Group assignment

**Problem**: Targeting the right customer to sell Term Deposit plan of a bank.

**Data**: Dataset contains information of about 45000 customers of a bank who have accepted or rejected past campaigns.

**Solution**: Build a classification model which will predict the outcome of the campaign held by the bank.

**Data Description**

Total records: 45211

The total data set has 45211 rows . 39922 clients have not opted for campaigns in the past.   
5289 clients have opted for campaigns in the past.

The data is highly skewed.

Our goal is to identify relations between different variables specified below and the outcome of the campaign ( customer buying / not buying ) .

The total variables are 17.   
Most of the variable are categorical variables.

Our focus would be to increase precision.

**Variables:**

1. Age: Age of the customer
2. Job: Type of Job (admin, blue-collar, entrepreneur, housemaid, employed, services, student, technician, unemployed, unknown )
3. Marital: ( Divorced, single, married )
4. Education: (Primary, secondary, tertiary)
5. Default: (Yes, No)
6. Balance: Average yearly balance of the client
7. Housing: Does client have housing loan ?
8. Loan: Does client have personal loan ?
9. Contact: Communication type ( Telephone, cellular, unknown)
10. Day: Last contact day
11. Month: Last contacted month of the year
12. Campaign: Number of contacts performed during last campaign
13. Pdays: number of days since the client was last contacted
14. Previous:: number of contacts before the last campaign
15. Poutcome: Outcome of previous campaign
16. Y: Target variable

**Initial EDA:**Trying to identify if there is any visible relationship between a few variables and the outcome (y)

Some linearity is seen in marital – if single or married  
If previous campaign outcome has been a success.

If bank has contacted the client in March, December, October or September.

There are some rows where contact is unknown.   
We determine to keep those rows, simply because we would lose around 500 rows which have “yes” for client’s acceptance of campaign.

Removing Loan and default since it is not adding any value to our dataset. Most of the value for “y” column are “No”

Applying K NN algorithm for the remaining co variates.

Calling a client repeatedly for a campaign is counter intuitive.   
Every call to a client is negatively impacting the future campaigns by 2702 times.

If a client has a housing loan, it is unlikely that they ll agree on term plan.

|  |  |
| --- | --- |
| A graph with numbers and lines  Description automatically generated | A graph with numbers and points  Description automatically generated |
| Fig(a) : K value vs Precision | Fig(b) : K value vs Accuracy |

Comparison of Fig(a) vs Fig(b) shows clearly that accuracy although increases when K value increases in KNN, precision decreases.

Best fit model is when k = 1 Since precision is our primary focus. Let us see the value of recall for K values.

A graph with dots and numbers

Description automatically generated

K values vs Recall.   
Recall increases as K value increases. However, given k as 1, recall is at 90%.

Using Lasso Logistic Regression to fit logit regression model on the data set.

Lasso doesn’t perform well on our skewed dataset. Precision is 0.

Hence, we discard the model and try regular logistic regression.

For logistic regression, we must now choose the cut off for our probabilities to get an acceptable precision and recall value.

|  |  |  |  |
| --- | --- | --- | --- |
| Cut off | Precision | Sensitivity | Specificity |
| 0.1 | 94.8% | 65% | 72.9% |
| 0.2 | 92.6% | 93% | 43.8% |
| 0.3 | 91.5% | 96% | 32% |

We observe, as our cut off increases, our precision decreases and specificity increases.

Let us observe important variables that determine the outcome “y” = 1

If a client is a student, there is 34.8% chance that clients accepts the campaign.

If the client’s education is tertiary, there is 27.11% chance.

If the client is contacted in the month of September or October or December there is a high chance.

Which clients to target and when?

Any client with secondary or tertiary educated unmarried to be contacted in the month of September or October or December

If a client has accepted previous campaign.

Any client who has been recently contacted

Any client who is unemployed or retired or a student.