**NANYANG TECHNOLOGICAL UNIVERSITY**

**NANYANG BUSINESS SCHOOL**



**BC2407 ANALYTICS II: ADVANCED PREDICTIVE TECHNIQUES**

**SEMESTER PROJECT:**

IMPROVING CONVERSION RATES OF TELEMARKETING CAMPAIGNS

**ACADEMIC YEAR 20/21 SEMESTER 2**

**SUBMITTED BY TEAM 2 SEM GROUP 5**:

CHEE WEI KIAT COLIN

LOH XIN YANG SITTIPHAN

MAPLE LIM YIN YIN

SITTI NURARFAZIRAH BINTE SHEIKH ARZIMI

VENKAT SUBRAMANIAN

**PREPARED FOR:**

PROF. CHEW CHEE HUA, NEUMANN

PROF. LIU PENG

[**Executive Summary**](#_c9w5lhx7quek) **3**

[**Introduction**](#_7b01umipxy9n) **4**

[Business Problem](#_1999es50q6ur) 4

[Dataset, Data Cleaning and Data Preparation](#_hgurlvoi0l9d) 5

[Assumptions](#_99kv6rtip94) 5

[**Analysis and Insights**](#_mzv85fkeyp7k) **6**

[Exploratory Data Analysis](#_p471ic8sc4qg) 6

[Proportion of deposit cases](#_67i34wtyrph) 6

[Job against Deposit](#_9cpngzommwbs) 6

[Marital Status against Deposit](#_tnzxvc29kn64) 6

[Education Status against Deposit](#_4noswqjt02rr) 6

[Job against Marital Status](#_598qxcbztqle) 6

[Balance against Deposit](#_3xcmgojag4g2) 6

[Job against Balance](#_9ert445rgqz3) 6

[Marital Status and Job against Balance](#_x97nset7b72f) 7

[Housing loan and personal loan on Deposit](#_4bzq8jqv3ued) 7

[Jobs against general\_loans](#_8o7dtvbue1t6) 7

[Job with loans against Balance on married prospects](#_wm82hpl14su3) 7

[Previous and Pdays against deposit](#_nvzagwiscuhr) 7

[Campaign against Deposit](#_qux7hba9pmzc) 8

[Duration against Deposit](#_kvs61obkhpnb) 8

[Month against Deposit](#_n4pf0clqdn2e) 8

[Months against General Loans](#_4hgjinkotnx) 8

[Month with Loans against Balance](#_x62hsnarllx1) 8

[Balance against Duration on Deposit](#_cwxedof7kqvx) 8

[Age against Day](#_ewcmj82qayx3) 9

[Previous with Campaign on Deposit](#_ooga5gcooy2m) 9

[Customer\_status against Loans on deposit](#_h3hjyiykmabx) 9

[Effect of Campaign with previous on deposit](#_jjisaa3a367h) 9

[Methodology](#_ywp651jyzuuh) 9

[Customer Segmentation](#_um94nc50beo2) 10

[Modelling](#_q44hztepu7ui) 11

[Logistic Regression](#_grjzo8cl6hl7) 11

[MARS](#_fgsyladnyzrn) 12

[CART](#_pl68fvgifa40) 13

[Random Forest](#_sf46q88yd6xy) 13

[Neural Network](#_kq4osaipj755) 14

[Evaluation](#_ruqcuyygupn2) 15

[**Recommendations**](#_6opea8179b0g) **16**

[Target based on seasonality](#_ndkfcxcc29pb) 16

[Target based on loan status & age group](#_h6b4mbku48v) 16

[Target based on results of previous campaign](#_7q7x6lnjx572) 17

[Target based on expected duration to win a prospect](#_t3tjak5ij8ov) 17

[Hire skilled telemarketers and have a well-crafted script](#_t79fs2hwt8t0) 17

[**Discussion**](#_o7hcoi4lh9cd) **18**

[Limitations of MARS model](#_raiqhdiqh783) 18

[Model only considers the variables from dataset](#_saudmzd6e6zw) 18

[Duration of current call is taken into account](#_4kxufrtfilmf) 18

[Difficulty for non-data-analysts to interpret the model](#_us249cspgomu) 18

[Usage of Automated Functions](#_s0icrlo1idqp) 18

[Applicability of MARS model](#_7vyxyfkpfv4o) 18

[Applicability to other industries](#_7ggzjs96fxy3) 18

[Applicability to other types of marketing](#_s77wr0y9l412) 18

[**Conclusion**](#_o2kziejxoao8) **19**

[**References**](#_4u6dmyvc8n2j) **20**

[**Appendix**](#_1dxu1xbskzi0) **21**

[Data Dictionary](#_z8xo2kn3d2hd) 21

[Tables & Figures](#_3og99ojoaiuo) 22

# Executive Summary

Outbound telemarketing such as cold-calling faces legal restrictions and is less desired by prospects due to its intrusive nature, yet it remains an effective tool in sales and marketing especially for insurance companies and financial institutions. Telemarketing has increased in prominence since COVID-19 requires social distancing and safe business practices.

This report analysed data from a Portuguese banking institution with the aim to improve conversion rate of a telemarketing campaign, which can reduce unnecessary telemarketing costs.

During the Data Exploration stage, the following relevant information was found. Those who tend to subscribe include: students, retirees, unemployed, single people, those who completed tertiary education, those who have a slightly higher bank balance, have no loans, have been contacted before, were on the call for longer. Months with the most conversions are March, September, October, and December. However, there may be other unknown determining factors and the level of importance is unknown at this stage.

The models built aim to use the variables in the dataset to classify the deposit variable into “no” (negative) and “yes” (positive). The dataset was split into 70% train set and 30% test set. The team evaluated the models based on sensitivity followed by accuracy.

Several models were utilised and fine-tuned to predict outcome. Logistic Regression had an accuracy of 83.7% and sensitivity of 84.3%. The MARS model tuned using a 10-fold CV and CART had an accuracy of 83.2% and sensitivity of 90.1%. CART had an accuracy of 83.7% and sensitivity of 86.5%. Random Forest had an accuracy of 85.7% and sensitivity of 89.4%. Neural Network had an accuracy of 82.5% and sensitivity of 86.3%. Since sensitivity is the most important metric of evaluation and the Tuned MARS model’s accuracy was not too low, the Tuned MARS model was recommended.

Using the Tuned MARS model, recommendations were suggested for the bank telemarketing campaign. Firstly, the bank should call in December, March, September, October, April, followed by February. Secondly, it should target students and retirees with no loans. Thirdly, it should target those who were interested in previous campaigns. Fourthly, it should try to reach a balance per duration of more than 1.555, and prioritise those with a shorter expected call duration. Lastly, telemarketers hired should be skilled enough to grab the attention of prospects and create desire for the product within the first 57 seconds of the call.

The first limitation to the analysis is that the model only considers the variables from the dataset even though there may be other determining factors to whether the prospect subscribes or not. Secondly, the duration of the current call is taken into account despite it being unknown until after the prospect’s decision is made. Thirdly, the MARS model is not intuitive for most non-data-analysts. Lastly, the use of automated functions make calculating the exact splitting points difficult.

This analytical model can be applied to other industries such as real estate, B2B and fundraising telemarketing. It can also be applied to other types of marketing such as digital marketing.

# Introduction

## Business Problem

Telemarketing is a direct marketing technique aimed at selling products over the phone or the internet. Telemarketing activities include inbound and outbound telemarketing. While inbound telemarketing results in more favourable outcomes, outbound telemarketing, especially cold-calling, is less desired due to its intrusive nature (Kenton, 2020). Outbound telemarketers also face legal restrictions, with some countries implementing a Do Not Call (DNC) provision, which makes it illegal for companies to call individuals listed in the DNC registry who are not their existing prospects (Hayes, 2019). In addition, the advent of automated and pre-recorded telephone calls or robo-calls make outbound telemarketing even less desirable due to the rise of phone scams (Lorette, n.d.). This poses a greater challenge for outbound telemarketing efforts as prospects often react negatively to outbound telemarketing phone calls and more people are opting into the DNC registry. However, if done right, outbound telemarketing is an effective tool (Akhrin, 2020).

Businesses that rely heavily on outbound telemarketing strategies include insurance companies and financial institutions. Due to the COVID-19 pandemic, more businesses are turning to outbound telemarketing as a safe way to connect with new prospects, with an expected compound annual growth rate (CAGR) of 4.10% between 2020 to 2024 for businesses providing outbound telemarketing services (Business Wire, 2021). One of the benefits of outbound telemarketing include the human interaction aspect, which provides the ability to build rapport with prospects in order to create trust between the prospects and the telemarketers. For example, for a financial advisor, trust seems to be the most crucial element when choosing an advisor (Lake, 2021). While putting up advertisements and/or social media profiles allow financial advisors to provide further information that can prove their credibility, relying on these methods alone is a passive approach and will not guarantee prospects calling them.

Thus, with outbound telemarketing becoming increasingly prevalent during this period, there is an opportunity to analyse this marketing strategy to improve its conversion rates, reducing unnecessary financial and non-financial expenses (NI Business Info, n.d.). Conversion rate refers to the percentage of all prospects contacted who eventually converted to customers i.e. purchased the product (Convoso, 2020). By improving the conversion rate, the cost per lead will be reduced which improves overall profitability and effectiveness of the outbound telemarketing campaign. To do so, it is important to understand the steps taken in an outbound telemarketing phone call. These include 1) pre-qualifying lead generation, 2) introduction, 3) sales pitch and 4) closing the sale. The first step is the most important step as it determines whether or not a prospect would hang up the phone. Pre-qualifying leads allow the telemarketer to only call those who are likely to be interested in their products, increasing the chance of the prospect staying on the call.

In this project, the team’s objective is to develop analytical model(s) in order to perform pre-qualifying of leads effectively, so as to improve the conversion rate of an outbound telemarketing campaign, which is the key business outcome measure for a telemarketing campaign.

## Dataset, Data Cleaning and Data Preparation

**Dataset**

The Bank Marketing Dataset from UCI Machine Learning Repository was used, accessible via this URL: <https://archive.ics.uci.edu/ml/datasets/bank+marketing>. The dataset was generated from a bank marketing campaign based on phone calls. The objective of the dataset is to predict whether or not a prospect will subscribe to a term deposit account based on certain attributes. The dataset contained 11,162 records with 17 columns. (Appendix Fig. 1). The data has 16 independent variables and one response variable named deposit.

**Data Cleaning**

The dataset did not contain any NULL or N/A values. However, a summary of the dataset revealed that there is an “unknown” category in the *job, education* and *contact* columns (Refer to Appendix Fig. 1). It appears that the missing values in those columns were categorised as “unknown”. The “unknown” category across the various columns do not belong to the same clients, hence, deleting records with “unknown” values will result in a significant reduction of records, which may make the dataset less viable for analysis. By making an assumption that the missing values are indeed missing at random, we replaced the “unknown” values with the mode of each column. To allow the analytical models to recognise whether a certain value was previously “unknown”, we created new columns e.g. *education\_unknown* which returns 1 if the value was converted and 0 if the value was not converted.

The team also had to convert the attributes with character or string data into categorical variables using the factor function. For the variable *month*, the team also reordered the levels of the variable.

## Assumptions

To make use of this data, a few assumptions were made:

1. Some values in the *balance* attribute are negative. It was assumed that negative balances mean that the prospect either has a loan with the bank or a bank overdraft.
2. Each row is assumed to represent the details of one unique prospect.
3. The attribute *poutcome* has a value of “unknown” if the specific prospect has not been contacted before. However, there were 2 prospects who were contacted before with “unknown” *poutcome*. It was assumed that the outcome of the previously-held campaign was not known even until the end of the campaign (Refer to Appendix Fig. 2).
4. As discussed in Section 1.2, it was assumed that the values are missing randomly.
5. There could be many other possible variables influencing the conversion rate that is not in the dataset. It was assumed that only the variables in the dataset affects the conversion rate.

# Analysis and Insights

## Exploratory Data Analysis

### Proportion of deposit cases

The number of prospects who have said yes and the number of prospects who have said no are almost equally distributed (Refer to Appendix Fig. 3).

### Job against Deposit

We saw that most of the prospects who have said yes to the current campaign tend to be students, retirees and the unemployed. Prospects who are entrepreneurs, housemaids and work in blue-collared jobs tend to say no (Refer to Appendix Fig. 4).

### Marital Status against Deposit

Prospects who are married and divorced tend to say no, while prospects who are single tend to say yes (Refer to Appendix Fig. 5).

### Education Status against Deposit

Prospects who have completed tertiary education tend to say yes, while the prospects who have only completed primary and secondary education tend to say no (Refer to Appendix Fig. 6).

### Job against Marital Status

This chart explains how marital status and job affects the outcome variable (Refer to Appendix Fig. 7). Majority of students tend to be single, thus explaining the insights from Fig. 4 & Fig. 5. However, it is possible to conclude that marital status is a more important factor compared to jobs. It is evident from Fig. 7 that jobs that are relatively dominated by married individuals (all jobs except for students), tend to say no. However, there are anomalies in categories such as unemployed and retirees, where these individuals tend to say yes but most of them are married. This may suggest that there are other more important underlying factors.

### Balance against Deposit

From the plot we can see that the graph for both response cases are very skewed (Refer to Appendix Fig. 8). We tried to normalise the distribution by using a combination of logarithm function as well as Gaussian function found in LambertW package. Prospects who said yes tend to have slightly higher balance than the prospects who said no (Refer to Appendix Fig. 9).

### Job against Balance

Since balance and balance\_log (a new column derived from normalising the balance) have many outliers we will be mainly focusing on the median balance. From the plot we can see that the median balance of retirees are relatively higher than the rest (Refer to Appendix Fig. 10). This could be a possible reason as to why retirees tend to say yes to the campaign.

Additionally, across all the jobs, prospects who say yes tend to have a higher median balance against than those who said no (Refer to Appendix Fig. 11). Therefore it could be assumed that balance plays a vital role in the success of the campaign.

### Marital Status and Job against Balance

Earlier it was observed that even though most of the unemployed and retired prospects tend to be married, they still say yes. This could be due to their balance. Plotting the balance against the jobs with the marital status, we observe that the retirees and unemployed who are married have a higher balance than its other marital status. (Refer to Appendix Fig. 12). This might be the reason why most retired and unemployed prospects accepted the offer even though they are married.

But from the plot, we could also observe that married prospects who work in blue-collared jobs, services and entrepreneurs have a higher balance than their counterparts in other marital status but tend to say no to the campaign.

### Housing loan and personal loan on Deposit

Prospects who have housing loans or personal loans tend to say no (Refer to Appendix Fig. 13 & Fig. 14). Since the distribution is around the same, the team has combined the 2 attributes and created an attribute named general\_loan to see the effects of prospects with no loans and with both loans.

It can be noted that prospects who have said yes to the campaign tend to have neither housing or personal loans (Refer to Appendix Fig. 15). Prospects who have either loans tend to say no. This could be due to their financial freedom. Prospects with loans have a liability to pay their loans so they do not have the freedom to deposit their money or subscribe to such campaigns.

### Jobs against general\_loans

Another reason as to why some jobs tend to say yes could be due to the presence of loans as well. Jobs that tend to have loans like blue collar jobs and services tend to say no. Also, a significant number of students, unemployed and retired prospects are shown to have no loans, which may explain why they tend to say yes. However jobs such as managers and technicians also tend to have no loans, yet they say no to such campaigns (Refer to Appendix Fig. 16).

### Job with loans against Balance on married prospects

It is shown from this figure that jobs with the most married prospects tend to not have any loans (Refer to Appendix Fig. 17). By comparing the balance of these prospects across different jobs and loans, we observe that balance and marital status has higher explainability power in determining the outcome variable compared to jobs or loans. This is suggested when comparing 2 distinct jobs such as housemaids and entrepreneurs, specifically in the area of ‘no loans’ where both jobs have the same mean for balance. Therefore, showing that low balance could be the reason as to why these jobs and why married couples tend to say no.

### Previous and Pdays against deposit

The data for previous and pdays was very skewed, since most of the data points were obtained from prospects that were newly contacted. Even with different functions to transform the variable, these 2 variables remained very skewed. Therefore we split the prospects into 2 segments. If the pdays is -1, we will classify these prospects as not being contacted before, else we will classify them as contacted before. It is observable that prospects who have been contacted before tend to say yes while prospects who are newly targeted tend to say no (Refer to Appendix Fig. 18).

### Campaign against Deposit

The data for campaign calls during the current campaign was very skewed as well (Refer to Appendix Fig. 19), hence, we transformed the data using a logarithmic transformation. The distribution of campaign\_log against different responses seems to be the same. This would imply that it does not affect the response variable. (Refer to Appendix Fig. 20)

### Duration against Deposit

The duration of the calls were very skewed as well. We used a logarithmic transformation to normalise the data. It is observed that prospects who are likely to say yes, tend to have a longer duration than those who decline the offer. (Refer to Appendix Fig. 21 & Fig. 22)

### Month against Deposit

By comparing the success rate of deposits across different months, some months are more successful than the others (Refer to Appendix Fig. 23). For example, the 3rd quarter of the year is seen to experience relatively higher levels of conversion rates. This might be attributed to the reason that working individuals may be receiving their end of year bonus or due to the fact that it is near the festive season, thus they will feel more inclined to spend.

### Months against General Loans

The reason why some months were more popular could be due to target selection (Refer to Appendix Fig. 24). The prospects targeted during these months tend to have a higher proportion of no loans. Popular months such as February, September, October and December have a higher proportion of prospects who do not have any loans. Based on section 2.1.10, people with no loans tend to say yes. This also explains why the month May had a lower customer conversion, since most of the people targeted in May had loans. The months of January, June and August also have a higher percentage of prospects with no loans, yet those months have low conversion. However, the reason why months such as January have low conversion rates can be explained by the fact that individuals tend to spend less after the holiday season (Riffkin, 2015). Additionally, since months like June and August are in the summer period, individuals may be spending their earnings in other areas instead.

### Month with Loans against Balance

From the plot, it is understandable that during the months of January, June and August the prospects with no loans have a lower balance than the rest of its counterparts (Refer to Appendix Fig. 25). Therefore these months had a lower conversion rate even though they targeted people with no loans.

### Balance against Duration on Deposit

The plot shows that the prospects who accepted a deposit have a negative relationship between duration\_log and balance\_log whereas for the no cases there is a positive correlation (Refer to Appendix Fig. 26).

The team thought that dividing balance with the duration take would be a better representation for both response cases. The resulting attribute was very skewed. To reduce the skewness we use an automated transformation function, Gaussianize, to normalise the result. By plotting the result against the response cases we can clearly see the difference (Refer to Appendix Fig. 27).

### Age against Day

When age was plotted against days, it was observed that no cases had a negative correlation whereas the yes cases had a positive correlation (Refer to Appendix Fig. 28).

Similar to section 2.1.18, the team thought a better representation will be dividing age with the day. The resulting age\_per\_day was transformed using a combination of logarithmic and Gaussianize function since it was very skewed. The distribution of the attribute is slightly different across the response variable (Refer to Appendix Fig. 29).

### Previous with Campaign on Deposit

The team wanted to discuss the effect of previous and campaign calls on prospects who have been targeted before. The team compared the number of calls previously held and the number of calls in the current campaign, and created 3 categories: “previous<campaign”, ‘previous>campaign” and “not contacted again”. It was also observed that if the customer had more calls during this campaign than the previous campaign he might be more likely to decline the offer. This could be due to irritation faced by the prospects (Refer to Appendix Fig. 30).

### Customer\_status against Loans on deposit

In section 2.1.13 we saw that most of the newly contacted prospects tend to say no. This may be due to society’s common negative stigma on telemarketers that deter newly contacted prospects to say yes. However, if we looked at the loans held by the different segments we can clearly observe that those prospects who said yes had a higher proportion of no loans (Refer to Appendix Fig. 31).

### Effect of Campaign with previous on deposit

The team wanted to see the effect of the number of calls on the response variable. The team saw that if the prospects were contacted more in the current campaign than in the last campaign (campaign > previous), more prospects say no when compared to prospects who have lesser calls in the current campaign than in the previous campaign. This could be due to the fact that when prospects are faced with more calls, they are easily irritated and more likely to decline the campaign. (Refer to Appendix Fig. 30)

From the analysis the team hypothesizes that Balance\_log, duration\_log, general\_loan would be the important variables that differentiate prospects who say yes and those who say no. For the effect of different variables on deposit, refer to Appendix B.

## Methodology

The dataset consists of 2 response (deposit) variables: ‘yes” and “no”. The team aimed to develop a model to improve the conversion rate by classifying prospects based on their predicted response. Since this is a classification problem, the team will be focusing on MARS, Logistic Regression, CART, Random Forest and Neural Networks to classify the response of the prospects. The team has defined “no” response as negative and the “yes” response as positive.

In order to improve their conversion rate, telemarketers should target prospects who are most likely to accept the campaign. The team evaluates the effectiveness of the model on 2 criterias. Firstly, the model should have a high sensitivity (true positive ratio), ensuring the model can identify prospects who are most likely to say yes. Furthermore, the model should also classify prospects who are not likely to say yes accurately. Failure to do so would lead to telemarketers channeling valuable resources to target these prospects because they have been wrongly classified as potential prospects. Therefore, the second criteria is to have a high accuracy is necessary for the model as well.

Sensitivity is the determining metric for this case because predicting an accurate percentage of those who convert is critical to a telemarketers’ success. The prediction of those who will say yes is more important than the prediction of who will say no. Thus, the team will put more importance on sensitivity than on accuracy.

The team has split the data set into 70% train sets and 30% testset. Different models will be trained on different train data based on its ability to handle missing values. But in order to compare the results of different models, these models will be only tested on the same testset.

## Customer Segmentation

The data has currently segmented the prospects based on job, marital status, education and whether they have been targeted before. Details such as campaign, duration, month and day can only be known after the customer has been contacted. Therefore the team thought it will be beneficial for the company to have a more generalised segmentation based on the details that they have prior to their first call. Therefore the team performed K-means Clustering with age and balance\_log. By calculating the total within the sum of squares the team concluded that the optimal number of clusters is 3 (Refer to Appendix Fig. 32).

The team also wanted to see the effect of these cluster ID on the response variable. It seems that prospects who belong in cluster 2 tend to say yes, while prospects who belong in cluster 3 tend to say no (Refer to Appendix Fig. 33). The cluster ID of the different prospects are also fed into the models to provide more information.

For visualisation of the clustering, refer to Appendix Fig. 34.

### Modelling

### Logistic Regression

The team built a generic logistic regression model on the train set that has no missing values. The model selection was automated using the step function which removes the variable that has the highest Akaike Information Criteria (AIC). AIC deals with overfitting and underfitting issues by ensuring that the trade-off between goodness of fit and the simplicity of the model is addressed. It does so by adding a penalty for the complexity of the model. The resulting variables are the variables that reduced the overall AIC of the model and have some statistical significance. AIC also removes multicollinearity between independent variables.

The response variable of the train set was predicted using a standard threshold of 0.5. If the probability is greater than 0.5, the test data is classified as “yes”. Otherwise, it would be classified as “no''. The results are shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted No | Predicted Yes | Total |
| Actual No | 1487 | 275 | 1762 |
| Actual Yes | 280 | 1307 | 1587 |
| Total | 1767 | 1582 | 3349 |

The model predicted 2794 out of the 3349 data points correctly, which makes it 83.4% accurate. Of the results, 82.4 % of positive cases predicted accurately and 84.4% of negative cases predicted accurately.

The team went on to experiment to see if a threshold of 0.5 is suitable to classify the predicted probability of the response variable. To pick out the appropriate threshold, the team built a CART model based on the predicted probability that was derived from the logistic model to predict the response variable. After pruning the model, the optimal threshold that was identified was 0.47. If the probability is less than 0.47, the outcome should be “no”. The results of the logistic regression together with a threshold of 0.47 is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted No | Predicted Yes | Total |
| Actual No | 1465 | 297 | 1762 |
| Actual Yes | 249 | 1338 | 1587 |
| Total | 1714 | 1635 | 3349 |

The accuracy improved by 0.3% to an accuracy of 83.7%. The sensitivity increased by 1.9% to 84.3%.

Therefore, by using the optimal threshold of 0.47 the accuracy of the model increased and the sensitivity of the model is increased more than proportional. Therefore, the logistic model with the optimal threshold that was found from CART is a better model to be for comparison.

### MARS

Another model that the team implemented is Multivariate adaptive regression spline (MARS). Since MARS is unable to handle missing values, it was trained on the train set that does not have any missing values. The team built the MARS model using a logistics regression as specified in earth documentation. The output of this model was similar to logistic in the sense it predicted the probability of the response variable. Using a standard threshold of 0.5 the confusion matrix of the model is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted No | Predicted Yes | Total |
| Actual No | 1494 | 268 | 1762 |
| Actual Yes | 274 | 1313 | 1587 |
| Total | 1768 | 1581 | 3349 |

The accuracy of the model was determined to be 83.8%, while the sensitivity is 82.7% and specificity is 84.8%.

The team wanted to further improve the accuracy of the model. The team used the train method found in caret package to train the MARS model using 10 fold cross validation to improve the Accuracy of the model. The optimal degree for the MARS was 3 and the nprune was 45. (Refer to Appendix Fig. 35) Similar to the logistic model, the team built a CART model based on the predicted probability to predict the deposit values. It was observed that the threshold of 0.39 is a more suitable than 0.5. The accuracy of the model increased to 85.7%, the sensitivity increased to 90.1% and the specificity decreased to 80.1%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted No | Predicted Yes | Total |
| Actual No | 1427 | 335 | 1762 |
| Actual Yes | 143 | 1444 | 1587 |
| Total | 1427 | 1922 | 3349 |

### 

### CART

Since CART is able to handle missing values using surrogates, the team removed the values that were initially replaced with the mode of that column that were set to just NAs (refer to section 1.2) and we also removed additional columns such as education\_unknown, job\_unknown, contact\_unknown. The tree was grown to its maximum by allowing the initial cp to be 0. Later on the tree was pruned using an optimal cp value of 0.002362622. After which the team used the pruned tree to predict the outcome of the response variable in the testset. The accuracy is 83.7% and the sensitivity is 86.5%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted No | Predicted Yes | Total |
| Actual No | 1431 | 331 | 1762 |
| Actual Yes | 214 | 1373 | 1587 |
| Total | 1645 | 1704 | 3349 |

### Random Forest

Random forest is able to handle missing values in 2 ways, one of them is to fill the missing values based on the mode of the column and the other way is to predict using the rfImpute function. The team worked on 2 Random forest, one was trained on the dataset the team created. The other random forest was trained on the dataset that was generated from the rfImpute function. Comparing the model against the testset, the model that was trained using the dataset that the team created had a higher accuracy and sensitivity of 85.7% and 89.4% respectively.

(with rfimpute)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted No | Predicted Yes | Total |
| Actual No | 1442 | 320 | 1762 |
| Actual Yes | 184 | 1403 | 1587 |
| Total | 1622 | 1727 | 3349 |

(with actual data)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted No | Predicted Yes | Total |
| Actual No | 1453 | 309 | 1762 |
| Actual Yes | 169 | 1418 | 1587 |
| Total | 1622 | 1727 | 3349 |

Therefore the next few analyses on the random forest will be based on the random forest that is trained on the original dataset. The team built the random forest with the default number of trees which is 500 trees. It seems that 500 is an optimal number of trees since the error tends to be stable after 200 trees (Refer to Appendix Fig. 36). The team wanted to find the optimal number of variables to be selected in each split of the random forest. By using the tuneRF function and a mtry of 10, the out of bag error is the lowest. Using this newly found mtry as the base, the team tried to tune the model using 10 fold cross validation to boost the accuracy of the model with the caret package. It was found that the optimal mtry was 28 (Refer to Appendix Fig. 37). This model has accuracy of 85.8 % and sensitivity of 88.7 %. The confusion matrix of the model is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted No | Predicted Yes | Total |
| Actual No | 1442 | 320 | 1762 |
| Actual Yes | 184 | 1403 | 1587 |
| Total | 1626 | 1723 | 3349 |

### Neural Network

The last model that the team worked on is neural network. Since this is a classification problem, the team decided to use the nnet package instead of neuralnet. Nnet only has one hidden layer and the number of nodes specified in that layer can be changed. nnet configures the most optimized model by automatically assigning the decay of the weights ( from 0.1 to 0.5 ) that would return the highest train accuracy (Refer to Appendix Fig. 38). Furthermore, this model is unable to handle missing values therefore we will be training this model with the train set that does not have any missing values.

The team used 10 fold cross validation (using caret) to train the neural network with and without normalising the numeric values in the train set. For the nnet that was trained on non normalised data the optimal weight decay of 0.5 and 2 nodes in the hidden layer while the nnet that was trained on normalised data needed 7 nodes in the hidden layer with a weight decay of 0.5. From the results, it seems that working with a neural network that does not have normalised data seems to be better with a higher accuracy of 81.8% and higher sensitivity of 84.3%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted No | Predicted Yes | Total |
| Actual No | 1402 | 360 | 1762 |
| Actual Yes | 249 | 1338 | 1587 |
| Total | 1620 | 1729 | 3349 |

The results shown with normalisation, accuracy is 53.5% and sensitivity is 81.2%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted No | Predicted Yes | Total |
| Actual No | 503 | 1259 | 1762 |
| Actual Yes | 297 | 1290 | 1587 |
| Total | 1131 | 2218 | 3349 |

Therefore, the model without normalisation is better in terms of accuracy and sensitivity.

The team thought that the presence of too many variables could result in redundant noise and decrease its accuracy. Therefore we selected the top 4 variables that had a higher mean Decrease Gini than the rest of the variables. The accuracy is 76% and sensitivity is 78.4%. If the variables are normalised, the accuracy is worse at 47.4%. Thus, selecting the top 4 variables with the highest mean Decrease Gini does not improve the model.

Even after normalising the numerical data of these 4 variables, the model only had an accuracy of 47.4% and a 100% sensitivity. The neural network predicts every customer as a potential customer which is similar to the telemarketer’s current marketing strategy. This model would not be a suitable model to propose since it does not add value to their current strategy.

## Evaluation

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy (%) | Sensitivity (%) | Specificity (%) |
| Logistic Regression | 83.7 | 84.3 | **83.1** |
| Tuned MARS | 85.7 | **90.1** | 80.1 |
| CART | 83.7 | 86.5 | 81.2 |
| Random Forest | **85.7** | 89.4 | 82.5 |
| Neural Network | 81.8 | 84.3 | 79.5 |

As seen from the table above, Random Forest followed by the tuned MARS had the highest accuracy among all the 5 models. However, the sensitivity of the tuned MARS model is higher than the Random Forest even though the accuracy of the model is slightly lesser than that of the random forest.

As discussed in section 2.2, the team will recommend a model based on sensitivity followed by accuracy. Tuned MARS has relatively good accuracy, albeit not the best. The higher sensitivity is worth the trade off. Therefore, the tuned MARS model is recommended.

# Recommendations

Since the tuned MARS model returns the highest value for sensitivity, our recommendations will mainly revolve around the results generated from this model. The respective hinges can be found in Appendix Fig. 39. If the hinge functions were above or below specific levels, this will affect the odds of success (deposit = yes) for each of the relevant variables, respectively.

# 

## Target based on seasonality

The first recommendation to improve the conversion rate is by targeting their calls based on the different months. The team found that the months which tend to have the highest customer conversion rates were February (OR of 1.11), March (OR of 1.57), April (OR of 1.14), June (OR of 1.14), September (OR of 1.26), October (OR of 1.28) and December (OR of 1.38) (Refer to Appendix Fig. 40) . This suggests that these firms should aim to churn out as many calls as possible during these months compared to other periods within the year. Additionally, telemarketing firms can also try to save labour costs, as they can also tailor their frequency of hiring according to the months that are shown to have higher demand from prospects. This is reasonable as telemarketing firms would mostly hire part-time or on demand workers as telemarketers.

## Target based on loan status & age group

It was found that those with no housing and personal loans tend to say yes. Given that this marketing campaign is for a term deposit account, which is an investment, it makes sense that prospects with more financial freedom, i.e. no loans, would subscribe. Additionally, based on the dataset, it was found that students (18 to 31 years old) and retirees (more than 53 years old) tend to say yes. This could be attributed to the fact that they do not have outstanding loans. Furthermore, the 18-31 years old age group have a higher odds ratio than the 53 years old and above age group. This may be due to the fact that the term deposit account is a long-term investment product which would appeal more to a younger age group. Therefore, telemarketing firms should target their calls on students and retirees with no loans.

## Target based on results of previous campaign

According to our analysis we have also found that prospects who tend to subscribe to such campaigns through telemarketing were actually previous prospects who have already shown interest in previous campaigns there were promoted by the company. This means that firms should try to focus more on their existing prospects when trying to market new campaigns. In addition to this, we have also discovered that the frequency of calls to these existing prospects should be kept at a minimum in order to increase the conversion rate of existing prospects to subscribe to the new campaign.

## Target based on expected duration to win a prospect

The balance per duration has to be greater than 1.555 for a prospect saying yes to have a higher odds ratio. Since duration is unknown before the call, we can derive the expected duration that a telemarketer should be on the call based on their current balance to convert a specific customer. Furthermore, telemarketers can prioritise leads based on the expected duration for a customer to be won. For example, the telemarketer may prioritise a prospect which has an expected duration of 10 minutes compared to another prospect with an expected duration of 30 minutes to increase the efficiency of the outbound telemarketing process.

## Hire skilled telemarketers and have a well-crafted script

Targeting prospects based on the above factors improves the pre-qualifying lead generation step of outbound telemarketing because the prospects being targeted are those who are more likely to subscribe to the term deposit, making the phone calls less intrusive. To further increase the likelihood of positive outcomes, it is imperative for telemarketers to have a well-crafted script and possess excellent telemarketing and persuasion skills in order to build rapport with prospects and influence the prospects to stay on the line long enough for the telemarketer to effectively sell the term deposit plan. It was found that the average phone call duration for those who eventually subscribed to the term deposit was 57 seconds. This means, it is extremely crucial for telemarketers to be skilled enough to grab the attention of the prospects and create desire for the product within the first 57 seconds of the call.

# Discussion

## Limitations of MARS model

### Model only considers the variables from dataset

Although the dataset includes various variables, in reality there can be many more possible variables that affect the conversion rate of prospects towards such campaigns. For example, as telemarketing requires communication between individuals, the ability of the telemarketers would also be an important factor in determining prospects conversion. Another variable that would affect the decision of prospects in subscribing to such campaigns could be attributed to the amount of other institutions that these prospects are already prospects of. This information would show firms the likelihood prospects would convert if they have already subscribed to other similar institutions.

### Duration of current call is taken into account

The model and analysis considered the duration of the current call that the telemarketer engages in. However, the duration of the current call is not known until the call is over and the prospect’s decision to make a deposit or not is known. Therefore, the actual duration of the current call cannot be used in the prediction of whether the prospect will make a deposit. Rather, the estimation of the duration can be used instead.

### Difficulty for non-data-analysts to interpret the model

The MARS model is not a widely known model like linear/logistic regression. Neither does it produce an intuitive chart like decision trees. Users who are not familiar with MARS might find it difficult to interpret and make use of the model.

### Usage of Automated Functions

Variables such as balance, balance per duration and age per day were normalised using automated functions such as Gaussianize. Some of the variable were normalised using an automated function -> this makes it harder to calculate the exact splitting points

## Applicability of MARS model

### Applicability to other industries

The MARS model does not apply to just the finance sector. It has strong potential to be applied in other industries such as real estate telemarketing. The MARS model can determine whether an owner is likely to be selling a property based on important variables such as value of property, age/occupation/salary of home-owner, location of property and even marital status, for a home-owner to sell his/her property. Another possible industry would be B2B telemarketing. Companies can use the MARS model to determine which other businesses to call based on factors such as revenue of the business (balance), number of loans the business have (general loans), industry the business is in (job). Other possible industries could be non-profits for fundraising purposes, utilities companies and telecommunications.

### Applicability to other types of marketing

Our model can also be applied to digital marketing to help companies better target their prospects for a new digital ad campaign. Possible factors to help determine the likelihood of a conversion include whether the user has clicked on an ad before (previous), number of clicks on previous ads (campaign), how long the user has stayed on the site (duration). Another possible area would be influencer marketing. Our model can help determine whether a follower would purchase an advertised product from a particular influencer based on factors such as the age/demographic of followers (age/job), number of times the follower viewed the influencer in the previous week (previous), number of times the follower viewed the influencer in the current week (campaign), how long does a follower view a post/video (duration).

# Conclusion

In conclusion, our team was concerned on increasing the efficiency and effectiveness of telemarketing through improving the conversion rate of outbound telemarketing campaigns. Through evaluating all the possible models, we are confident that the tuned MARS model will achieve the above mentioned by identifying important variables and critical points such as mean duration to improve the conversion rate of an outbound telemarketing campaign. However, we are aware of the limitations of the analysis such as the dataset being confined to only a set number of variables. This could have impacted the accuracy and sensitivity of the models, which may have affected the chosen model or even changed the models we used to understand our data. Nonetheless, from our analysis we believe with the current data and methodology our model could be utilized to optimize and transform the telemarketing industry for the better.

# References

# 

Akhrin, D. (2020). Council Post: A Case For Cold Calling In 2020. Retrieved 2 April 2021, from <https://www.forbes.com/sites/forbesbusinesscouncil/2020/02/26/a-case-for-cold-calling-in-2020/?sh=30fd3a625b60>

Business Wire. (2021). COVID-19 Impact and Recovery Analysis | Outbound Telemarketing Services Market Procurement Intelligence Report Forecasts Spend Growth of over USD 2 billion. Retrieved 2 April 2021, from <https://www.businesswire.com/news/home/20201127005160/en/COVID-19-Impact-and-Recovery-Analysis-Outbound-Telemarketing-Services-Market-Procurement-Intelligence-Report-Forecasts-Spend-Growth-of-over-USD-2-billion>

Convoso. (2020). 16 Essential KPIs for Lead Gen Call Center Profitability - Convoso. Retrieved 2 April 2021, from <https://www.convoso.com/blog/16-essential-kpis-for-lead-gen-call-center-profitability/#lead-conversion-rate>

Hayes, A. (2019). How Cold-Calling Works. Retrieved 2 April 2021, from <https://www.investopedia.com/terms/c/coldcalling.asp>

Kenton, W. (2020). Telemarketing. Retrieved 2 April 2021, from <https://www.investopedia.com/terms/t/telemarketing.asp>

Lake, R. (2021). Trust: An Advisor's Most Important Asset. Retrieved 2 April 2021, from <https://www.investopedia.com/financial-advisor/trust-advisors-most-important-asset/>

Lorette, K. Facts and Figures on Telemarketing. Retrieved 2 April 2021, from <https://smallbusiness.chron.com/figures-telemarketing-2206.html>

NI Business Info. Advantages and disadvantages of telemarketing | nibusinessinfo.co.uk. Retrieved 2 April 2021, from <https://www.nibusinessinfo.co.uk/content/advantages-and-disadvantages-telemarketing>

Riffkin, R. (2015). After Holidays, January U.S. Consumer Spending Falls to $81. Retrieved 2 April 2021, from <https://news.gallup.com/poll/181370/holidays-january-consumer-spending-falls.aspx>

# 

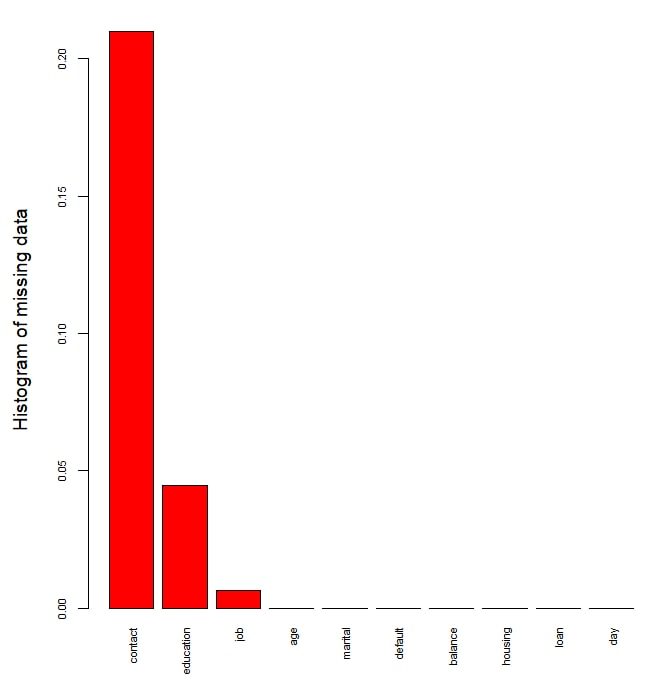
# Appendix

## Data Dictionary

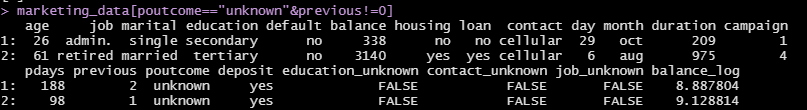
|  |  |
| --- | --- |
| Attribute | Description |
| age | age (numeric) |
| job | type of job (categorical: ‘admin', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown') |
| marital | marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed) |
| education | education (categorical: ‘primary’, ‘secondary’, ‘tertiary’) |
| default | has credit in default? (categorical: 'no','yes','unknown') |
| balance | bank balance (numeric) |
| housing | has housing loan? (categorical: 'no', 'yes', 'unknown') |
| loan | has personal loan? (categorical: 'no', 'yes', 'unknown') |
| contact | contact communication type (categorical: 'cellular', 'telephone', ‘unknown’) |
| day | last contact date of the month (numeric: 1 - 31) |
| month | last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec') |
| duration | last contact duration, in seconds (numeric) |
| campaign | number of contacts performed during this campaign and for this client (numeric, includes last contact) |
| pdays | number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) |
| previous | number of contacts performed before this campaign and for this client (numeric) |
| poutcome | outcome of the previous marketing campaign (categorical: 'failure', 'other', 'success', ‘unknown’) |
| deposit | has the client subscribed to a term deposit? (binary: 'yes','no') |

## Tables & Figures

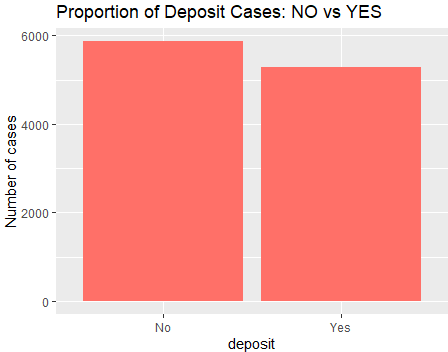
**Figure 1. Data Cleaning**



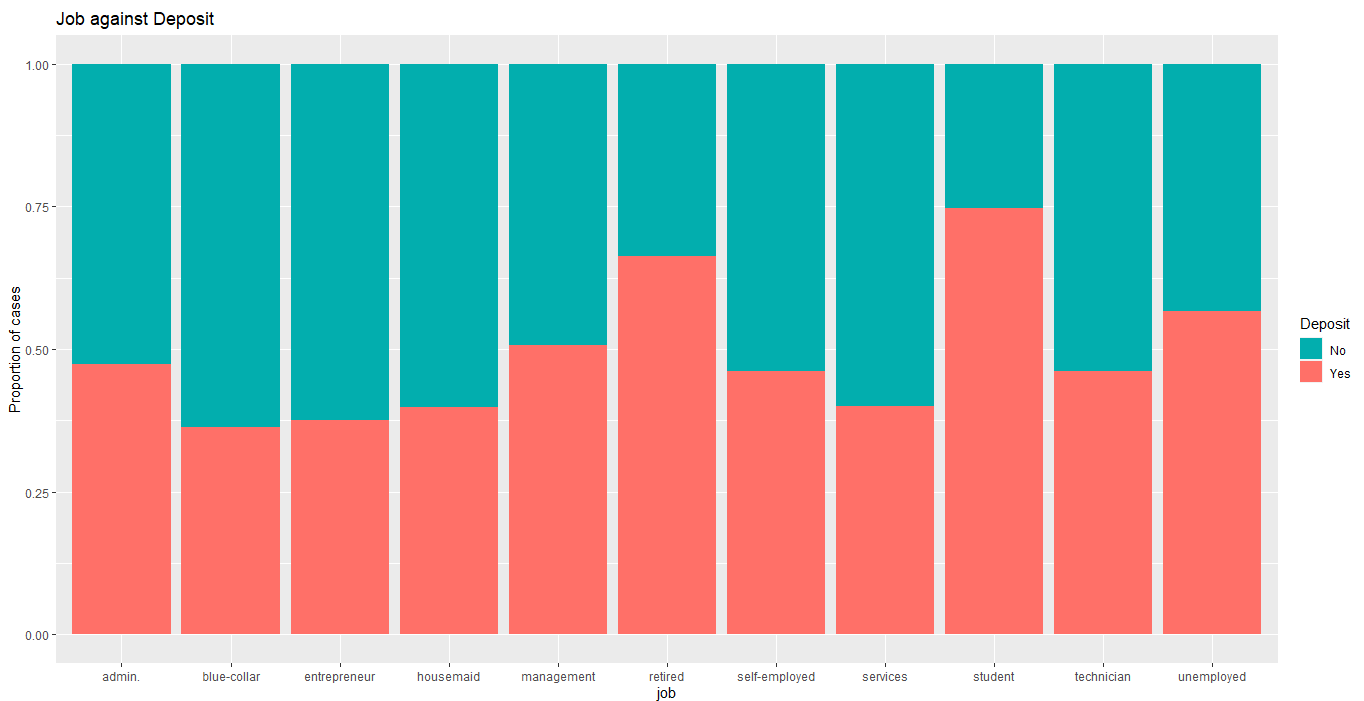
**Figure 2. Assumption - *poutcome***



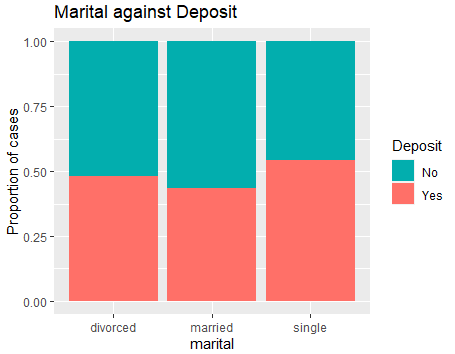
**Figure 3. Proportion of Deposit Cases**



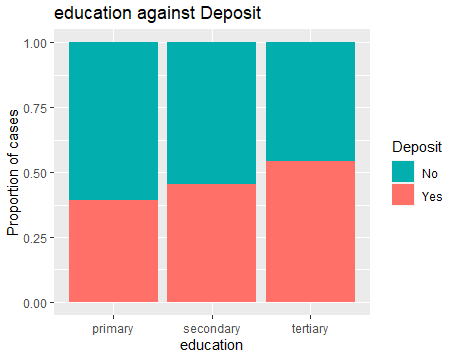
**Fig 4. Job Against Deposit**



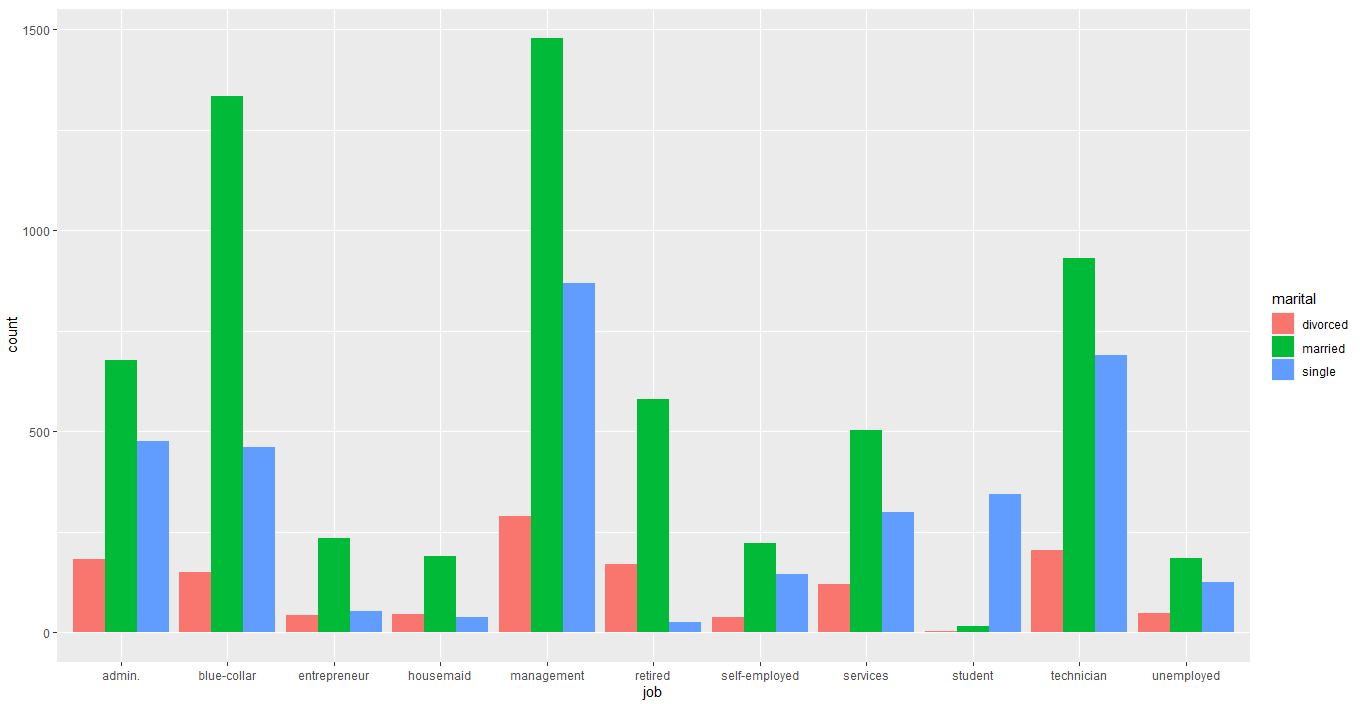
**Fig 5. Marital Status Against Deposit**

****

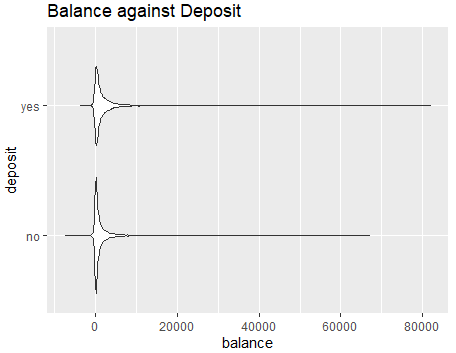
**Fig 6. Education Status Against Deposit**

****

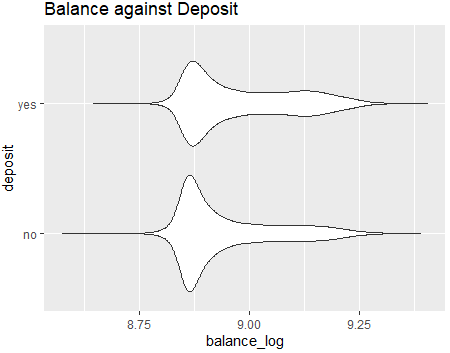
**Fig 7. Job Against Marital Status**

****

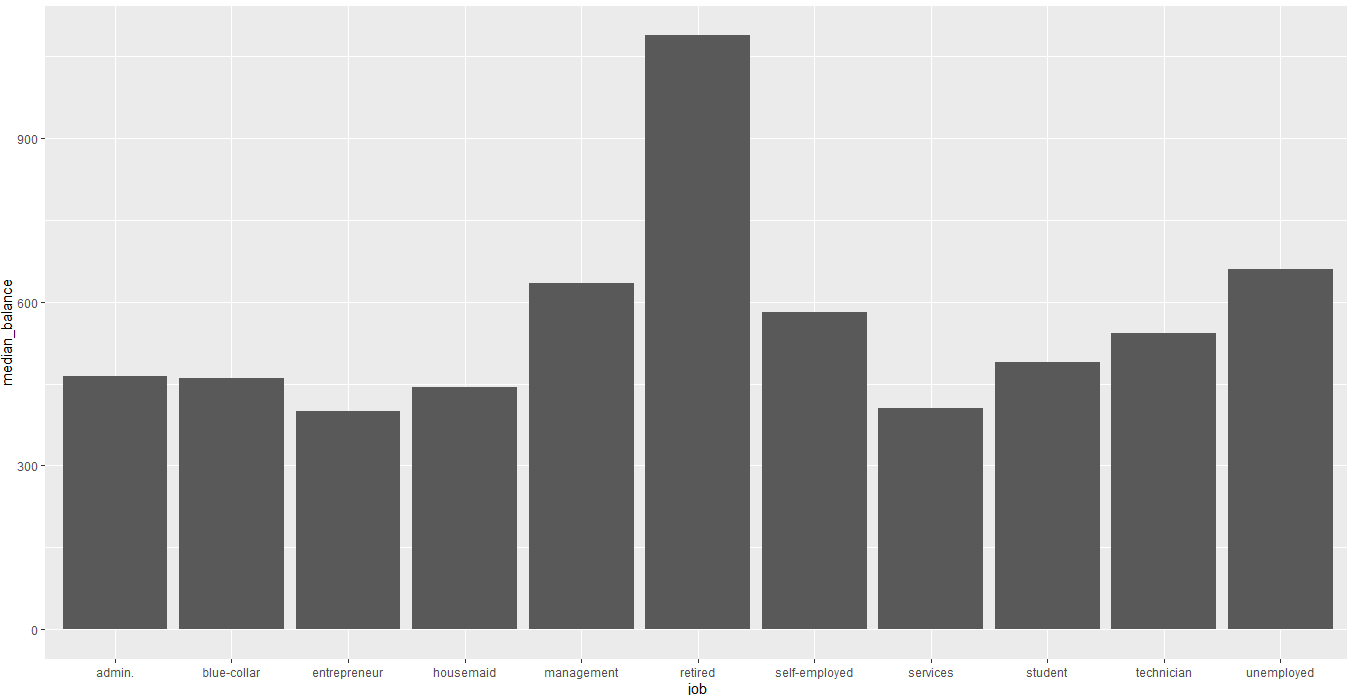
**Fig 8. Balance Against Deposit**



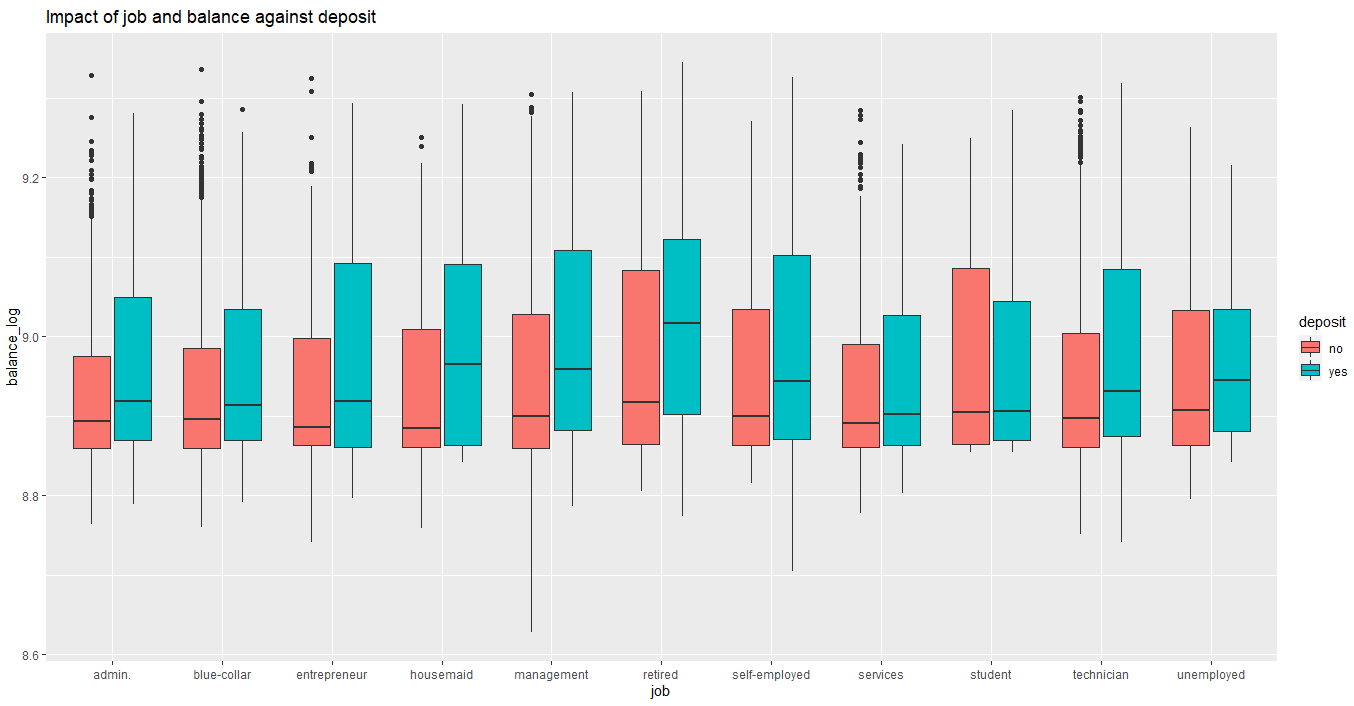
**Fig 9. Balance Against Deposit with Logarithm Function**



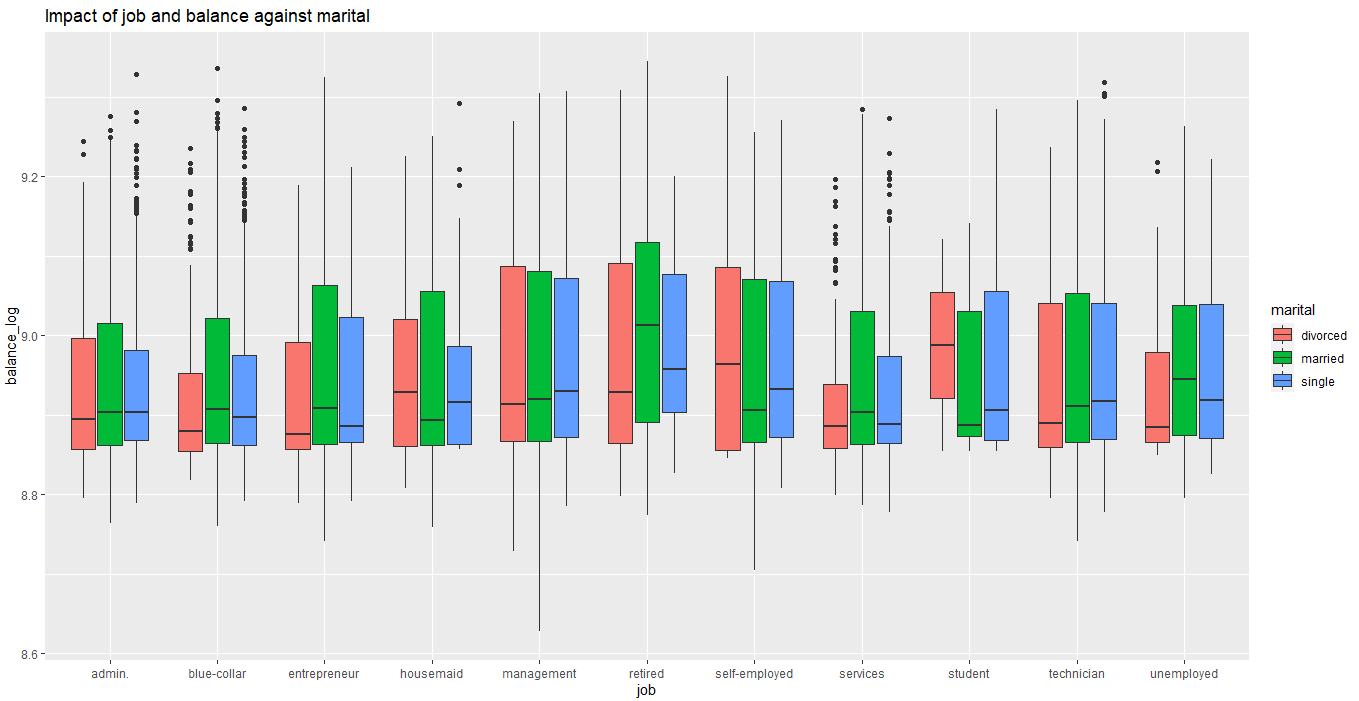
**Fig 10. Job Against Median Balance**

****

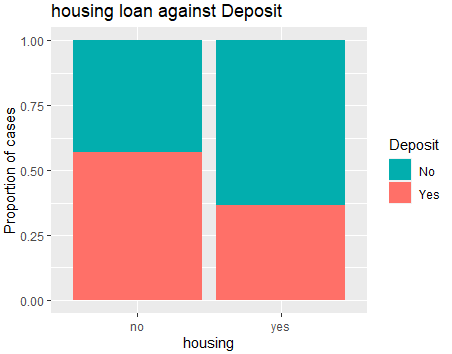
**Fig 11. Job Against balance\_log**



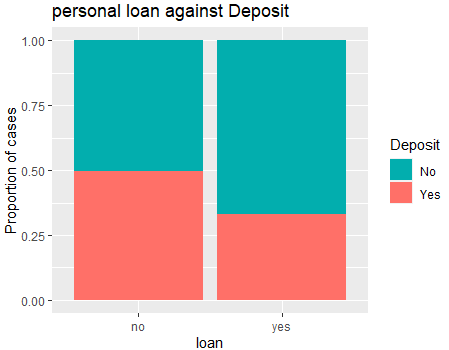
**Fig 12. Marital Status and Job against balance\_log**



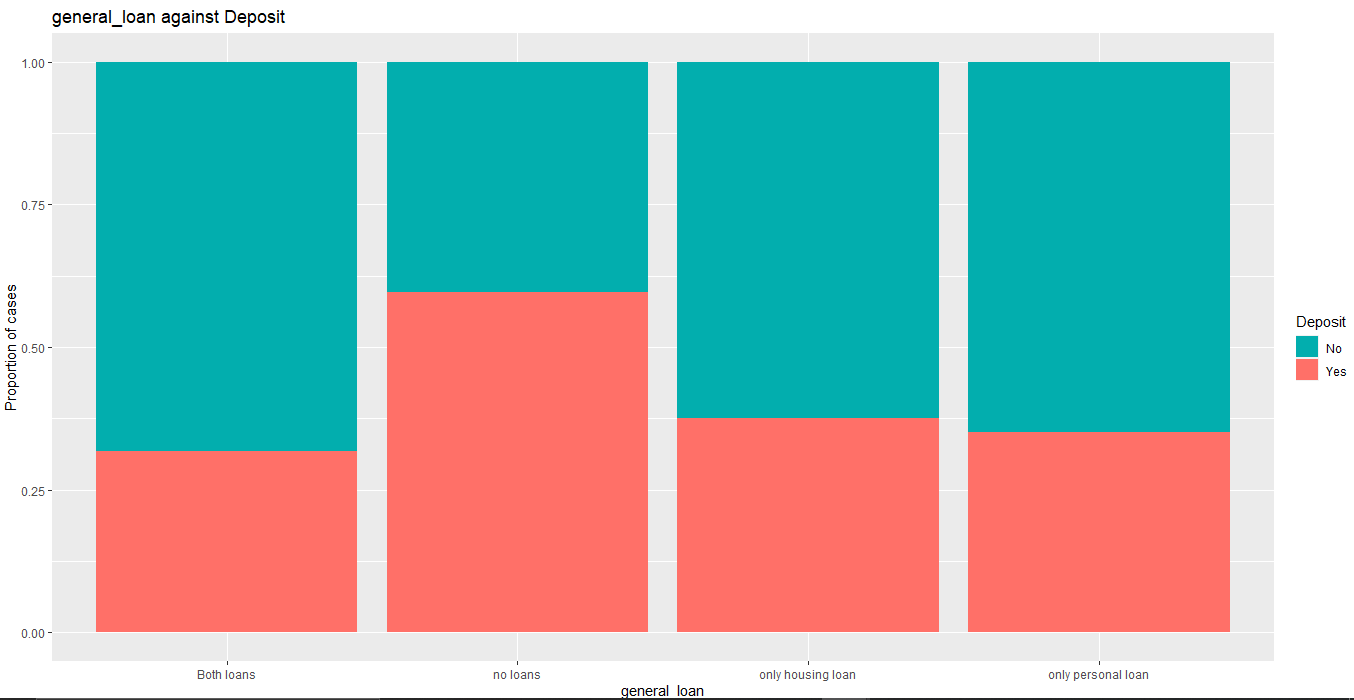
**Fig 13. Housing Loan against Deposit**



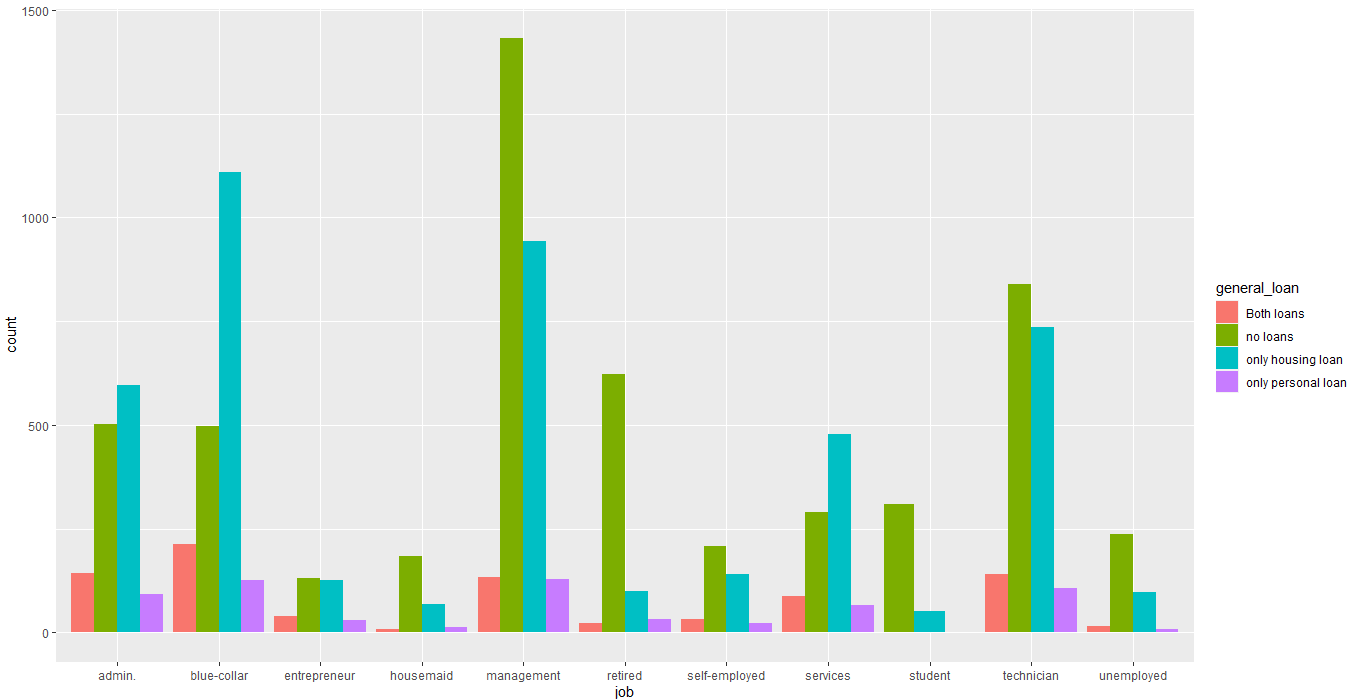
**Fig 14. Personal Loan against Deposit**



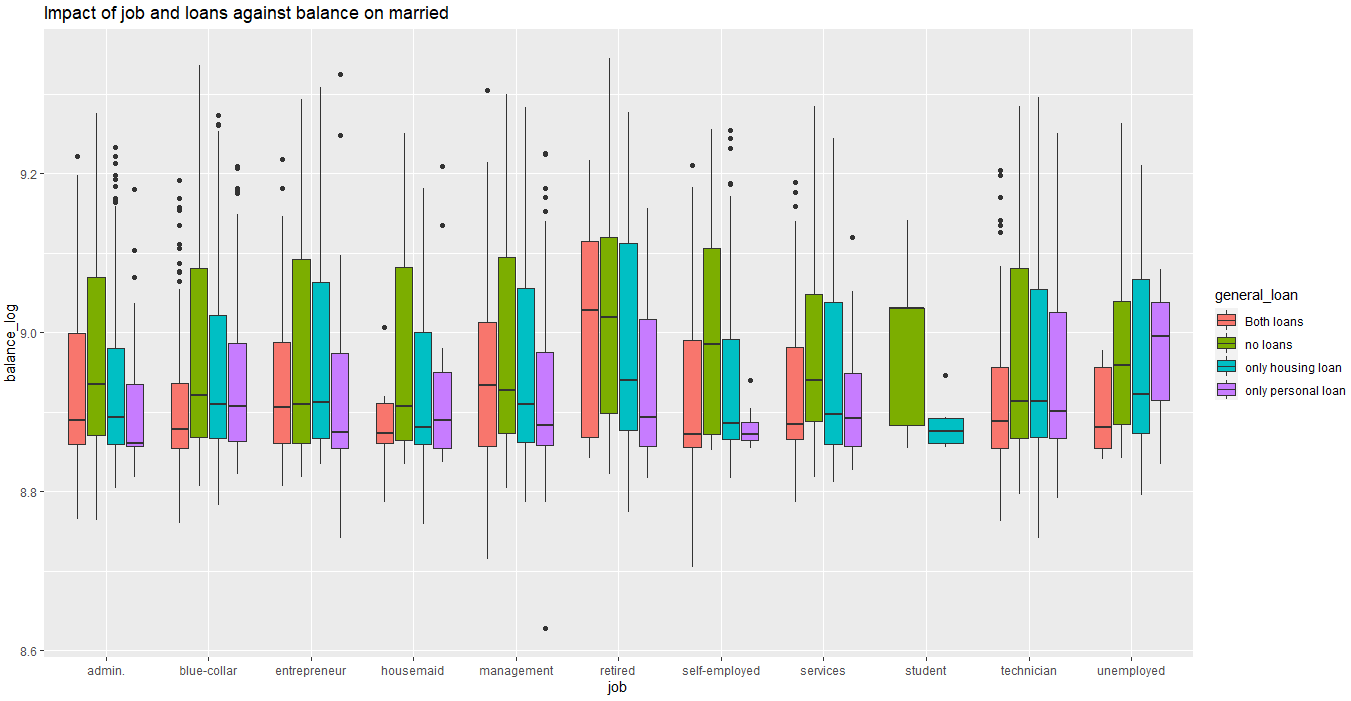
**Fig 15. General Loan against Deposit**



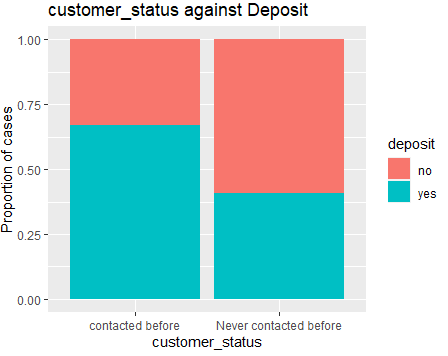
**Fig 16. Job against general\_loan**



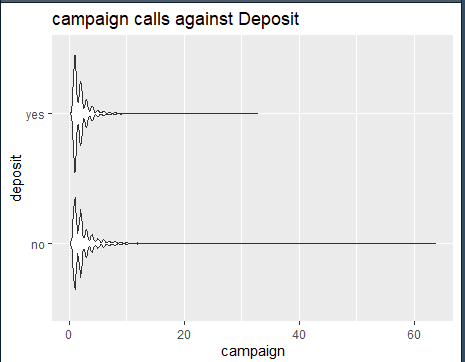
**Fig 17. Job with loans against Balance on married prospects**



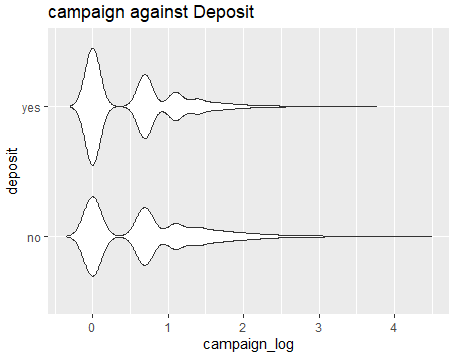
**Fig 18. Previous and p-days against Deposit**



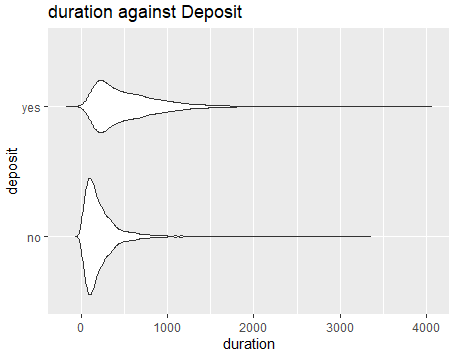
**Fig 19. Campaign against Deposit**



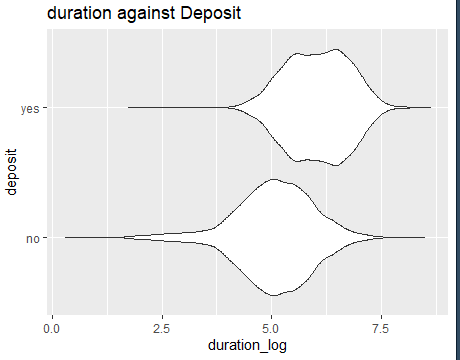
**Fig 20. Campaign against Deposit**



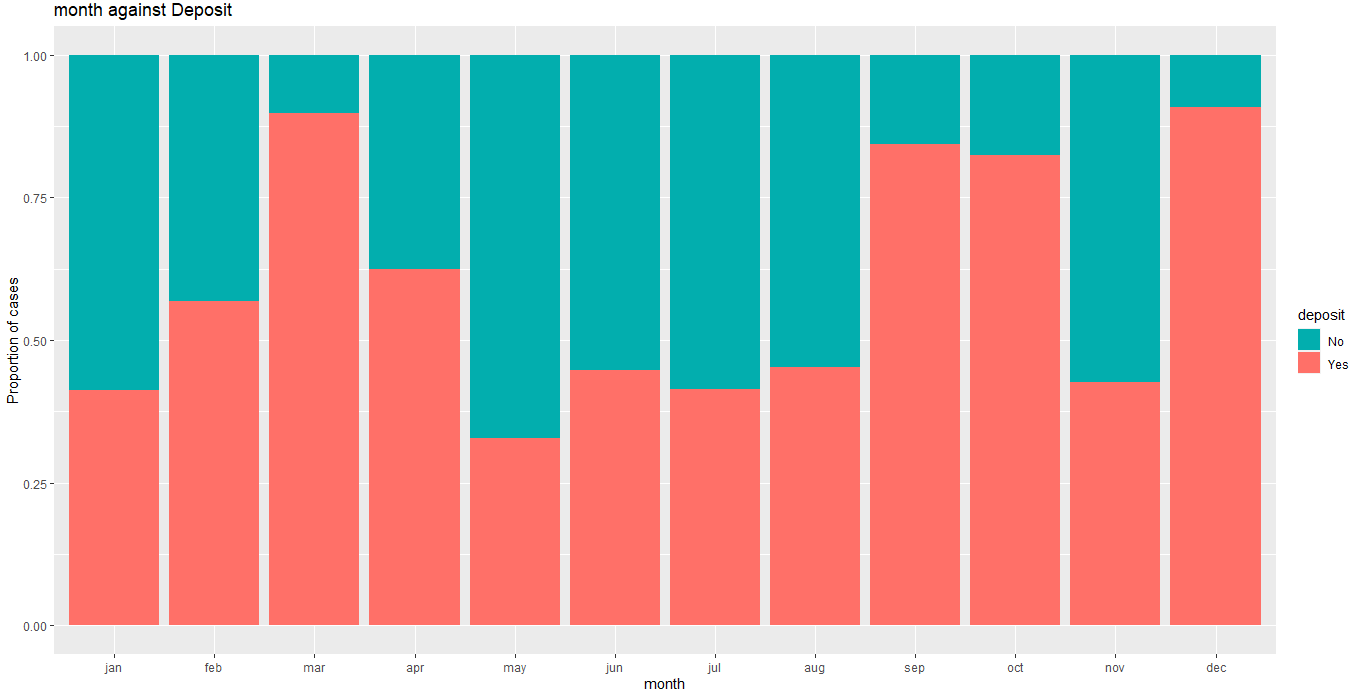
**Fig 21. Duration against Deposit**



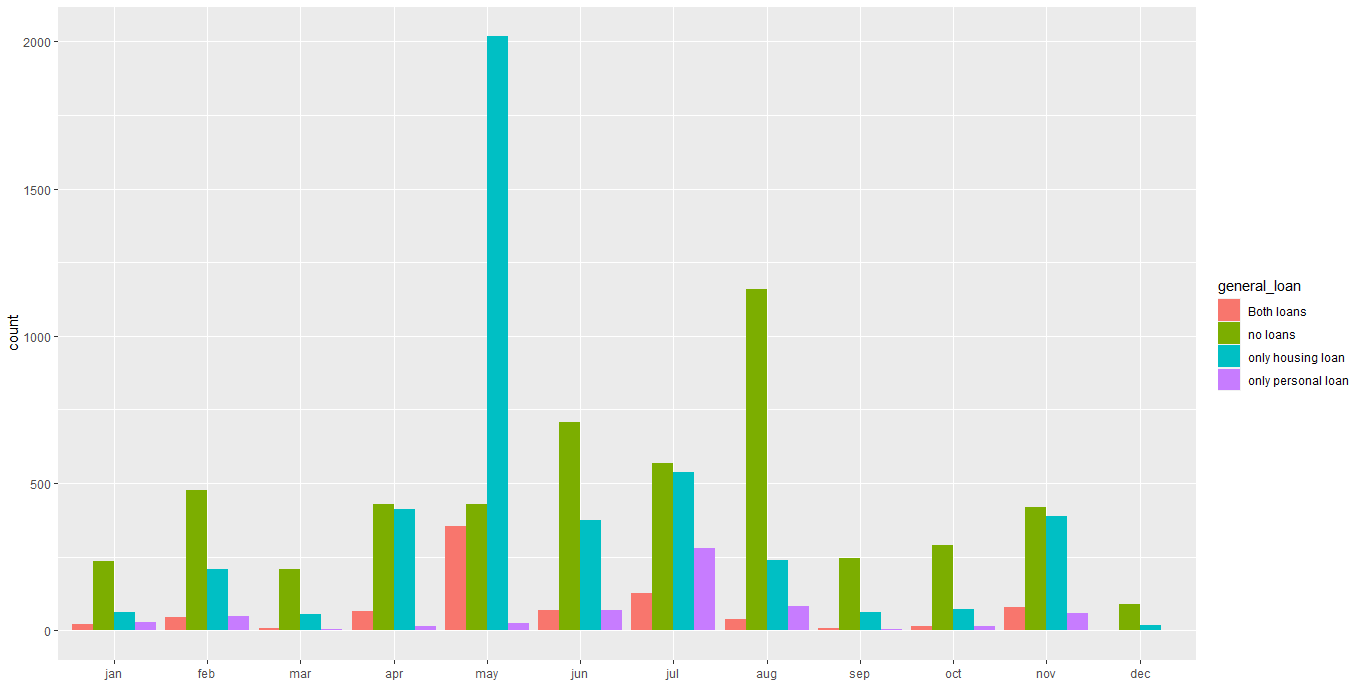
**Fig 22. Duration\_log against Deposit**



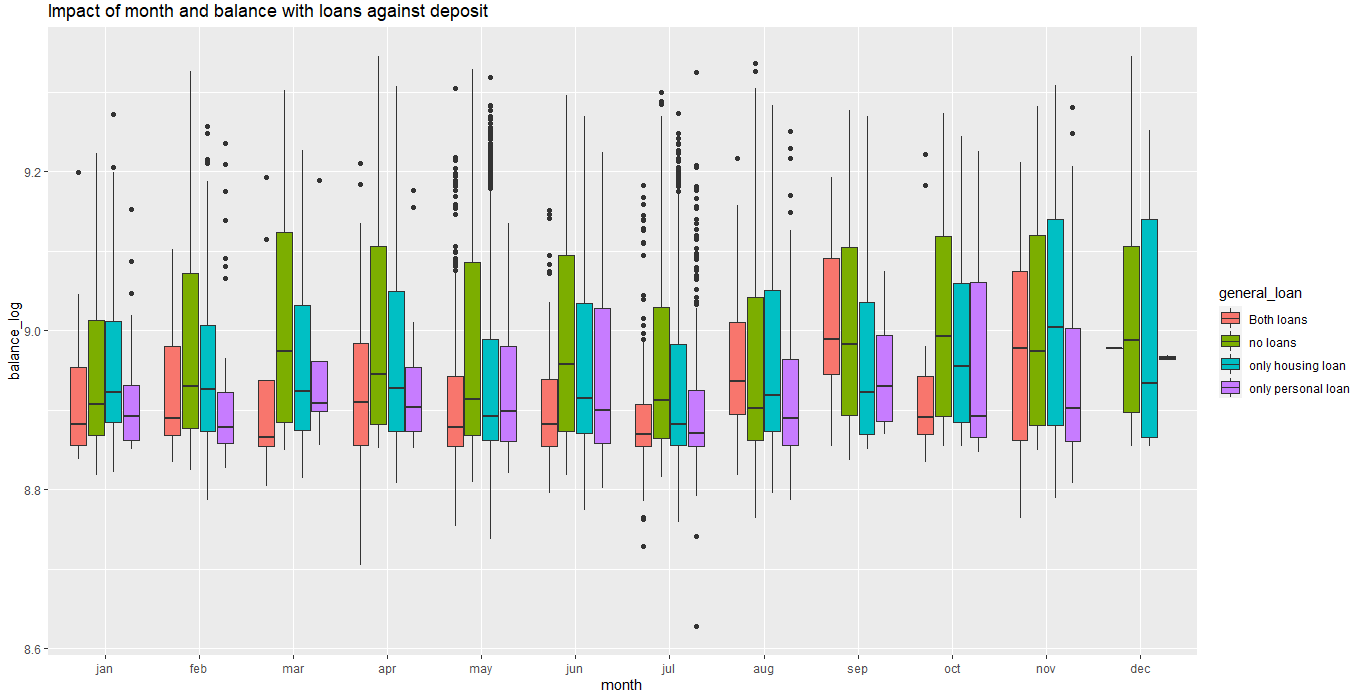
**Fig 23. Month against Deposit**



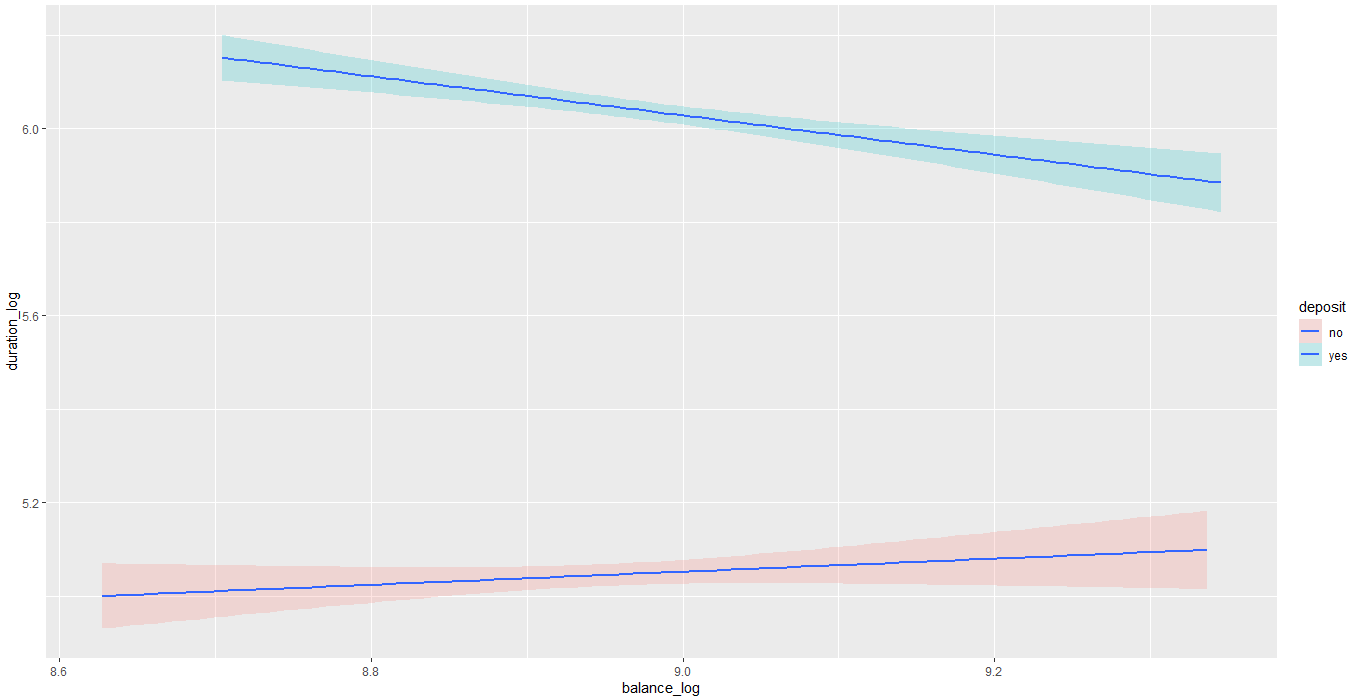
**Fig 24. Month against general\_loan**



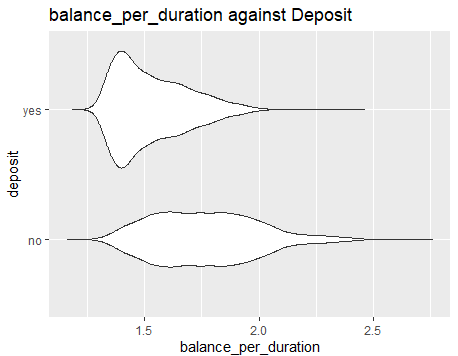
**Fig 25. Month with loans against Balance**



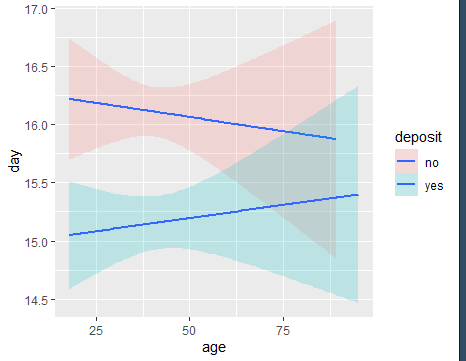
**Fig 26. Balance against Duration on Deposit**

****

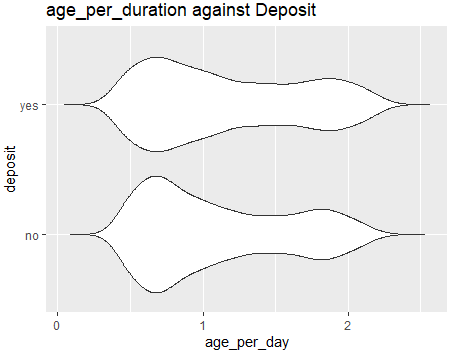
**Fig 27. Balance per duration against Deposit**



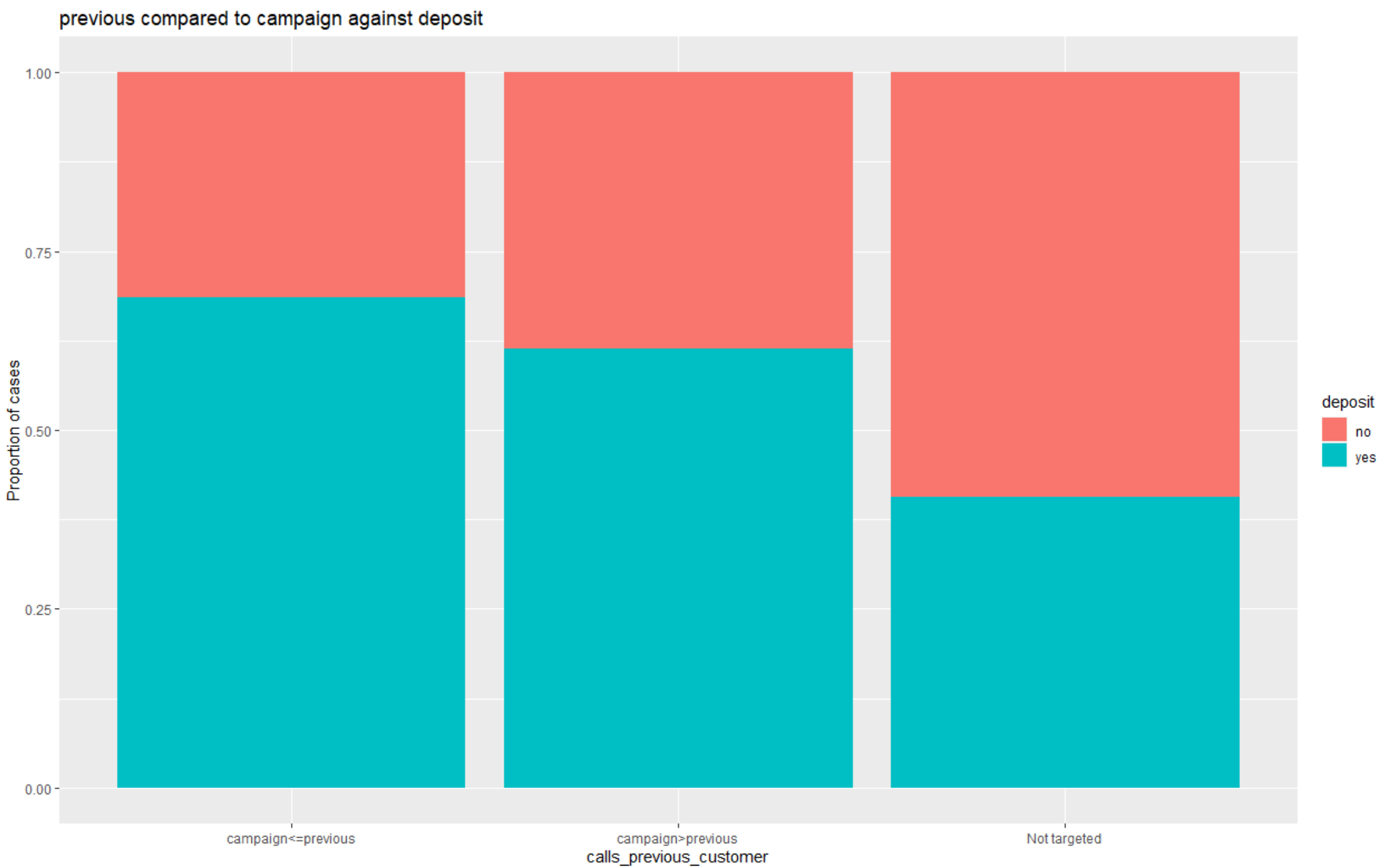
**Fig. 28. Age against Day**



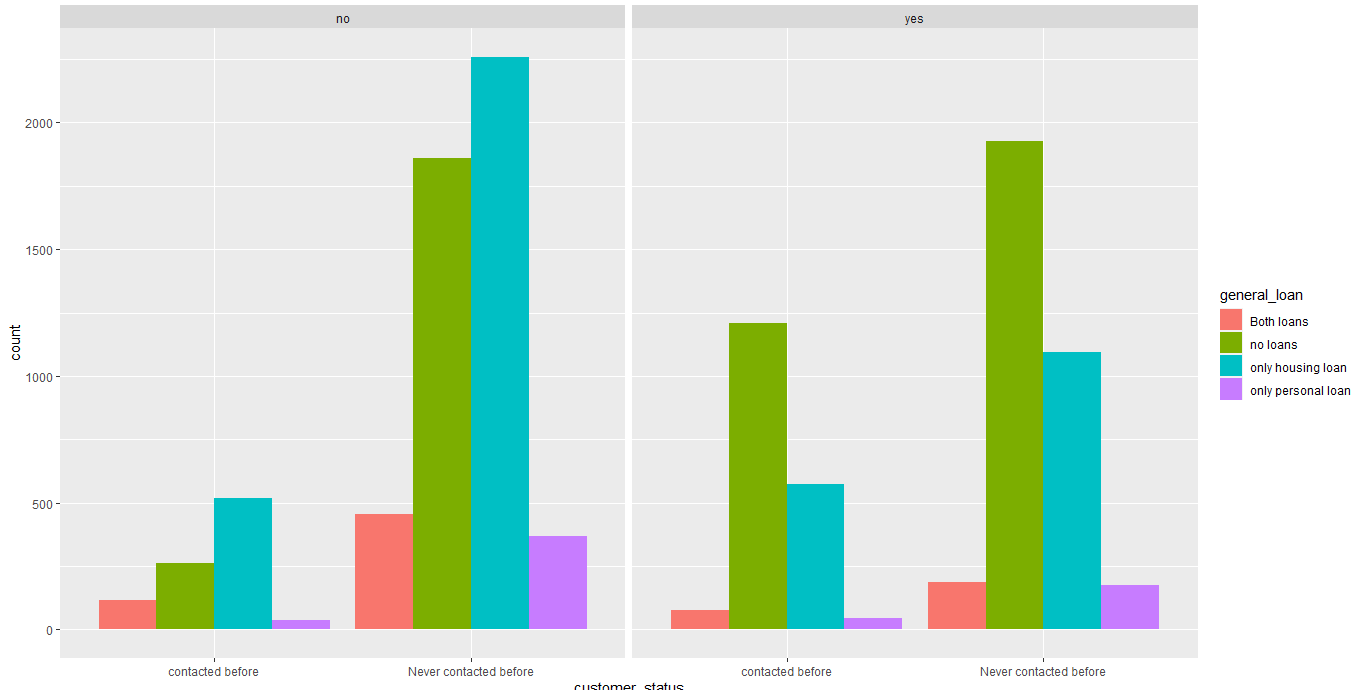
**Fig 29. Age per duration against Deposit**



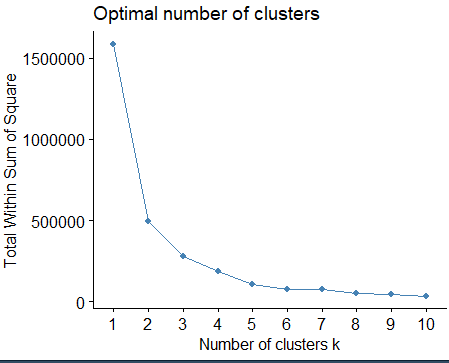
**Fig 30. Previous with Campaign on Deposit**



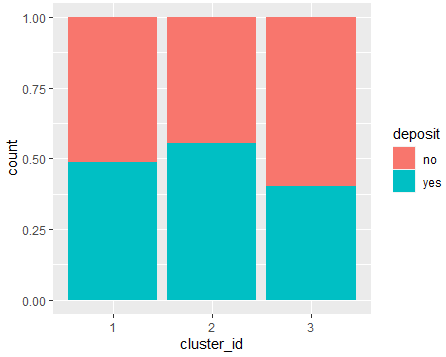
**Fig 31. Prospect Status against loans on Deposit**



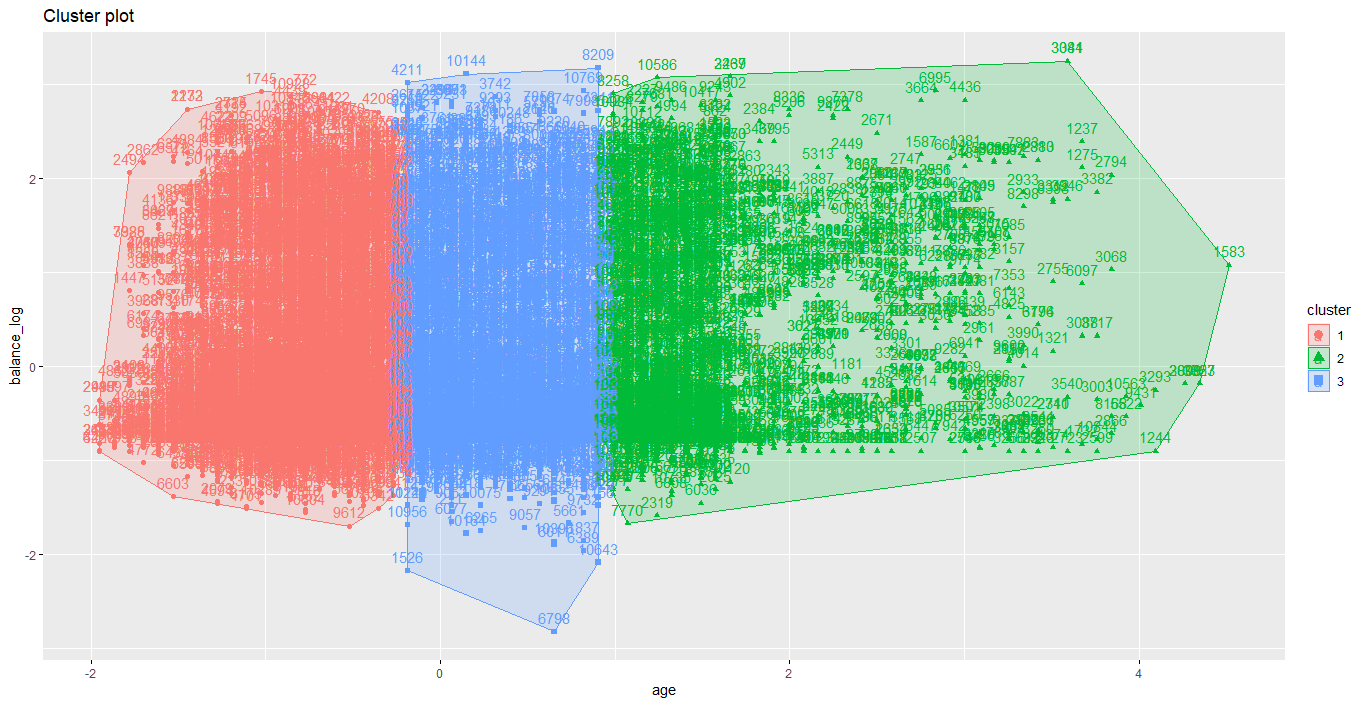
**Fig 32. Prospect Segmentation by Clustering**



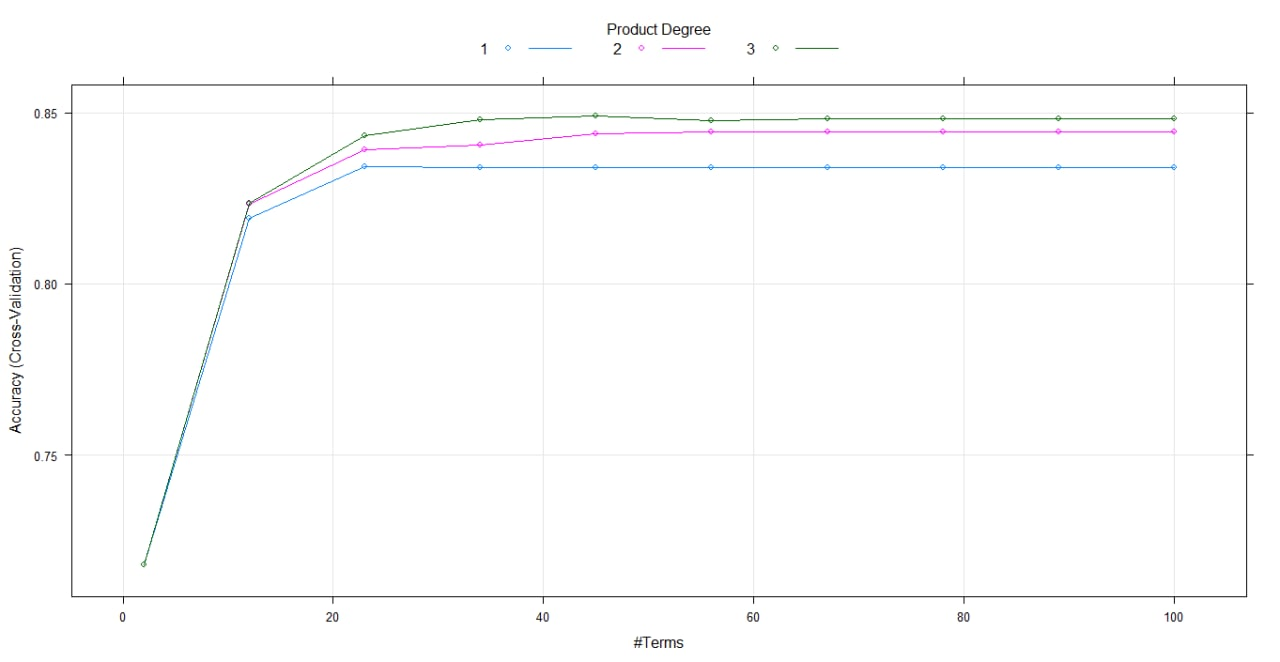
**Fig 33. Cluster ID on Deposit**



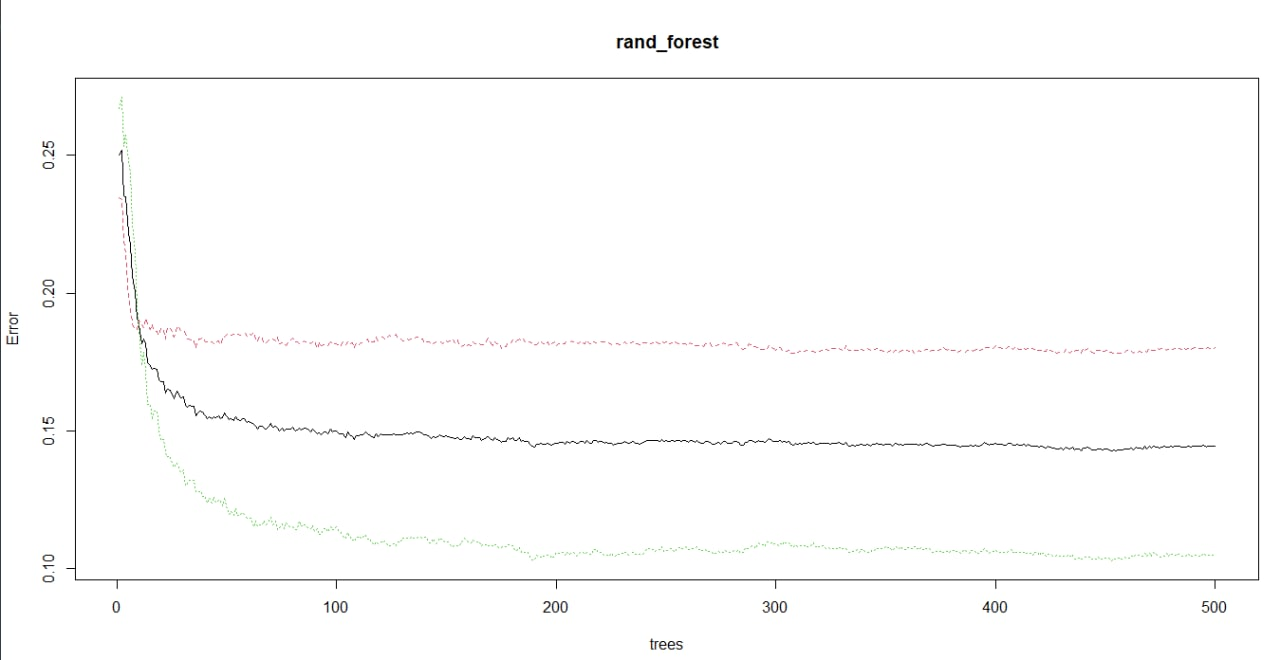
**Fig 34. Visualising Customer Segments**



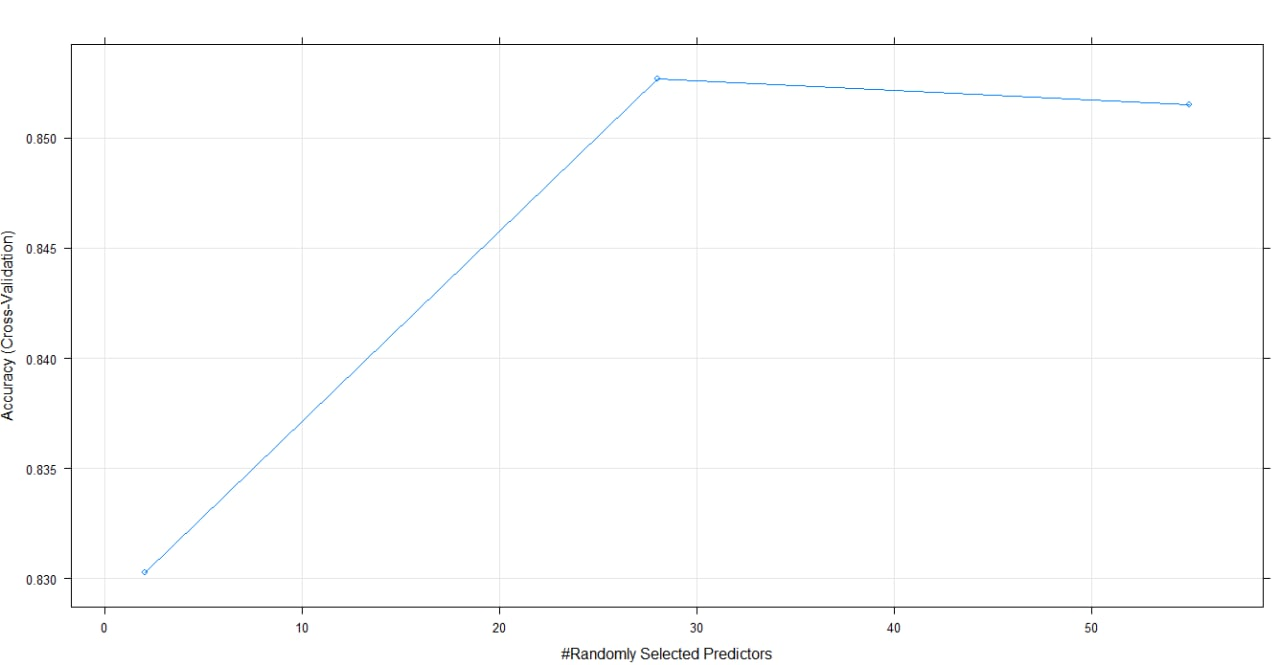
**Fig 35. MARS Model Evaluation (Degree =1,2,3)**

****

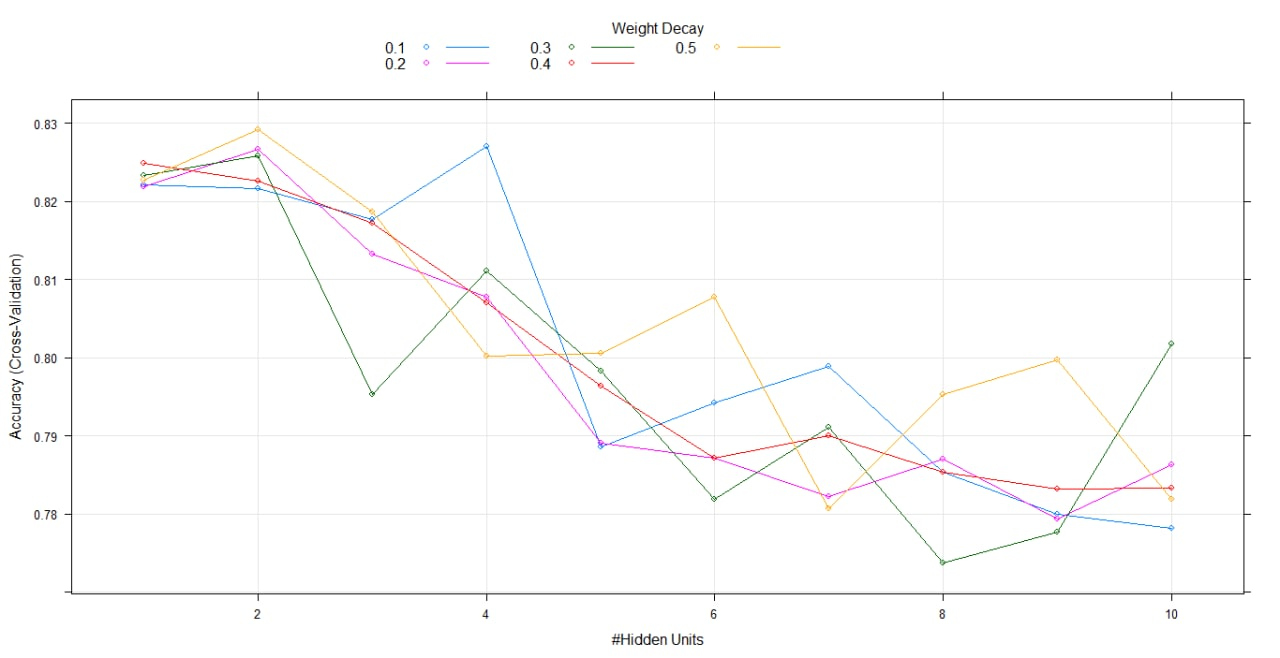
**Fig 36. Stabilization of errors for Random Forest Model**



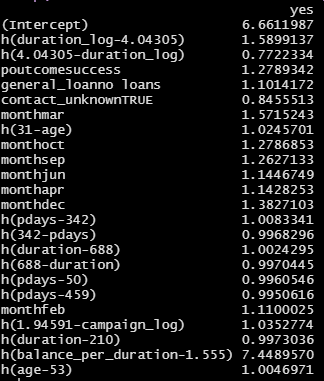
**Fig 37. Accuracy of Random Forest based on selected predictors**



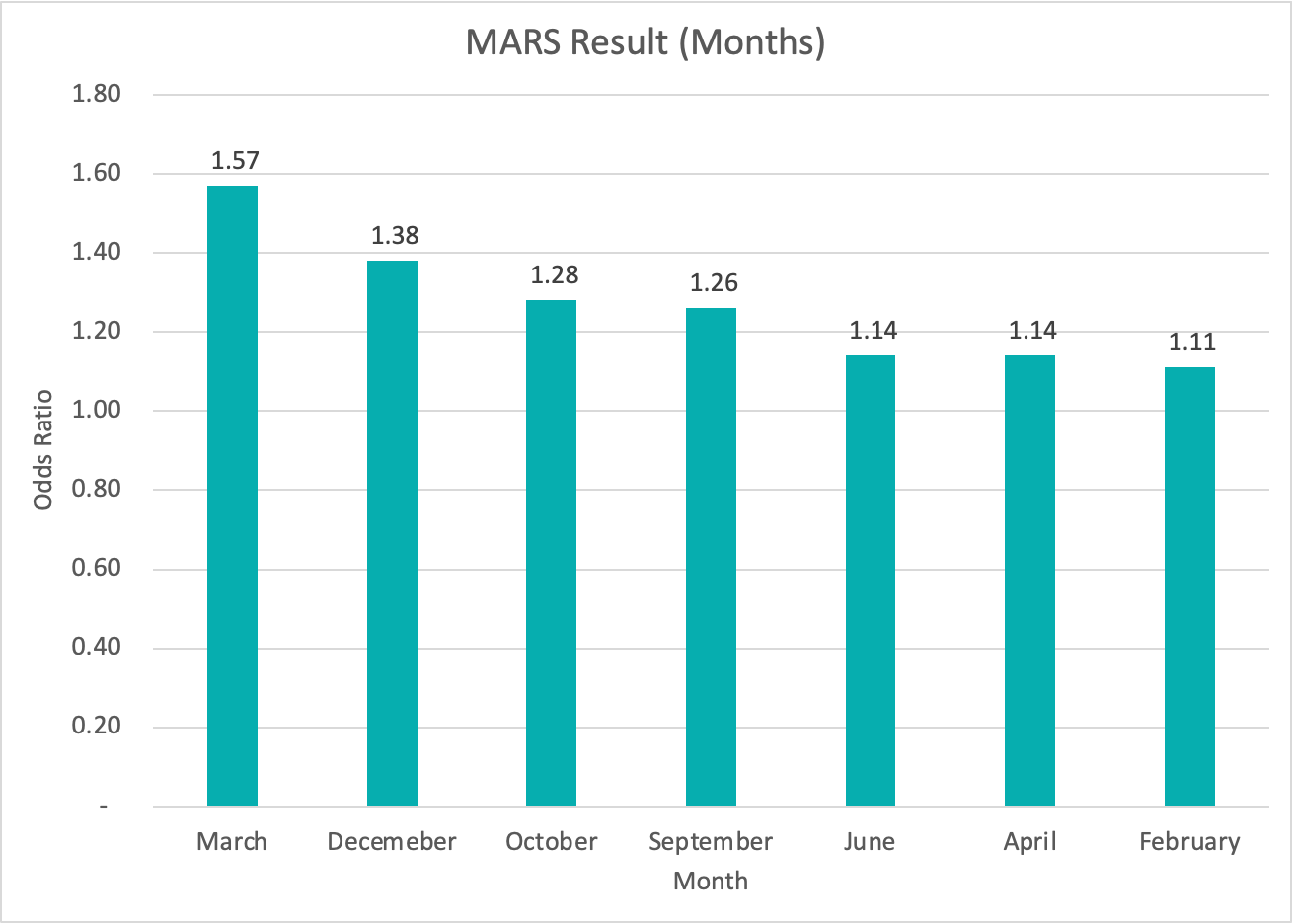
**Fig 38. Neural Network Weights Evaluation (Weight = 0.1 - 0.5)**

****

**Fig 39. Hinge for MARS Model**

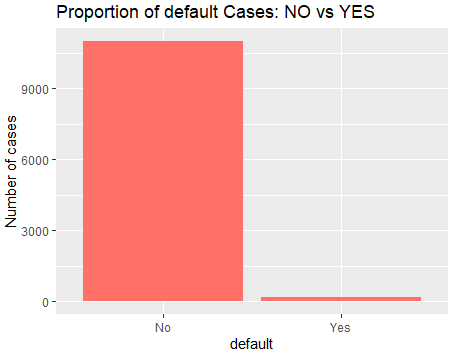


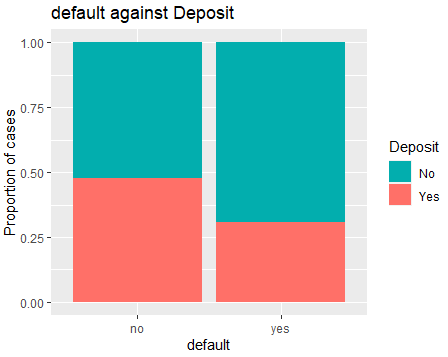
**Fig 40. Odds Ratio for Seasonality**

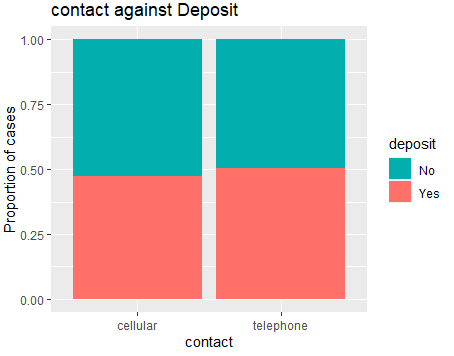
****

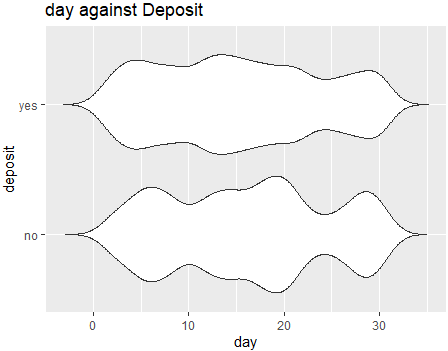
**Appendix B**

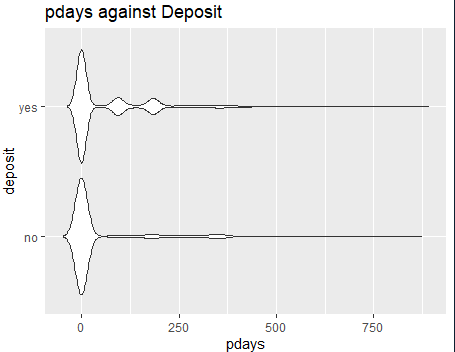












QQ plot of pdays

