LEAD SCORING CASE STUDY ANALYSIS

- By Darshana Singh

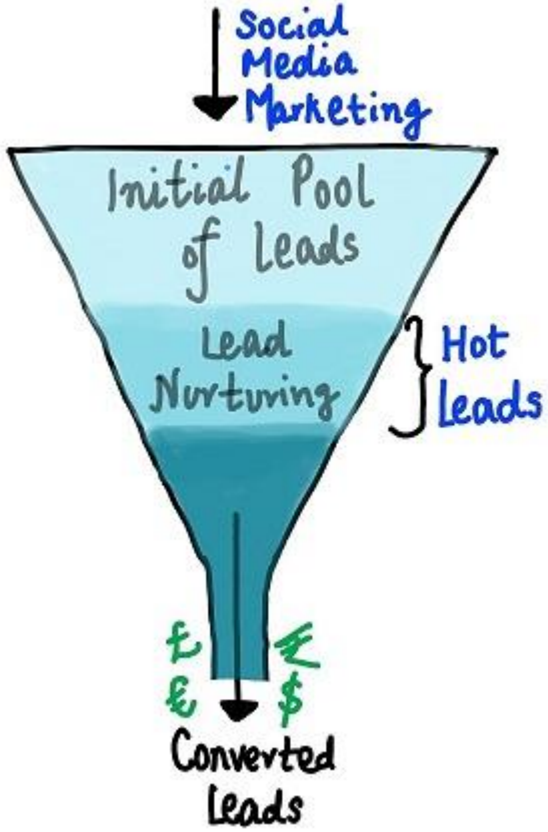
COMPANY BACKGROUND

An education company named X Education sells 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

Now, although 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:



As you can see, 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%.

MY DATA AND ANALYSIS APPROACH

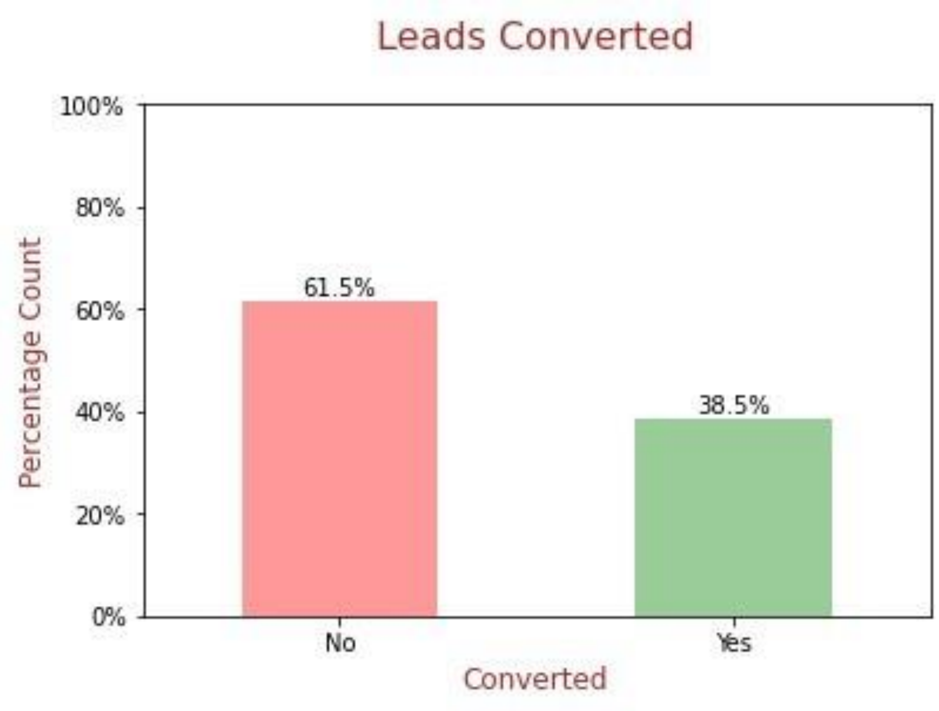
- Importing the data
- Data preparation & cleaning
- EDA (Exploratory Data Analysis)
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building (RFE R squared VIF and p-values)
- Model Evaluation
- Making predictions on test set

DATA CLEANING & PREPARATION

- First of all, data was analysed to understand the columns, data size and null value counts
- "Select" level in the data represented null values for some categorical variables, as customers did not choose any option from the list. So, we replace it for null.
- Columns with over 40% null values were dropped and below that were kept as it is.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Some columns that didn't add any insight were dropped - ('City', 'Tags', 'Country', 'What matters most
- to you in choosing a course’, 'Prospect ID’, 'Lead Number’ and 'Last Notable Activity’)
- Imputation was used for some categorical variables.
- Numerical data was imputed with mode after checking distribution.
- Skewed category columns were checked and dropped to avoid bias
- Outliers in 'TotalVisits’, 'Total Time Spent on Website’ and 'Page Views Per Visit' were treated.
- Low frequency values were grouped together to “Others”.
- Standardising Data in columns by checking casing styles, etc. (“Lead Source” has Google and google) so there are no unnecessary duplicates
- Validated that there were no discrepancies in the remaining data

EDA - EXPLORATORY DATA ANALYSIS (1/5)

All the exploratory data analysis plots/graphs/maps can be seen in the .ipynb file where they were used for different purposes and for bivariate analysis as well but here are a few to explain the analysis.

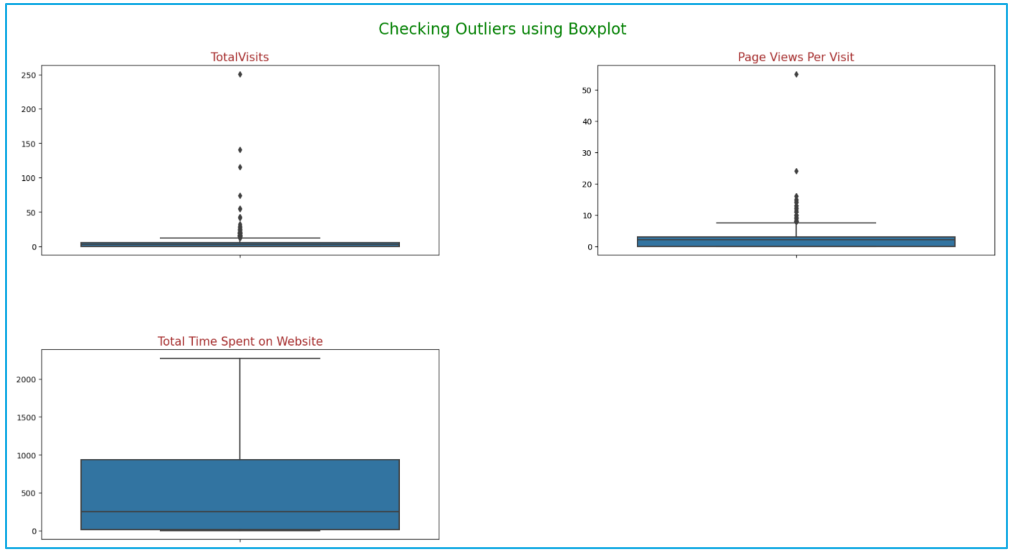


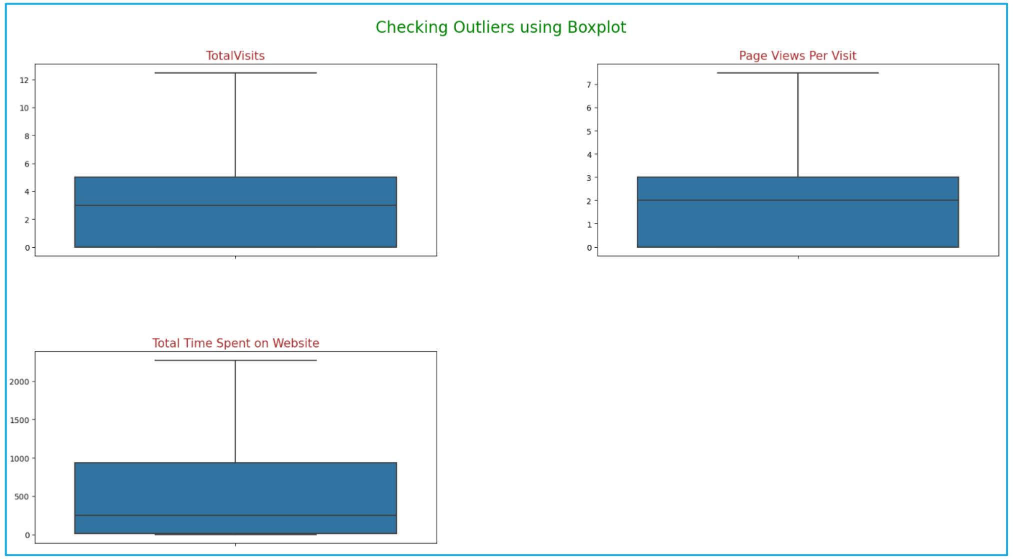
- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority)

Which is not something an organisation would like to see

EDA - EXPLORATORY DATA ANALYSIS (2/5)

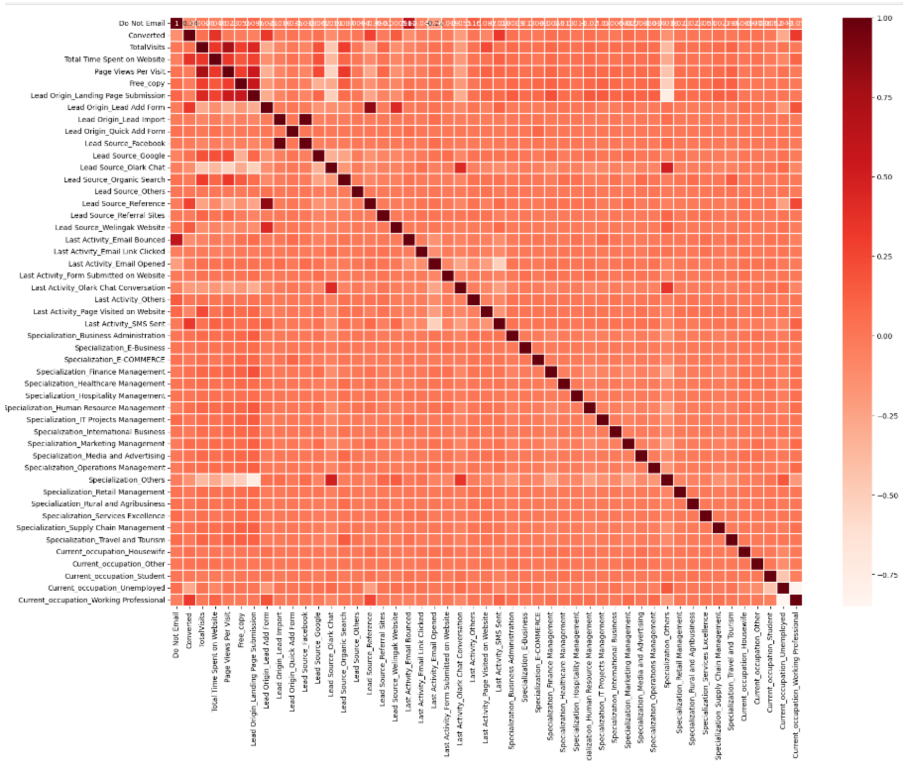
We used box plot to identify outliers in the data which were then treated succesfully

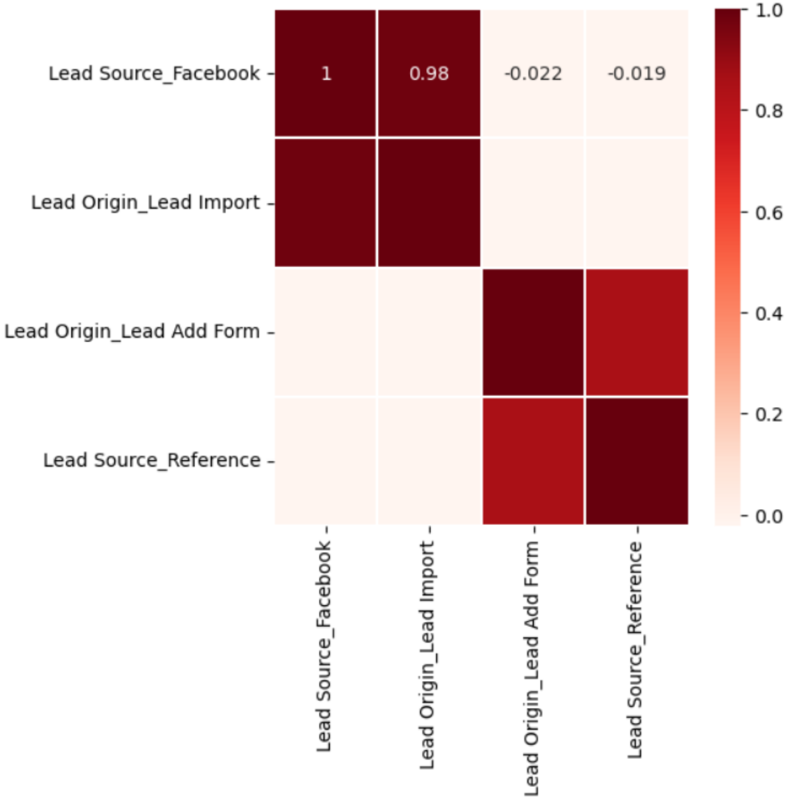




EDA - EXPLORATORY DATA ANALYSIS (3/5)

We used Heatmap to identify correlations in the data and dropped 'Lead Origin_Lead Import' and 'Lead Origin_Lead Add Form' because of high corrrelation





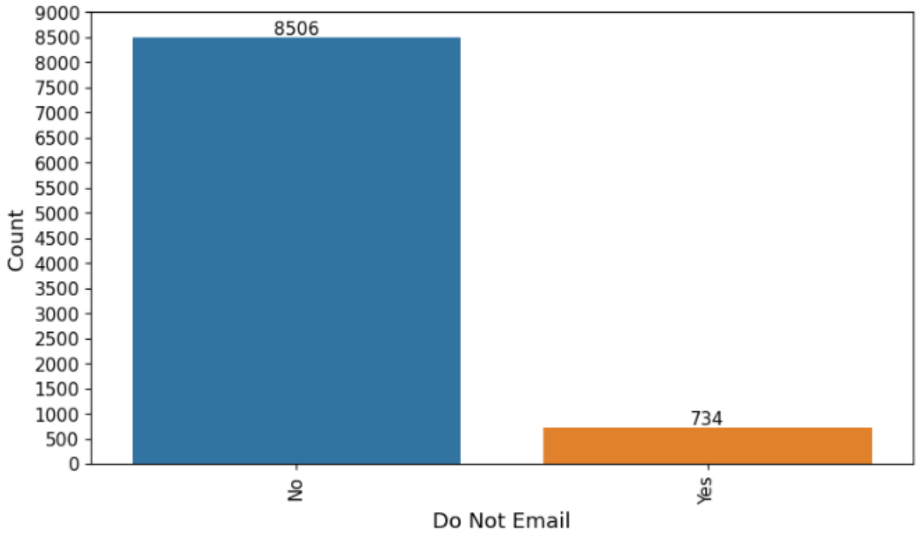
EDA - EXPLORATORY DATA ANALYSIS (4/5)

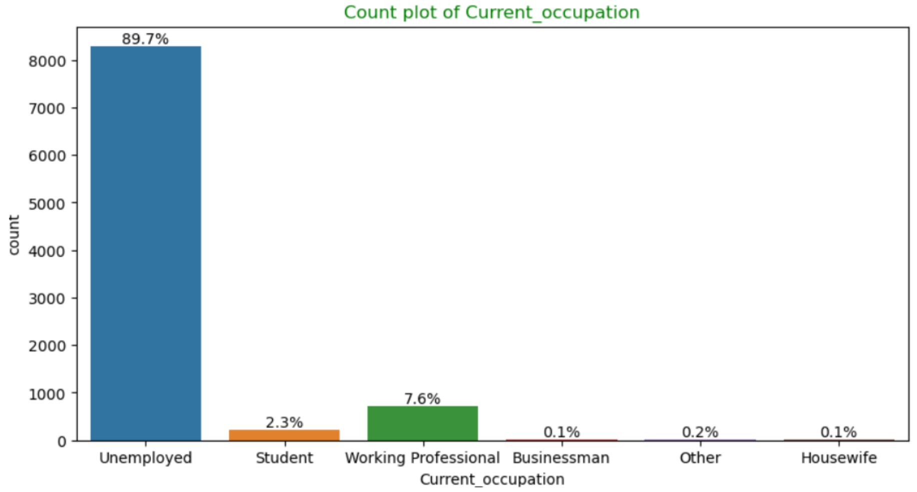
We used bar plots to identify skewness in the data & to gain insights as well.

The columns - ‘Do Not Call’, ‘Search’, ‘Newspaper Article’, X Education Forums’, ‘Newspaper’, ‘Digital Advertisement’, ‘Through Recommendation’ were highly skewed and hence were dropped from the dataset.

Insights into features of variables found were as follows:

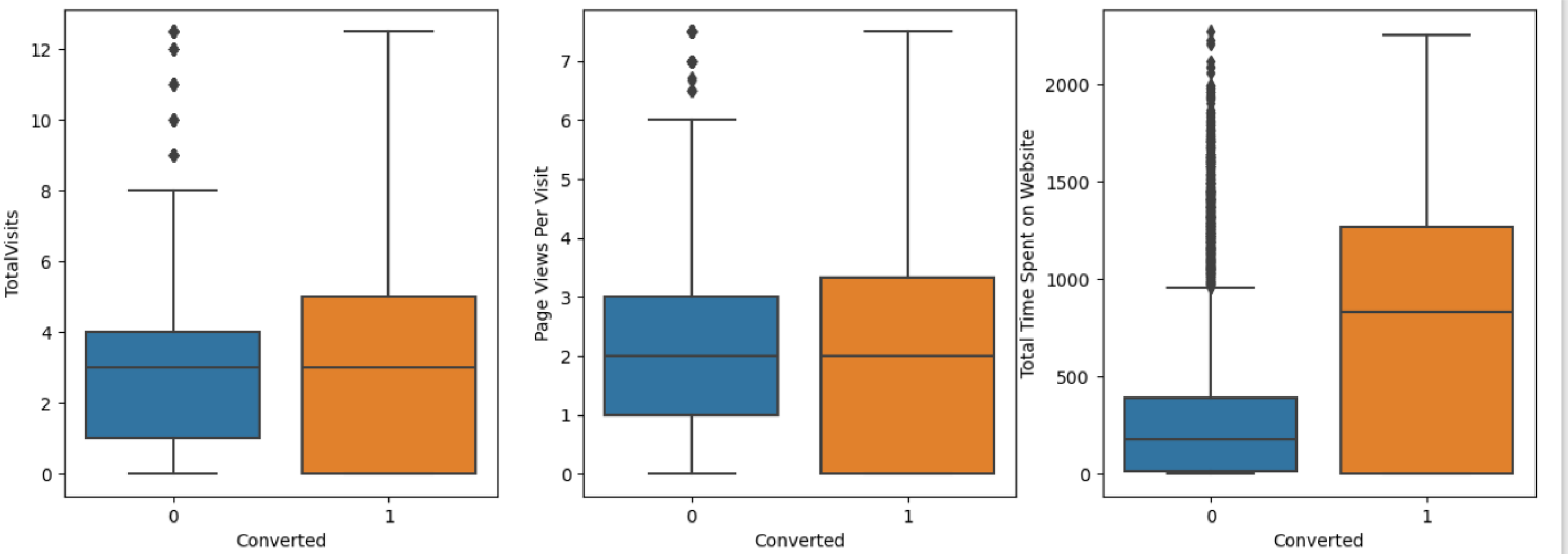
1. Lead Origin: "Landing Page Submission" identified 53% customers; "API" identified 39%.
2. Current_occupation: 90% of the customers as Unemployed
3. Do Not Email: 92% of the people has opted for not getting the email
4. Lead Source: 58% Lead source is from Google & Direct Traffic
5. Last Activity: 68% of customers contribute in SMS Sent & Email Opened Activity





EDA - EXPLORATORY DATA ANALYSIS (5/5)

Bivariate Analysis: We also use box plot to identify from where we were able to convert lead successfully and that was from time on Website. So, the more time the customer spends on website, more is the chance of getting that lead converted. Find more insightful EDA on the python .ipynb file



DATA PREPARATION BEFORE MODEL BUILDING

- Binary level categorical columns were already mapped to 1 / 0 in previous steps as shown in EDA
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation
- 1st Step for regression was the data was split into Train & Test Sets with 70:30 ratio
- Feature scaling was performed on the split dataset using standardisation method
- We checked for the correlations as seen above in EDA
- We started building model

MODEL BUILDING

The data set has many (large number of) features and dimensions which can reduce model performance and increase computation time.

Recursive Feature Elimination (RFE) is performed to select only the important columns i.e. for feature selection

Pre RFE, the data set had 48 columns and post RFE it has 15 columns.

Logistic Regression Model - 1 is a basic model where we dound higher p-values i.e. > 0.05

Manual feature reduction process was used in Logistic Regression Model by dropping columns with higher p-values

We also ensured that VIF values were <= 5. Although, No sign of multicollinearity with VIFs less than 5

After the iterations, we found that model 3 was the most stable model

Logistic Regression Model - 3 (LRMod3) is the final model used for model evaluation and making predictions.

MODEL EVALUATION

Based on the curve analysis, a cutoff probability of 0.35(approx.) is suggested as the optimal point for classification.

Confusion Matrix

```
[[3057  945]
 [ 492 1974]]
```

True Negative : 3057

True Positive : 1974

False Negative : 492

False Positve : 945

Model Accuracy : 0.7778

Model Sensitivity : 0.8005

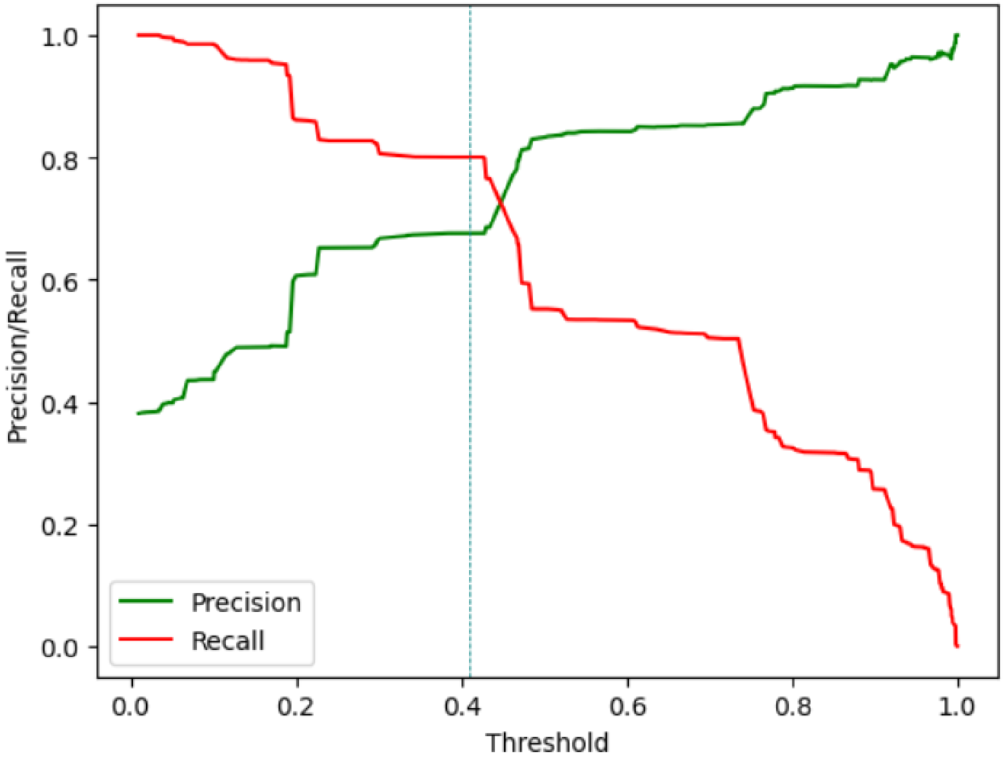
Model Specificity : 0.7639

Model Precision : 0.6763

Model Recall : 0.8005

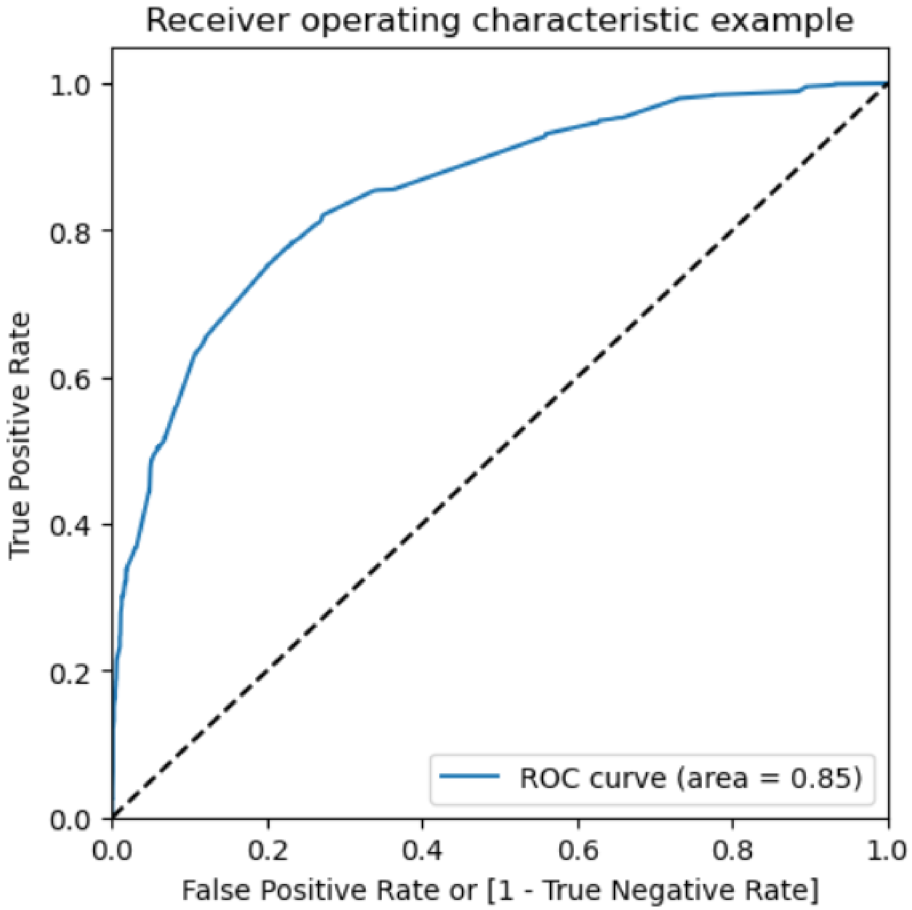
Model True Positive Rate (TPR) : 0.8005

Model False Positive Rate (FPR) : 0.2361



PREDICTIONS ON THE DATASET (1/2)

- The Area under ROC curve was found to be 0.85 out of 1, indicating that the model is a good predictor.
- The curve is plotted as close to the top left corner of the plot as possible, which indicates that the model has a high true positive rate and a low false positive rate at all threshold values.

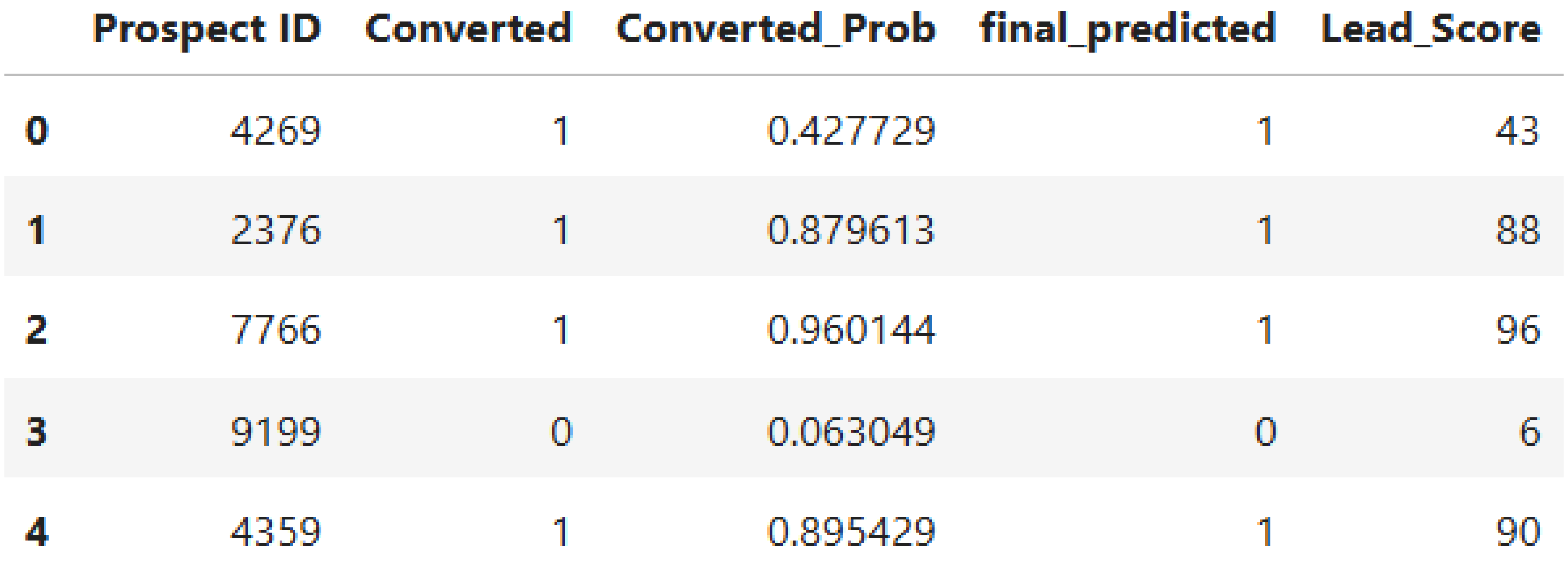


Area under ROC curve is 0.85 out of 1 which indicates a good predictive model

PREDICTIONS ON THE DATASET (2/2)

The customers with a higher lead score have a higher conversion chance

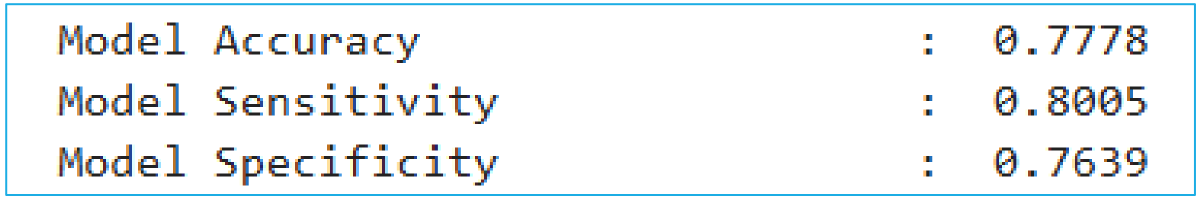
The customers with a lower lead score have a lower conversion chance.



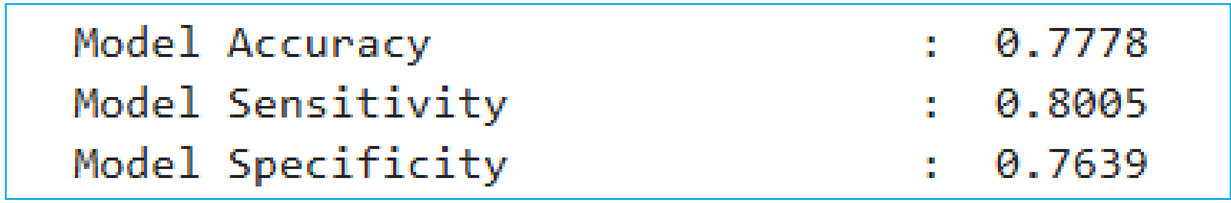
	Prospect ID	Converted	Converted_Prob	final_predicted	Lead_Score
0	4269	1	0.427729	1	43
1	2376	1	0.879613	1	88
2	7766	1	0.960144	1	96
3	9199	0	0.063049	0	6
4	4359	1	0.895429	1	90

CONCLUSION (1/2)

The evaluation metrics of the model are similar to each other, indicating that the model is performing consistently across different evaluation metrics in both the test and train datasets. This consistency suggests that the model is reliable and is not overfitting to the training data. It also implies that the model is generalizing well to new data, which is important for real-world applications. The similar performance across evaluation metrics also means that there are no significant biases in the model's predictions. This is a positive sign for the model's performance and provides confidence in its ability to make accurate predictions in the future.



```
Model Accuracy          :  0.7778
Model Sensitivity       :  0.8005
Model Specificity       :  0.7639
```



```
Model Accuracy          :  0.7778
Model Sensitivity       :  0.8005
Model Specificity       :  0.7639
```

CONCLUSION (2/2)

- 'Lead Origin_Lead Add Form', 'Current_Occupation_Working Professional' and 'Total Time Spent are effective factors that contribute to a good conversion rate.
- Working professionals and Unemployed customers tend to have higher conversion rates.
- Referral leads generated by old customers have a significantly higher conversion rate
- Google and Direct Traffic are channels that are showing promising conversion rates.
- Leads whose 'Last Activity' is ‘SMS Sent’ or ‘Email Opened’ tend to have a higher conversion rate.
- The 'Others' specialization category is the most common among customers followed by Finance Management, HR Management and Marketing Management.

RECOMMENDATIONS

- Features such as 'Lead Origin_Lead Add Form', 'Current_Occupation_Working Professional', and 'Total Time
- Spent on Website' have a high conversion rate and should be utilised more in lead generation efforts.
- Working professionals should be aggressively targeted as they have a higher probability of converting and are likely to have better financial situations to pay for services.
- Referral leads generated by old customers have a significantly higher conversion rate and should be incentivised with discounts or other rewards to encourage more referrals.
- Increasing the frequency of media usage such as Google ads or email campaigns can save time and increase the conversion rate.
- Leads whose 'Last Activity' is ‘SMS Sent’ or ‘Email Opened’ tend to have a higher conversion rate and should be
- targeted more frequently.
- Analysing the behavior of customers who spend more time on the website can help improve the user experience and increase conversion rates, and company should focus on creating engaging content and user- friendly navigation to encourage customers to spend more time on the website.
- Understanding the most popular specializations can help tailor course offerings and marketing campaigns to specific groups of customers. Providing targeted content and resources for popular specializations such as Marketing Management and HR Management can also help attract and retain customers in those fields.

THANK YOU!!!