X EDUCATION – LEAD SCORING CASE STUDY

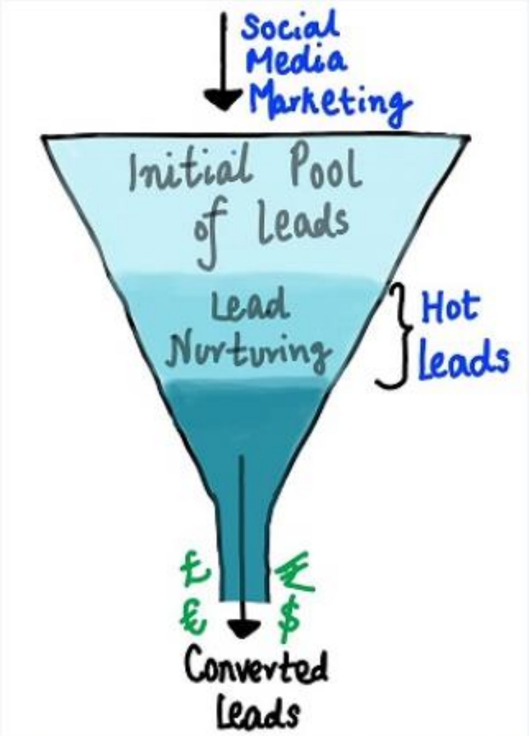
RUPESH CHACKO, YASH LUHARUKA

BACKGROUND

- X Education - An education company selling online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

PROBLEM STATEMENT

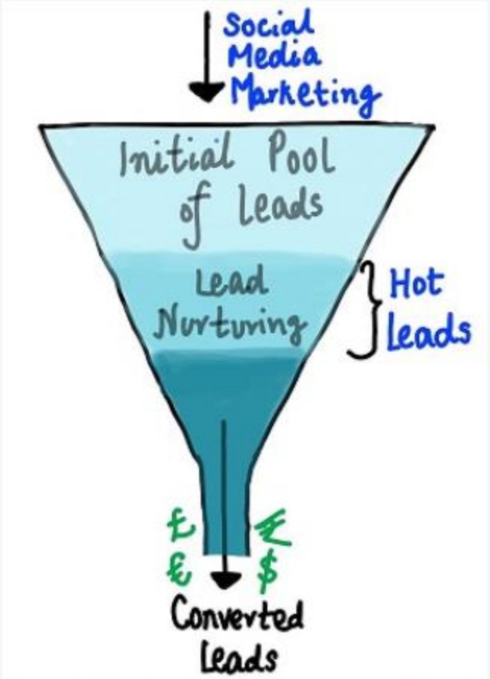
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:



Lead Conversion Process - Demonstrated as a funnel

PROBLEM STATEMENT (CONT.)

- According to diagram, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.



Lead Conversion Process - Demonstrated as a funnel

PROVIDED DATA

- Provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted. You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out for are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value (think why?).

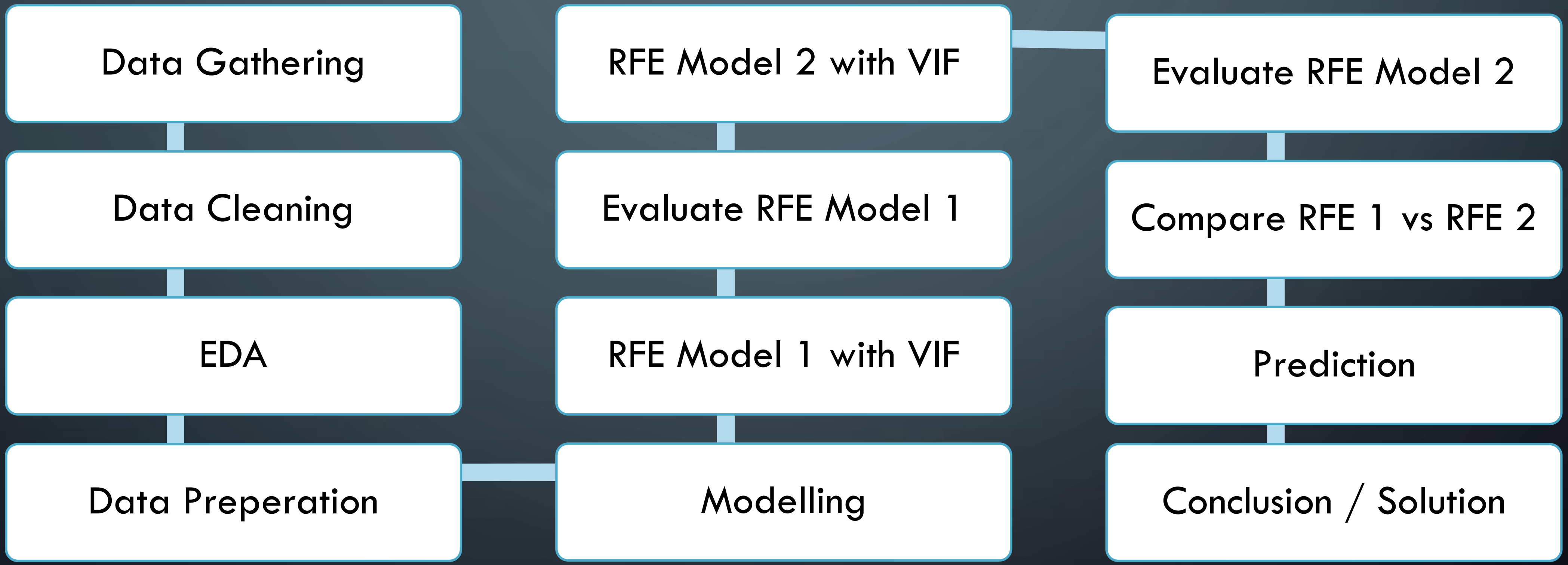
GOALS OF THE CASE STUDY

- There are quite a few goals for this case study.
  - Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
  - There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

EXPECTED RESULTS

1. A well-commented Jupyter note with at least the logistic regression model, the conversion predictions and evaluation metrics.
2. The word document filled with solutions to all the problems.
3. The overall approach of the analysis in a presentation
    1. Mention the problem statement and the analysis approach briefly
    2. Explain the results in business terms
    3. Include visualizations and summarize the most important results in the presentation
4. A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

LEAD PROCESS



Data Gathering

Data Cleaning

EDA

Data Preperation

Modelling

RFE Model 1 with VIF

Evaluate RFE Model 1

RFE Model 2 with VIF

Evaluate RFE Model 2

Compare RFE 1 vs RFE 2

Prediction

Conclusion / Solution

DATA GATHERING

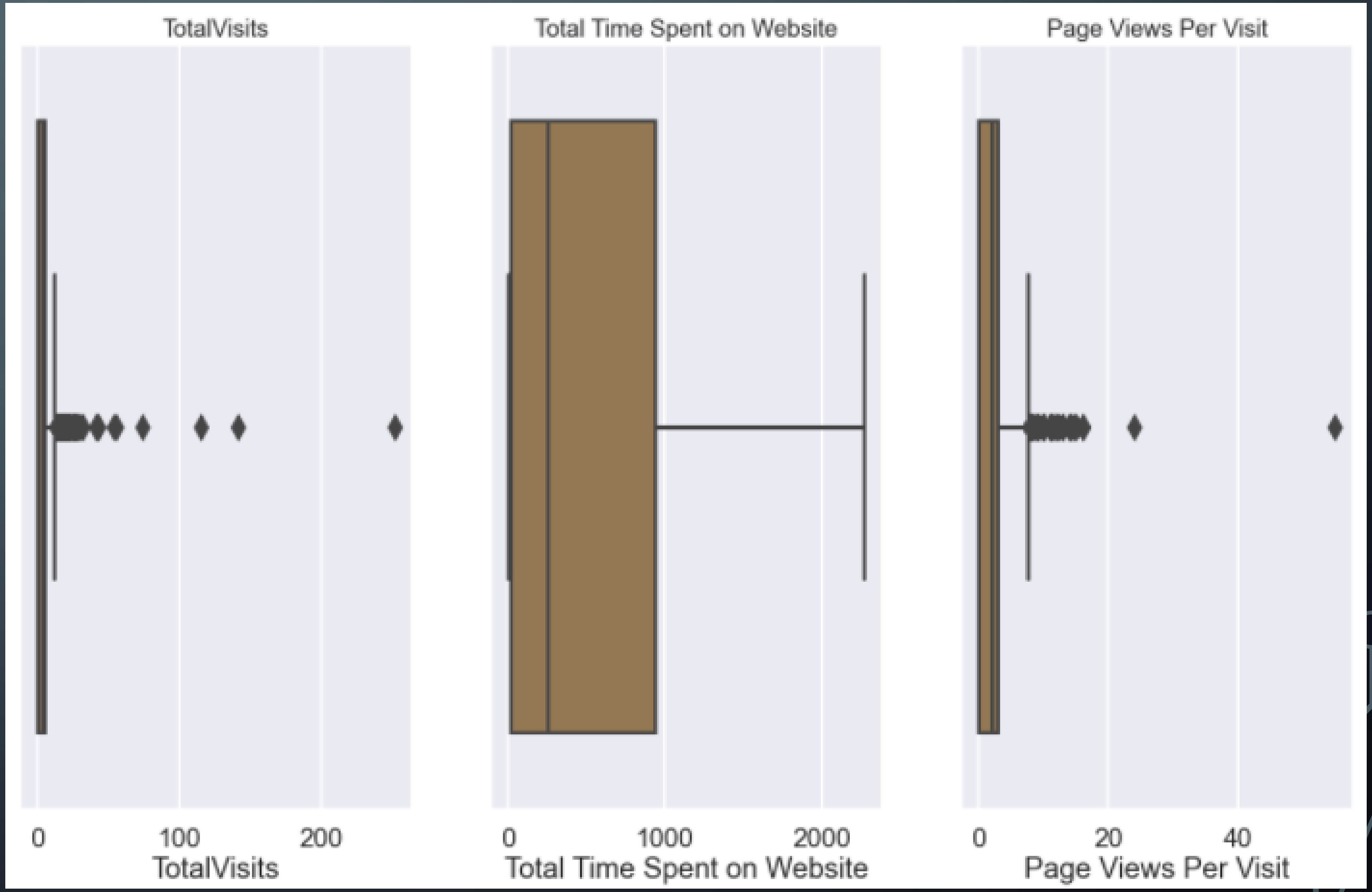
- Loading and & Observing the past data provided by the Company
- Gathering info on among the 37 columns 7 are numerical and remaining 30 are categorical variables.

DATA CLEANING

- Dropping redundant columns
- Find 'Select' label columns and categories
- Replacing ‘Select' label with nan values
- Columns with more than 30% of missing value to be dropped
- Columns with less than 30% of missing value to be imputed

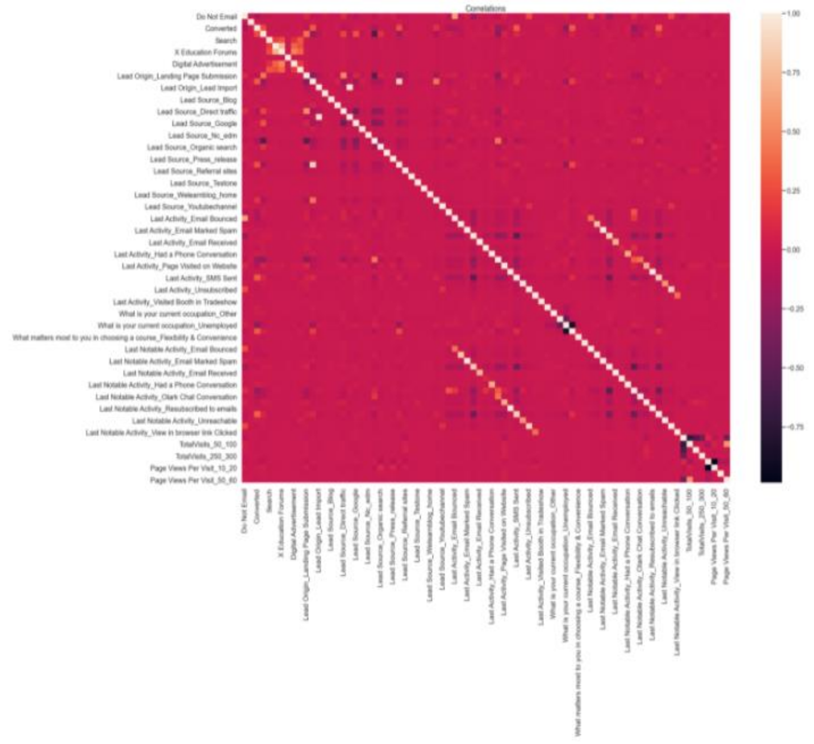
EDA

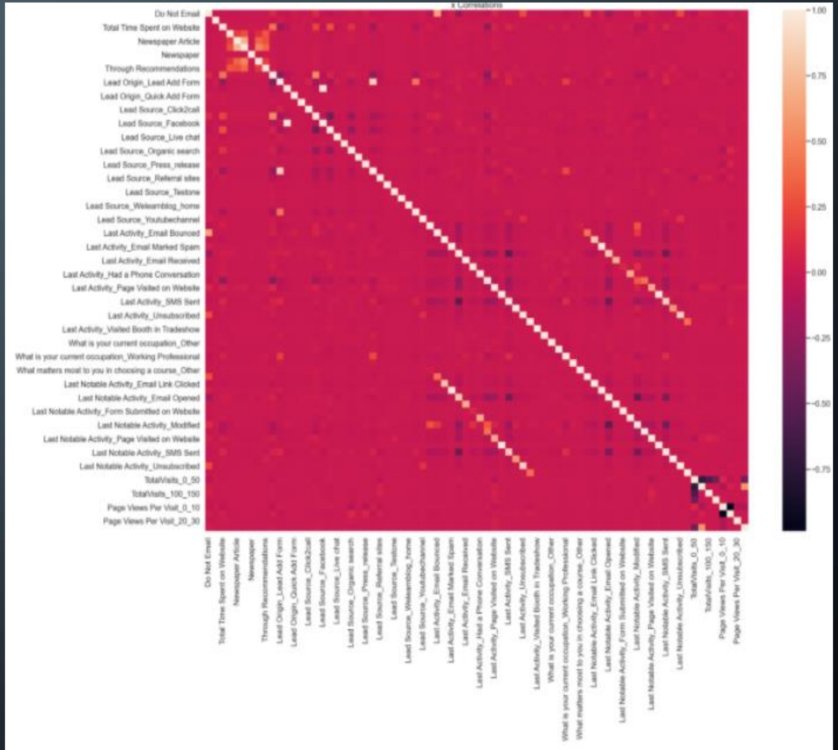
- Conduct Data Transformation
  - By converting all binary variables 0 and 1
- Univariate and Bivariate Analysis
- Assign numerical variables to categories
- Creation of Dummies



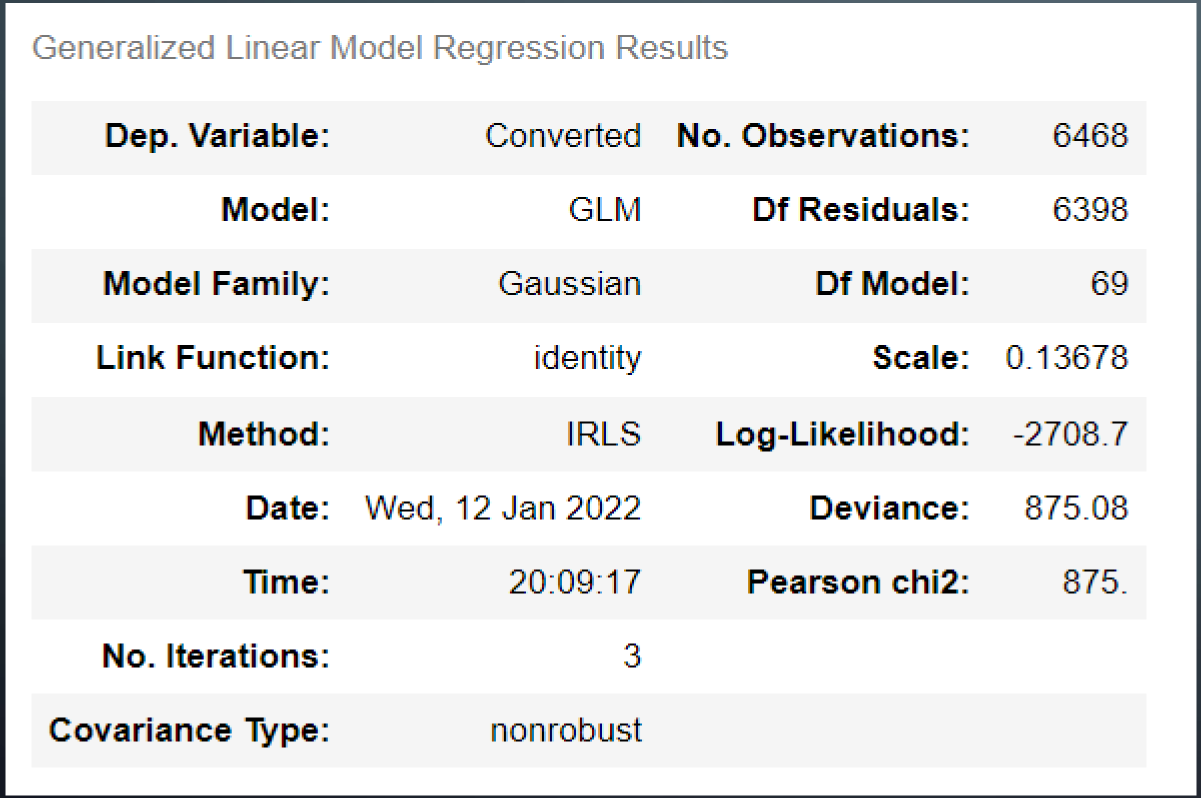
DATA PREPARATION

- Outlier Treatment
- Standardization
- Correlation





MODELLING



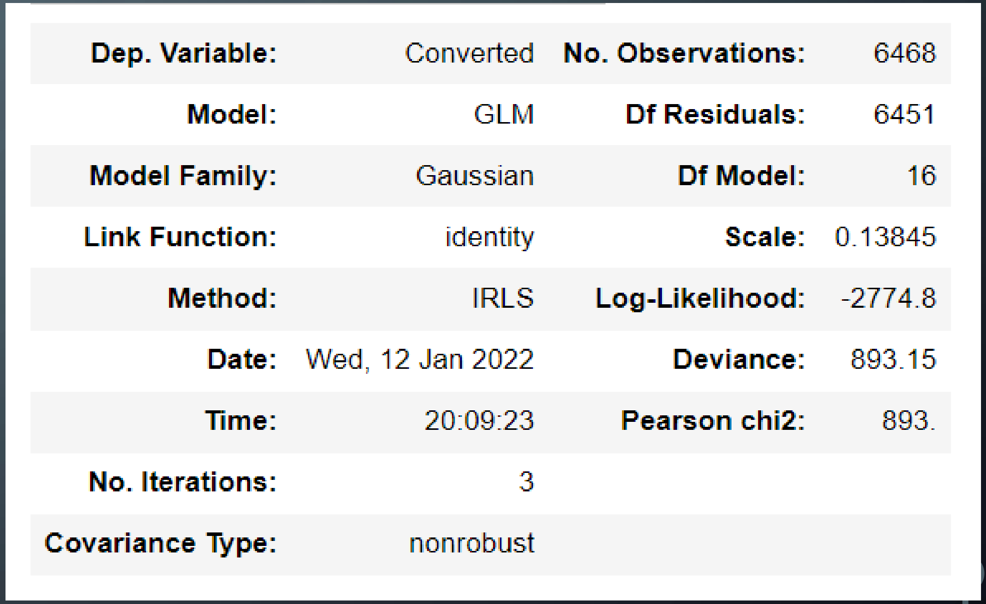
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6398
Model Family:	Gaussian	**Df Model:**	69
Link Function:	identity	**Scale:**	0.13678
Method:	IRLS	**Log-Likelihood:**	-2708.7
Date:	Wed, 12 Jan 2022	**Deviance:**	875.08
Time:	20:09:17	**Pearson chi2:**	875.
No. Iterations:	3		
Covariance Type:	nonrobust		

- As presented through the linear model, it can deduced that there are many variables, which have an insignificant p-value. Therefore, RFE selection of 70 variables is inefficient

RFE 1 WITH VIF

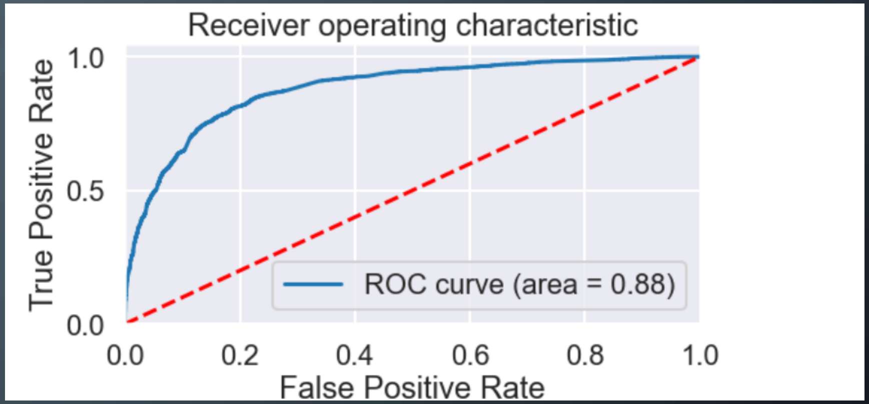
- Running RFE for 19 variables
- selection of 'True' in rfeme.support_ colums only that is True columns were selected for creating a model
- According to summary, there are high p-values, which has insignificant p-values therefore we will drop one after the other to create a new model through VIF
- At the right is the final RFE 1 model Generalized Linear Model Regression Results

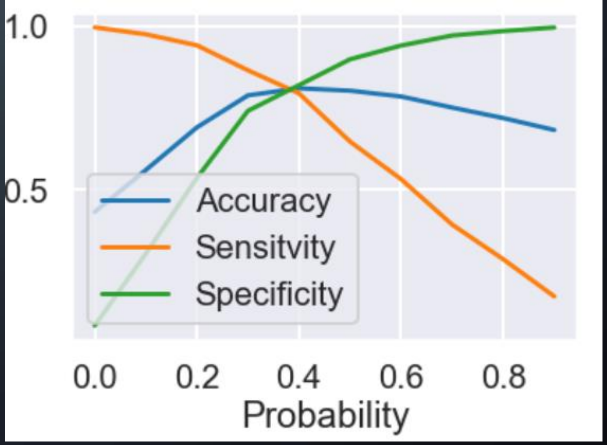


Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6451
Model Family:	Gaussian	**Df Model:**	16
Link Function:	identity	**Scale:**	0.13845
Method:	IRLS	**Log-Likelihood:**	-2774.8
Date:	Wed, 12 Jan 2022	**Deviance:**	893.15
Time:	20:09:23	**Pearson chi2:**	893.
No. Iterations:	3		
Covariance Type:	nonrobust		

EVALUATE RFE 1 MODEL

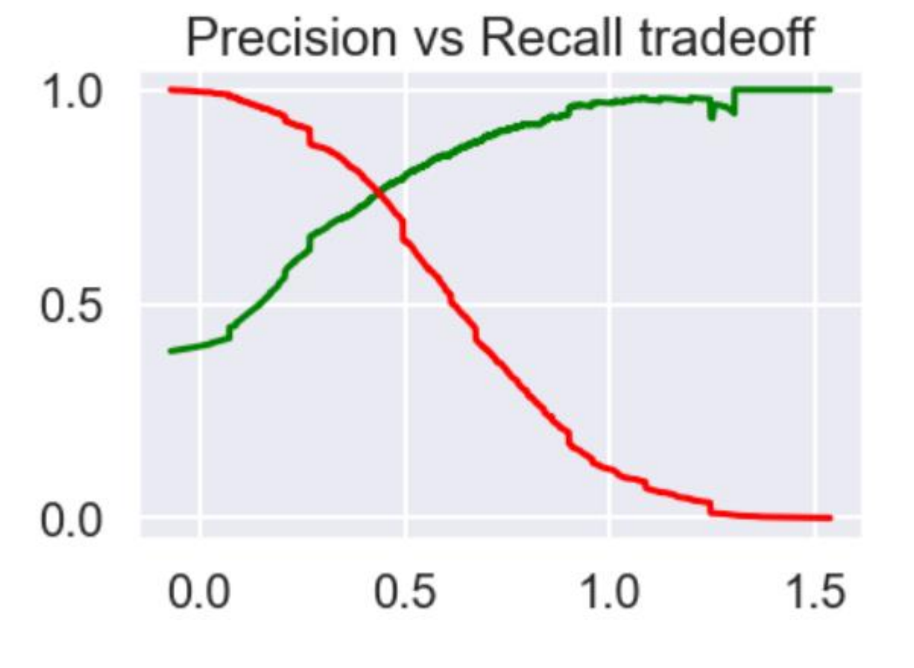
- Did ROC Curve Plotting
- Calculated accuracy, sensitivity and specificity with probability cutoffs
- Created probability points from 0 to 0.9 for accuracy , sensitivity and specificity. The probability cutoff = 0.4 as all the accuracy , sensitivity and specificity are having nearly same value which is an ideal point to consider for as we can't ignore any one from three.





EVALUATE RFE 1 MODEL (CONT.)

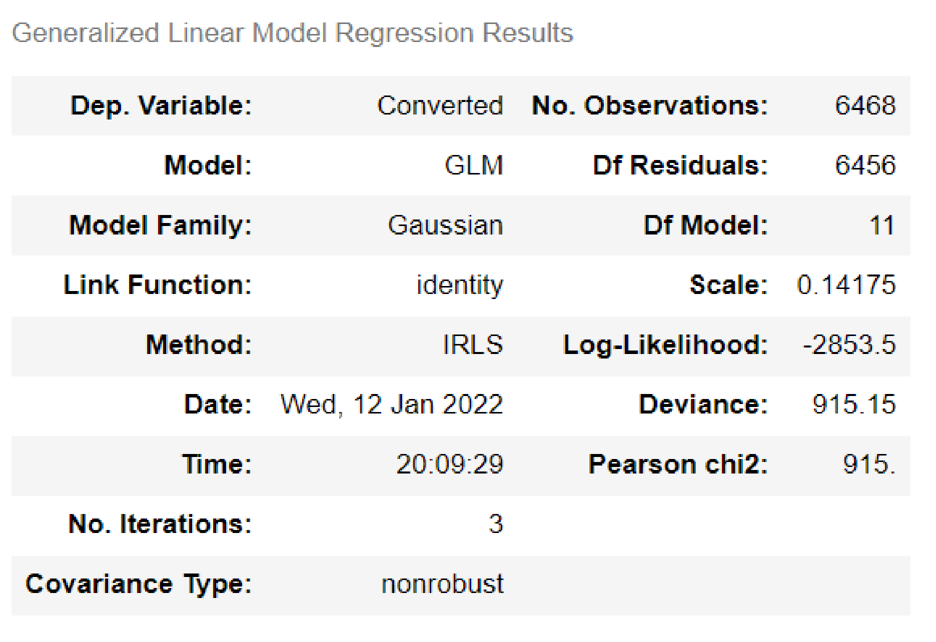
- Did a precision and recall analysis
- Precision percentage (Relevancy) is 73% approximately and recall percentage (Relevant results) is 79%
- It will be more on Recall than Percision as recall percentage is higher, which highlights less hot lead customers but don't want to left out any hot leads which are willing to get converted



Meeting point for trade off is 0.5

RFE 2 WITH VIF

- Running RFE for 15 variables
- selection of 'True' in rfeme.support_ colums only that is True columns were selected for creating a model
- According to summary, there are high p-values, which has insignificant p-values therefore we will drop one after the other to create a new model through VIF
- At the right is the final RFE 2 model Generalized Linear Model Regression Results

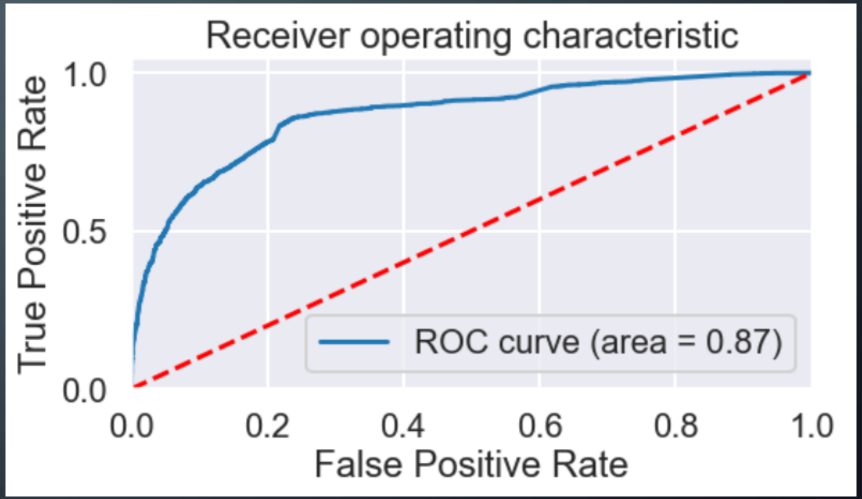


Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6456
Model Family:	Gaussian	**Df Model:**	11
Link Function:	identity	**Scale:**	0.14175
Method:	IRLS	**Log-Likelihood:**	-2853.5
Date:	Wed, 12 Jan 2022	**Deviance:**	915.15
Time:	20:09:29	**Pearson chi2:**	915.
No. Iterations:	3		
Covariance Type:	nonrobust		

EVALUATE RFE 2 MODEL

- Did ROC Curve Plotting

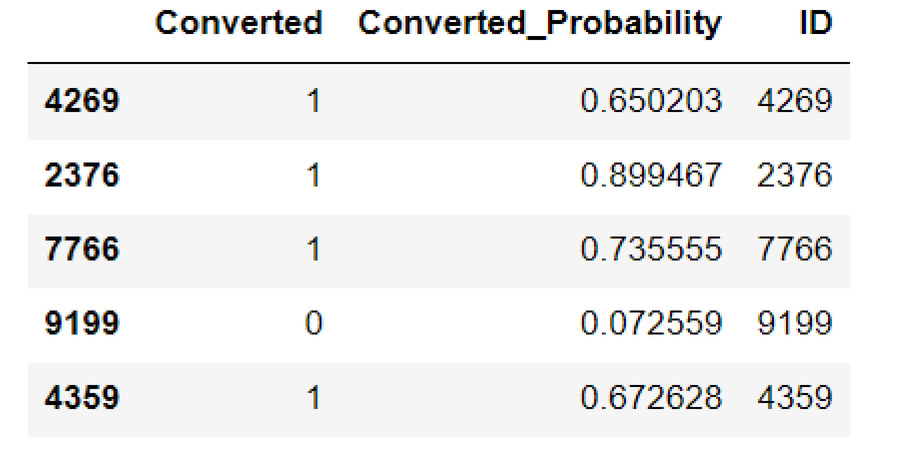


RFE 1 VS RFE 2

- For RFE Test - 1, auc score is 0.88 in ROC curve plot. For RFETest - 2, auc score is 0.87 in ROC curve plot. As AUC measures how true postive rates and false positive rates trade-off. It indicates the model stability as the larger the area, the model will be able to distinguish classes. RFE model 1 is better.

PREDICTION OF DATA SET

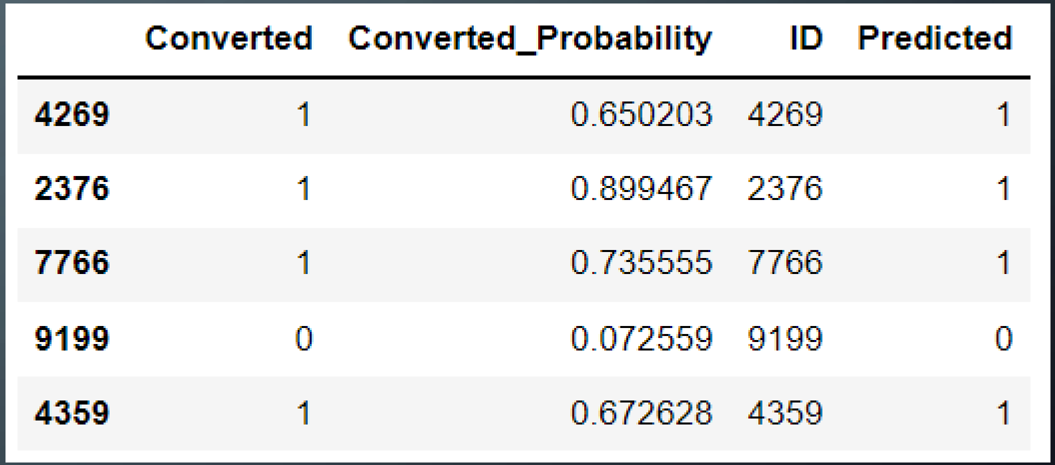
- Testing and Creating a new dataset and saving the prediction values in it



	Converted	Converted_Probability	ID
4269	1	0.650203	4269
2376	1	0.899467	2376
7766	1	0.735555	7766
9199	0	0.072559	9199
4359	1	0.672628	4359

MODEL EVALUATION

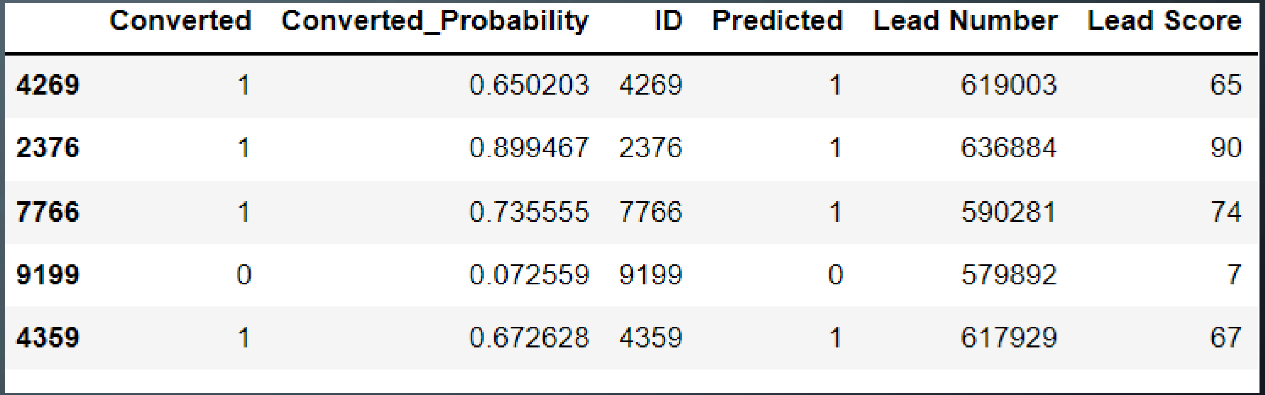
- Predict with probability cutoff as 0.4 by creating new columns in the final test dataset
- Accuracy score in predicting test dataset : 0.8152958152958153
- Precision score in predicting test dataset: 0.7541412380122058
- Recall score in predicting test dataset: 0.7899543378995434



	Converted	Converted_Probability	ID	Predicted
4269	1	0.650203	4269	1
2376	1	0.899467	2376	1
7766	1	0.735555	7766	1
9199	0	0.072559	9199	0
4359	1	0.672628	4359	1

CONCLUSION

- Acceptable range of Accuracy, Precision and Recall score
- Higher recall score than precision score
- The model meets the company requirements and is a stable model



	Converted	Converted_Probability	ID	Predicted	Lead Number	Lead Score
4269	1	0.650203	4269	1	619003	65
2376	1	0.899467	2376	1	636884	90
7766	1	0.735555	7766	1	590281	74
9199	0	0.072559	9199	0	579892	7
4359	1	0.672628	4359	1	617929	67