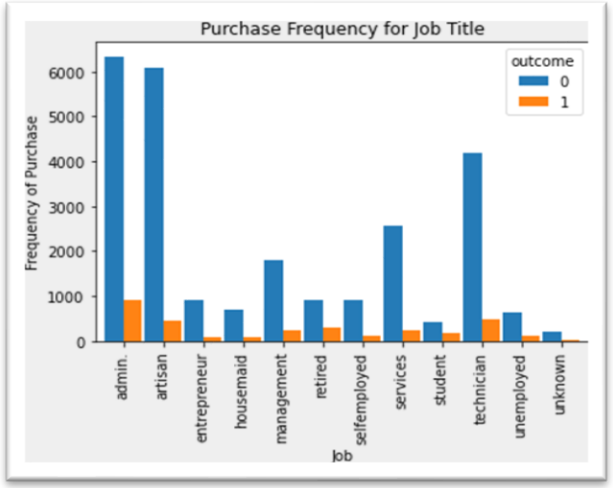
Predictive Modeling on a Bank Dataset

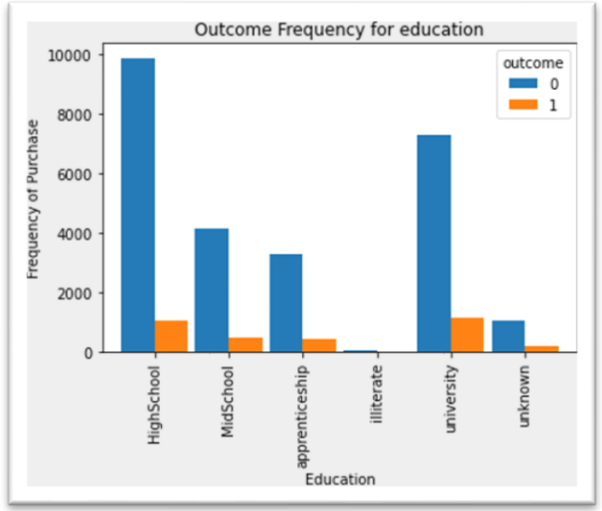
Team - Roshan Dubey, Saloni Tiwari, Yashveer Shrimal

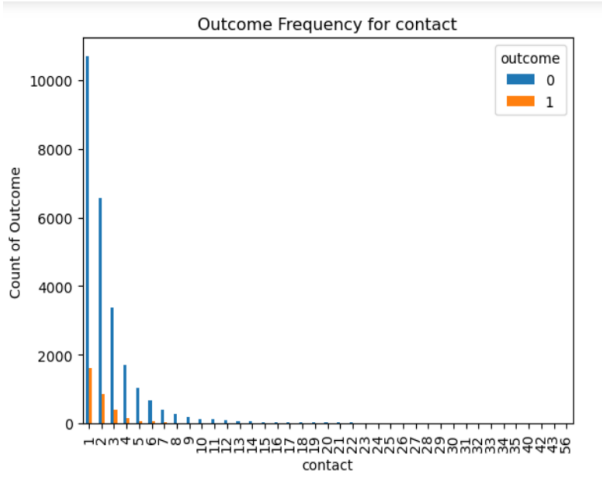
Data Understanding

- The dataset contains information on customer demographics and other relevant factors like whether they have a home or personal loan that affect their likelihood of opening a saving account.
- Among all the job types, ‘admin’, ‘artisan’ and ‘technician’ seem to have the highest positive outcome of our marketing campaign.
- Among all the educational qualifications, ‘University’ and ‘High School (9K and 12K)’ seem to have the highest positive outcome of our marketing campaign.
- Most of the customers contacted during the campaign are in the age range of 24-50 years.
- Almost 78% of the customers did not have credit in default.
- The distribution of customers with a home loan is quite even, approximately half of the customers have a home loan.
- Higher number of calls made does not necessarily help in a successful outcome.





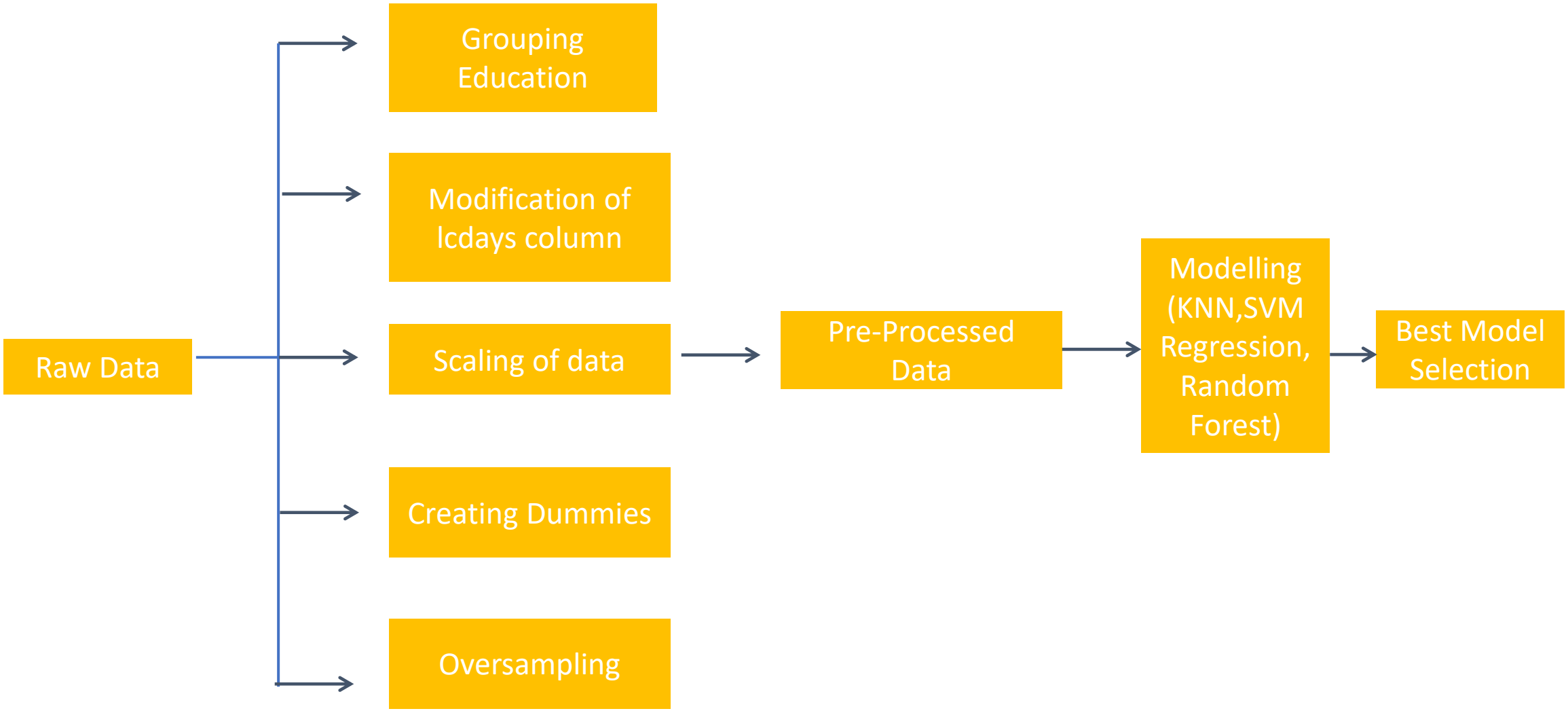




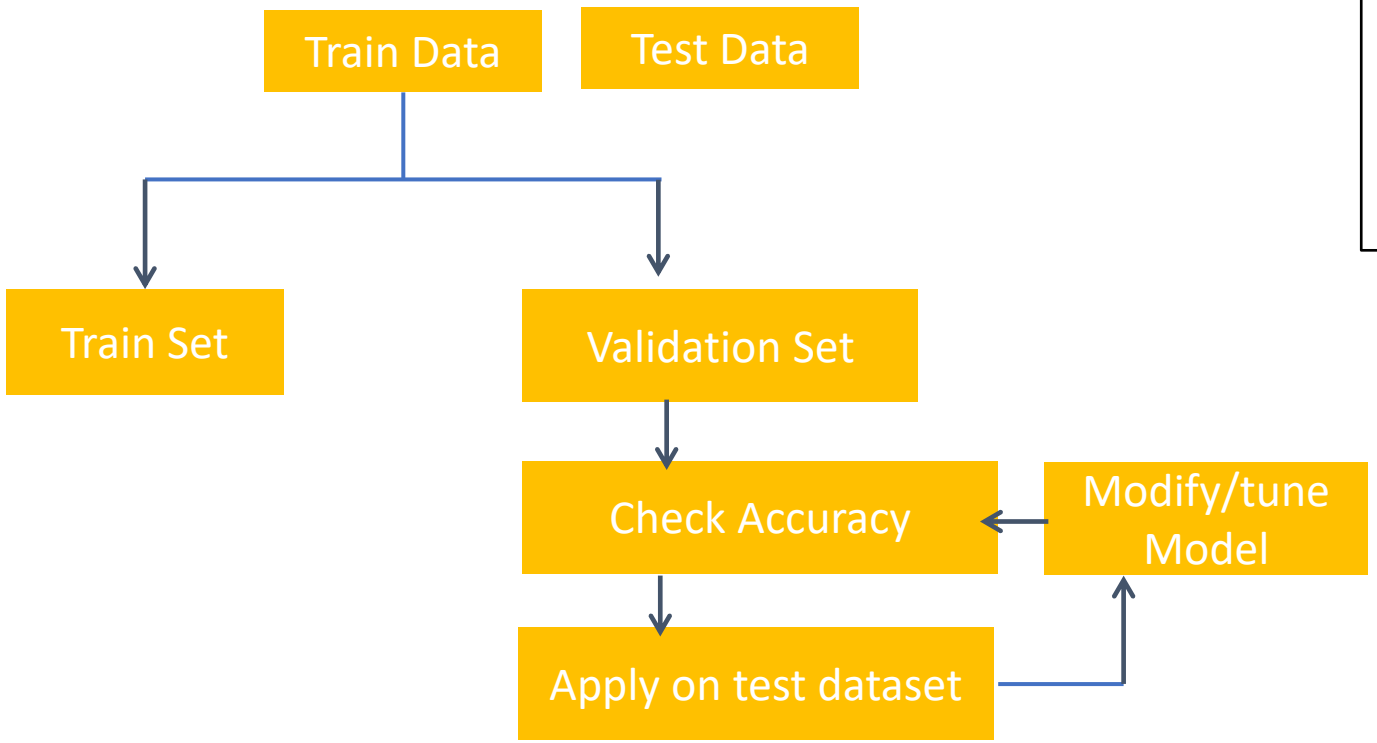
Data Preparation

- The data has education divided into 4K, 6K, 9K, and 12K, to reduce the complexity of our model we performed variable reduction and merged 4K & 6K to Mid-School and 9K & 12K to High-School.
- The column for ‘lcdays’ indicated that ‘999’ meant that the customer has not been contacted earlier, so to simplify this we replaced 999 with 0.
- The data had multiple columns which used different scales, (like age and euri3) for optimum results on our algorithm, we used the “StandardScaler()” function to scale all numerical variable.
- For handling categorical variables in our algorithm, we created dummies for them using the “get_dummies” function in the “pandas” library.
- Our dataset had an imbalance since the number of No/0 in the outcome column outnumbered the number of Yes/1, this could have impacted our model performance, to overcome this we performed over-sampling using SMOTE.

Data Modeling



Evaluation



The train data is split into two parts, train & validation in a ratio of 70:30. We train our model on the train set and check the model's accuracy on the validation set. Once we analyze the results, we modify or fine tune our model and check the performance again.

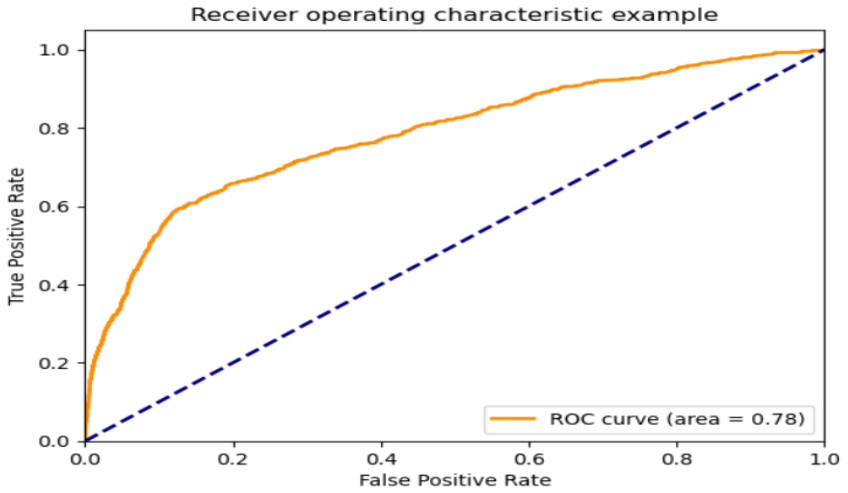


Figure : ROC Curve for Random Forest Model

Managerial Implication

- The marketing campaigns should be focused on the job types with the highest positive outcome, which are 'admin.', 'artisan', and 'technician'. This can help to increase the response rate of the campaign and ultimately result in a higher return on investment (ROI).
- The bank's credit risk is relatively low since the 78% of the customers are not credit defaulters. The bank has a creditworthy customer base hence it could ease the regulations to attract more customers.
- Higher number of calls does not necessarily mean that the customer will convert, hence the bank should focus on the quality of their sales call rather than the quantity.
- Most customers are in the range of 25-50 years of age; hence the bank could offer special promotions for this age range to attract a greater number of customers.
- Amongst various customers, married customers are more likely to open an account, hence the bank can offer specialized discounts or offer a sign-up bonus if a married couple wants to open an account together, or a joint account in both names.