Lead Scoring Case Study SUBMISSION REPORT

GROUP MEMBERS:

AKILESHWARI DJ AND VIVEK SHARMA

Business Objective

Problem statement: X Education markets its courses on several websites and search engines like Google and get a lots of leads but conversion rate of the company is around 30%.

Objective:
To help X Education to identify the most promising leads out of leads that they receive from people filling a form on website or through past referrals. These hot leads should categorize users that are most likely to convert into paying customers. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%. Also, the model should be able to adjust if the company's requirement changes in the near future.

Approach:
- To build a logistic regression model to predict lead conversion probability
- Decide on a probability threshold value above which a lead will be predicted as converted, whereas not converted if it is below it.
- Assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads.

Methodology

Understanding & Cleaning the Data: We have used the Jupyter Python Notebook to load and understand the Leads datasets. We have gone through the breadth & the depth of the features present in the datasets along with their definition to assess the data quality & it's spread at a high level. As part of the Data Cleaning process, we had found & treated all the irregularities in the dataset such as Missing Values; Outliers; Skewed Categorical features & Invalid data points.

Exploratory Data Analysis: After cleaning the date, we performed various types of Univariate, Bivariate & Multivariate analysis by plotting appropriate graphs with respect to the Target variable. This helped us to draw relevant insights & correlations present within the dataset.

Data Preparation: Here, we had firstly created the dummy variables of all the categorical features in the dataset. Then, we had split the dataset between training & test sets & after that, we performed the Standard Scaling of the independent features.

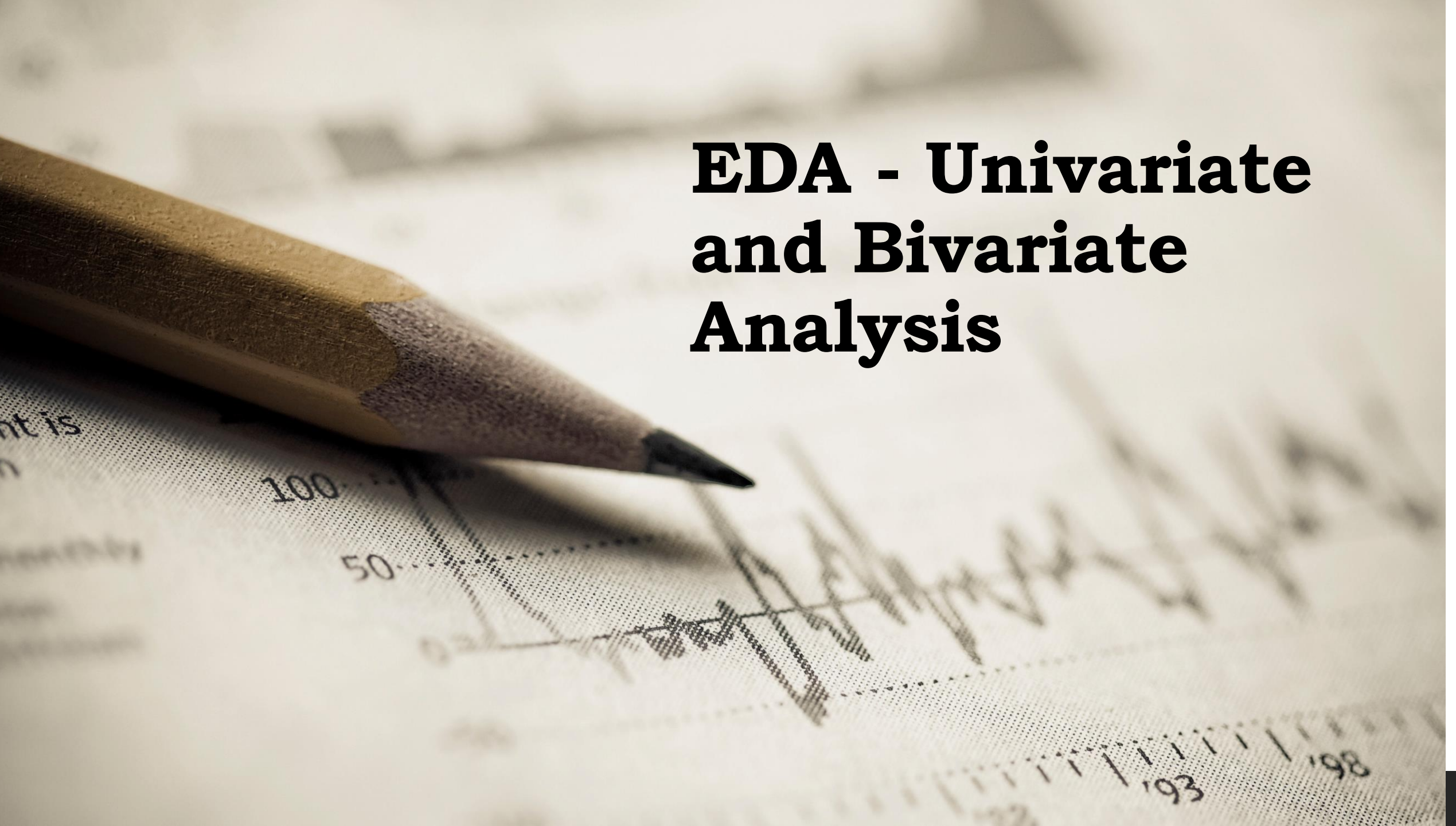
Model Building: Since the count of features were high, we first started with RFE to eliminate redundant features & then built the first Model. Over multiple iterations of refinement(through p-value & VIF), we concluded this step with a final model.

Model Evaluation: Here, with the final model, we first obtained the lead score and plotted the ROC Curve. After that, we determined the optimal cut-off to proceed further with the Evaluations Metrics.

Making Predictions: After evaluating the final model, we ran the model over the test dataset to make predictions & then, we reviewed the predicted results with the actual records. We finally concluded with our analysis findings & recommendations to the business.

Data Cleaning and preparation

- Dropping irrevlant and high missing value variables: Prospect ID','Lead Number','How did you hear about X Education', 'Lead Profile','Lead Quality','Asymmetrique Profile Score','Asymmetrique Activity Score', 'Asymmetrique Activity Index','Asymmetrique Profile Index','Tags','Last Notable Activity Segregating Categorical & Numerical Values
- Categorical Columns Null Value Treatment
- X-Education is online teaching platform. The city information will not be much useful as potential students can available any courses online despite their city. We will drop the column from analysis.
- It may be possible that the lead has no specialization or may be a student and has no work experience yet , thus he/she has not entered any value. We will create a new category called 'Others' to replace the null values
- As the data in" What matters most" is skewed, we can delete the column.
- 85.5% values are "Unemployed". If we impute the data as "Unemployed" then data will become more skewed. Thus, we will impute the value as "Unknown".
- Country data is heavily skewed as 95% of the data is mapped as India. Similar to City, Country data is not required for Model building as X-Education is online platform. We will drop the country columns too.
- As we are unsure what could be the Last activity, we will replace it with the most frequent activity "Email Opened"
- As Google seems to be most used Lead Source, we will replace null values with Google. There is a category 'google' which is same as 'Google' We will replace the values
- Numerical Columns Null Value Treatment
- Outliers Treatment
- Convert Binary Categories
- Creating a dummy variable for some of the categorical variables and dropping the first one
- here are 51 columns in Heatmap which makes it difficult to interpret. Let's review top 5 positively and negatively correlated features



EDA - Univariate and Bivariate Analysis

Univariate and Bivariate Analysis - Numerical

NUMERIC FEATURES ANALYSIS:

• After cleaning the data, we moved to the EDA for further insights.

• From the distribution plot(at the right), we can see that the spread of Total Visits &

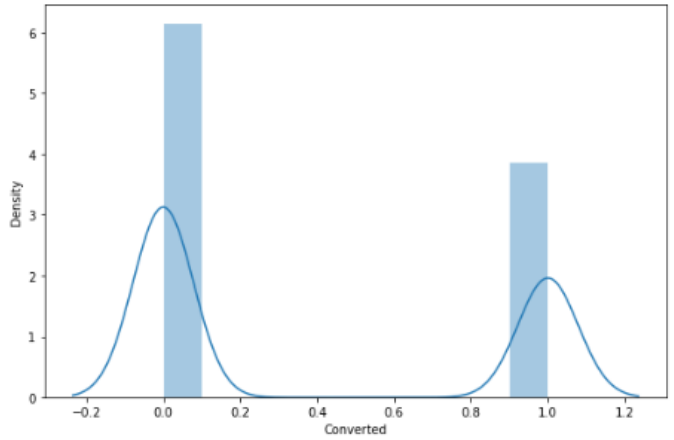
Total Time Spent per Visit both are right/positive skewed.

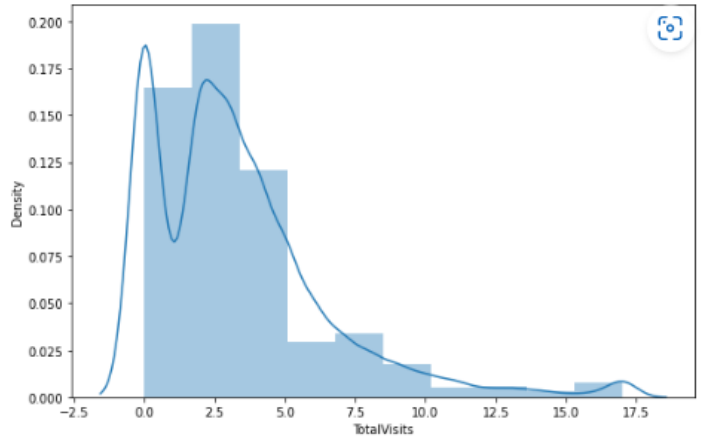
• It seems that most of the leads are not frequent visitors & their average

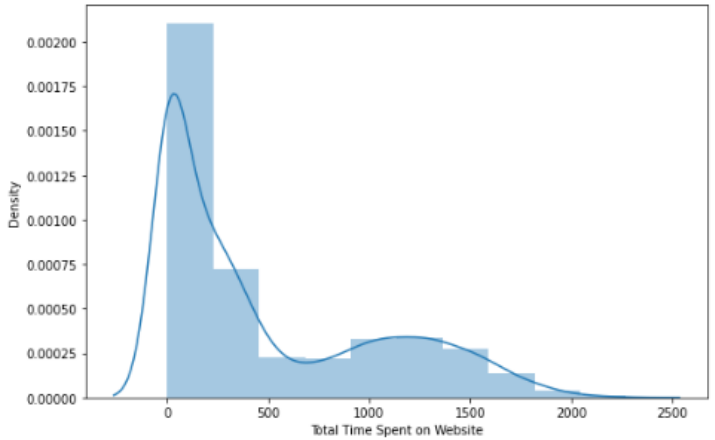
engagement(time spent & page views per visit) on he website is slightly on a lower side.

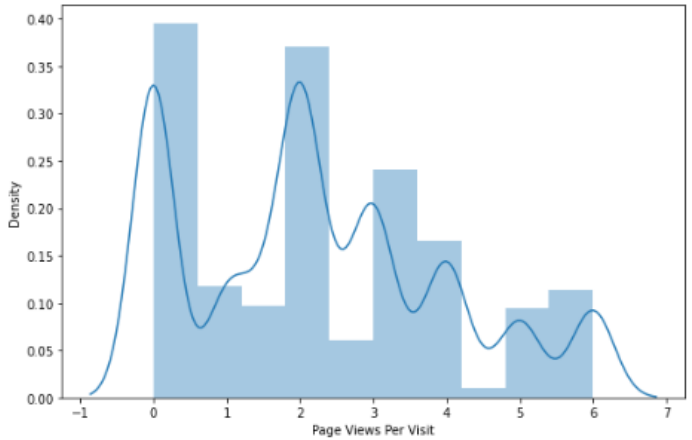
• Here, the Class Imbalance of the Target Variable(Converted) is = 1.6

This seems fine and we are good to proceed with further analysis









Univariate and Bivariate Analysis - Numerical

NUMERIC FEATURES ANALYSIS:

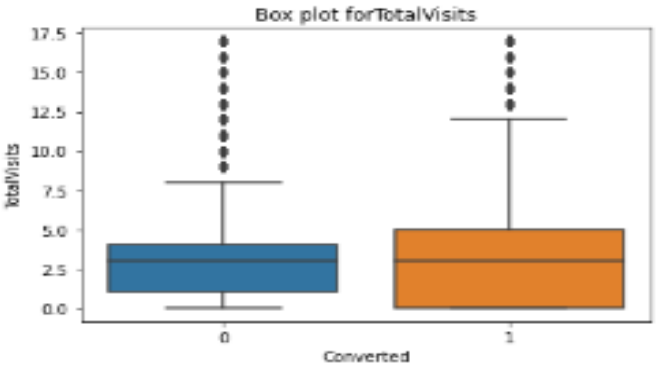
• Here are the box plots w.r.t. converted vs. non-converted leads.

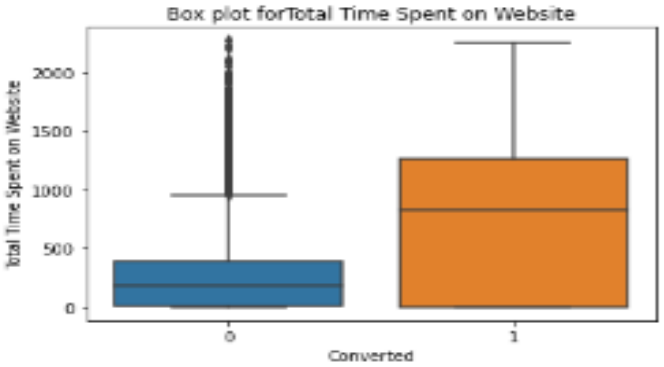
• It is evident that leads who got converted have visited & engaged more

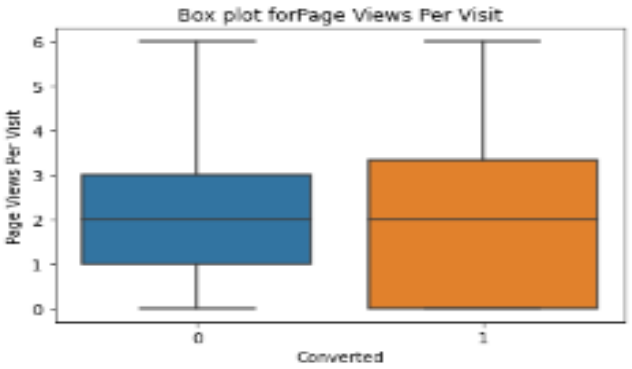
on the website & therefore have higher no. of visits & time spent on website compared to the non-converted leads

• For ‘Page Views per Visit’ metric, we can

say that it is slightly better for converted leads. But otherwise, there doesn’t seems to be a major difference we can notice for this metric between converted & non-converted leads





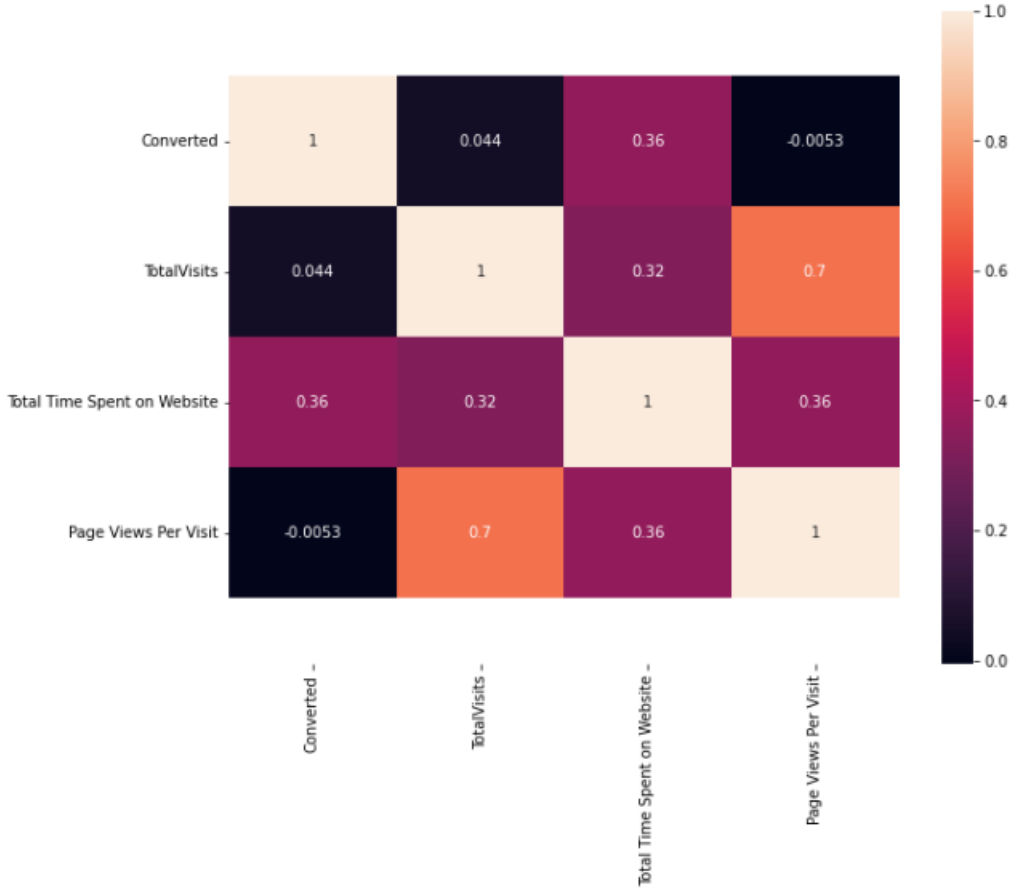


Univariate and Bivariate Analysis - Numerical

NUMERIC FEATURES ANALYSIS:

• In terms of correlation, from the correlation matrix, we have noticed a decent correlation of 0.7 between ‘Total Visits’ & ‘Page Views per Visit’ features.

Otherwise, there isn’t any major correlations we could see between the numeric variables.



Categorical Feature Analysis:

• Here, we plotted each categorical feature with respect to their respective

conversion rate.

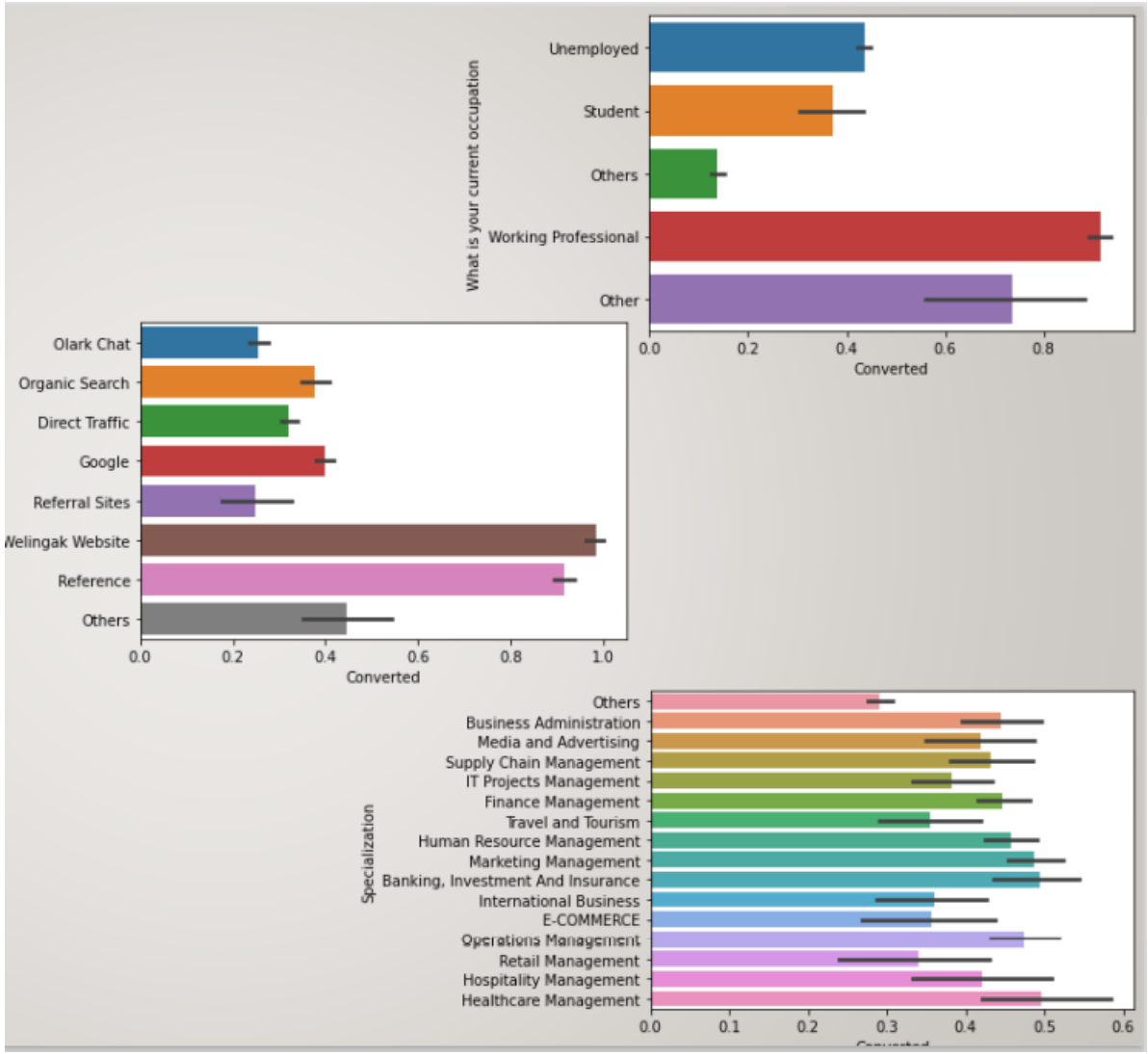
• It has been observed that ‘Working Professions’ leads have higher conversion rate than anyone else.

• Leads who have come from ‘Welingakwebsite’ or through some ‘Reference’ had relatively higher conversion rate

• Leads who have their specialization in

‘Healthcare Management’ or ‘Banking

Investments & Insurance’ have a higher chances of conversion.



Building the Model

Model Building

Model Building:

• We had built a model with all the features included and found that there were many insignificant variables present in our model.

• With the help of RFE, we reduced the no. of features to15 most significant features & built the model again.

• In the next model, from the summary stats, we decided to drop the insignificant feature with high p-value.

• In the 3rd iterations of model building & feature selection, we concluded with a statistically significant & most stable model with 14 statistically significant features. This model has all the p-values & VIFs within the permissible range.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6453
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2682.5
Date:	Sun, 07 Mar 2021	**Deviance:**	5365.1
Time:	19:28:40	**Pearson chi2:**	8.50e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	Features	VIF
0	const	5.04
2	Lead Source_Olark Chat	1.45
14	What is your current occupation_Working Profes...	1.34
13	What is your current occupation_Unemployed	1.32
8	Last Activity_Olark Chat Conversation	1.30
1	Total Time Spent on Website	1.25
3	Lead Source_Reference	1.14
9	Last Activity_SMS Sent	1.12
5	Do Not Email_Yes	1.09
10	Last Activity_Unsubscribed	1.07
6	Last Activity_Converted to Lead	1.06
12	What is your current occupation_Student	1.05
4	Lead Source_Welingak Website	1.03
7	Last Activity_Had a Phone Conversation	1.01
11	Specialization_Hospitality Management	1.01

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.2147	0.087	-25.354	0.000	-2.386	-2.044
Total Time Spent on Website	1.0594	0.039	27.116	0.000	0.983	1.136
Lead Source_Olark Chat	1.2232	0.103	11.872	0.000	1.021	1.425
Lead Source_Reference	3.4195	0.203	16.832	0.000	3.021	3.818
Lead Source_Welingak Website	5.2346	0.724	7.226	0.000	3.815	6.654
Do Not Email_Yes	-1.4200	0.168	-8.428	0.000	-1.750	-1.090
Last Activity_Converted to Lead	-1.2426	0.219	-5.674	0.000	-1.672	-0.813
Last Activity_Had a Phone Conversation	2.1693	0.676	3.208	0.001	0.844	3.495
Last Activity_Olark Chat Conversation	-1.2312	0.165	-7.445	0.000	-1.555	-0.907
Last Activity_SMS Sent	1.2195	0.074	16.481	0.000	1.075	1.365
Last Activity_Unsubscribed	1.1960	0.461	2.593	0.010	0.292	2.100
Specialization_Hospitality Management	-0.8886	0.319	-2.786	0.005	-1.514	-0.264
What is your current occupation_Student	1.2243	0.235	5.210	0.000	0.764	1.685
What is your current occupation_Unemployed	1.1323	0.085	13.352	0.000	0.966	1.298
What is your current occupation_Working Professional	3.6524	0.197	18.540	0.000	3.266	4.038

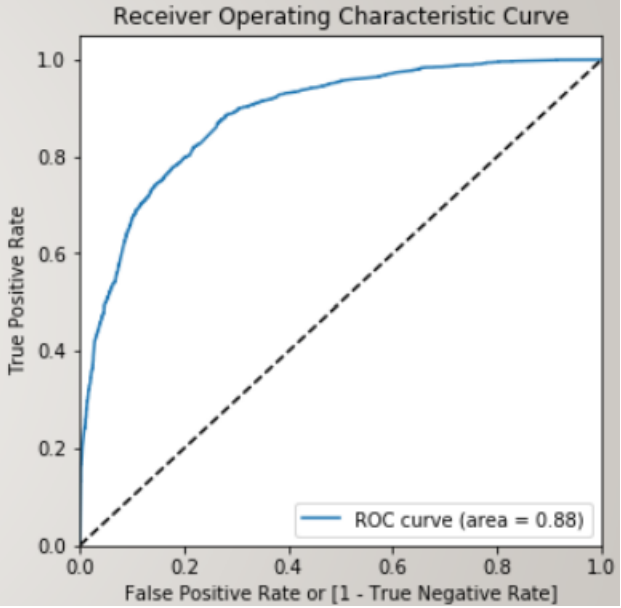
Evaluate the Model

Evaluate the Model

Model Evaluation – Lead Score and ROC Curve

- After building the final model, we predicted the probability of getting converted on our train dataset, which is termed as the “Lead_Score”(varies between 0 to 1). With this, we later determined the optimal probability cut-off to determine if the leads were converted or not.
- From the final model, after making predictions on it(on train dataset), we created the ROC Curve to determine the model stability with AUC Score (Area Under the Curve). As we can see from the graph plotted on the right side, this score is 0.88 which is a great score & represents quite stable & reliable model.
- In other words, as we see the ROC Curve is leaned towards the left side of the border, this means that the performance of our final model would be great.

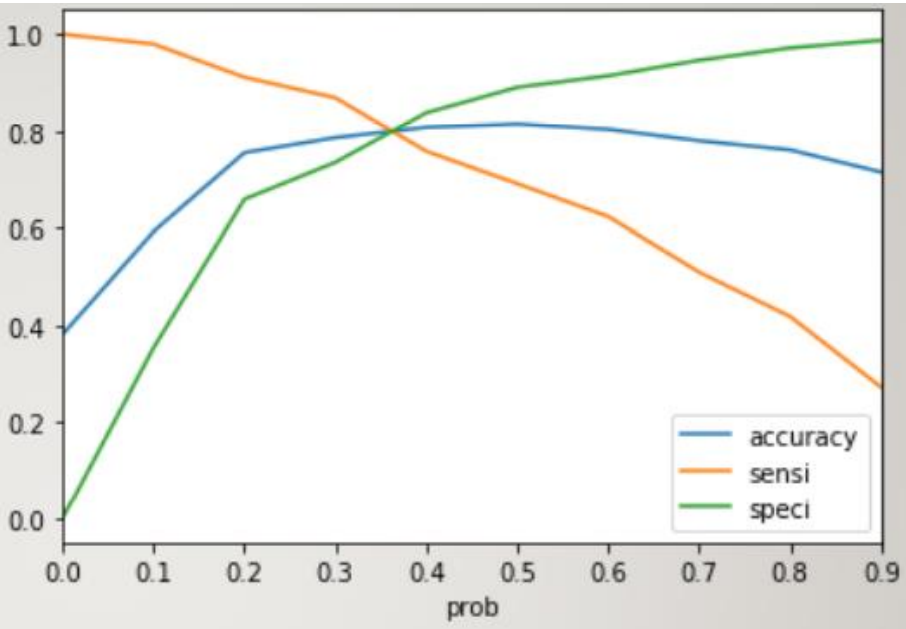
	Converted	Lead_Score
0	0	0.310625
1	0	0.254190
2	0	0.356241
3	0	0.829693
4	0	0.310625



Model Building

Model Evaluation : Optimal Cut-off

- In order to find the most optimal cut-off, we had plotted the graph between ‘Accuracy’, ‘Sensitivity’, and ‘Specificity’ at different probability/lead_score values.
- From this plot(on the right), we looked at the intersection point of accuracy, sensitivity and specificity which came out to be at 0.35, where all the score are in a close range which is the ideal point to select and hence it was selected.
- Therefore, the final probability/lead_score cut-off value that we decided was 0.35
- With this cut-off value, we had created the Confusion Matrix(see table at the right) to check the Accuracy, Sensitivity, Specificity, Precision & Recall of the model



Confusion Matrix on Training Dataset

(Cut-off = 0.35) ACTUAL	PREDICTED	
	0 (Non-Converted)	1 (Converted)
0 (Non-Converted)	3228	774
1 (Converted)	520	1946

Model Building

Model Evaluation : Evaluation Metrics

- With the probability cut-off value of 0.35 & the Confusion Matrix, we calculated the Evaluation Metrics here. >>
- Precision and Recall have a very important role in model evaluation & business decision making as it tells how our model will behave on unknown datasets.
- Our one of the Business Objectives was to achieve a Recall of 80% which means that the business wanted most of the hot leads to be identified so that the sales team can take appropriate actions to convert those hot leads.
- Therefore, our final model is now apt enough to identify all such hot leads for the sales team.

EVALUATION METRICS	SCORE (rounded)
Accuracy	80%
Sensitivity	79%
Specificity	81%
Precision	72%
Recall	79%


Prediction and
Recommendation

Model Prediction

Prediction

- After evaluating the final model, we ran the model over the test dataset to make predictions & then, we reviewed the predicted results with the actual records.
- The model is evaluated on the test dataset with the help of the Evaluation Metrics.
- The results shows that our model is very much stable even on unknown datasets

EVALUATION METRICS	SCORE (rounded)
Accuracy	80%
Precision	73%
Recall	80%

Conclusion and Recommendation

- Model performed equally well on the test dataset as it had on the training dataset. This shows that the **model is quite stable** and has a very good Accuracy & Recall.
- Based on changes in probability cut-off, the model is adjusting with the change in company's requirements in the near future.
- Below are the top 3 Important Features that the company should focus on to further increase the conversion rate of the leads:

1. Lead Source as "WelingakWebsite"
2. Lead Source as "Reference"
3. Current Occupation as "Working Professional"


Thank you!!