LEAD SCORING

Content

1 This is a Problem Statement and Objective

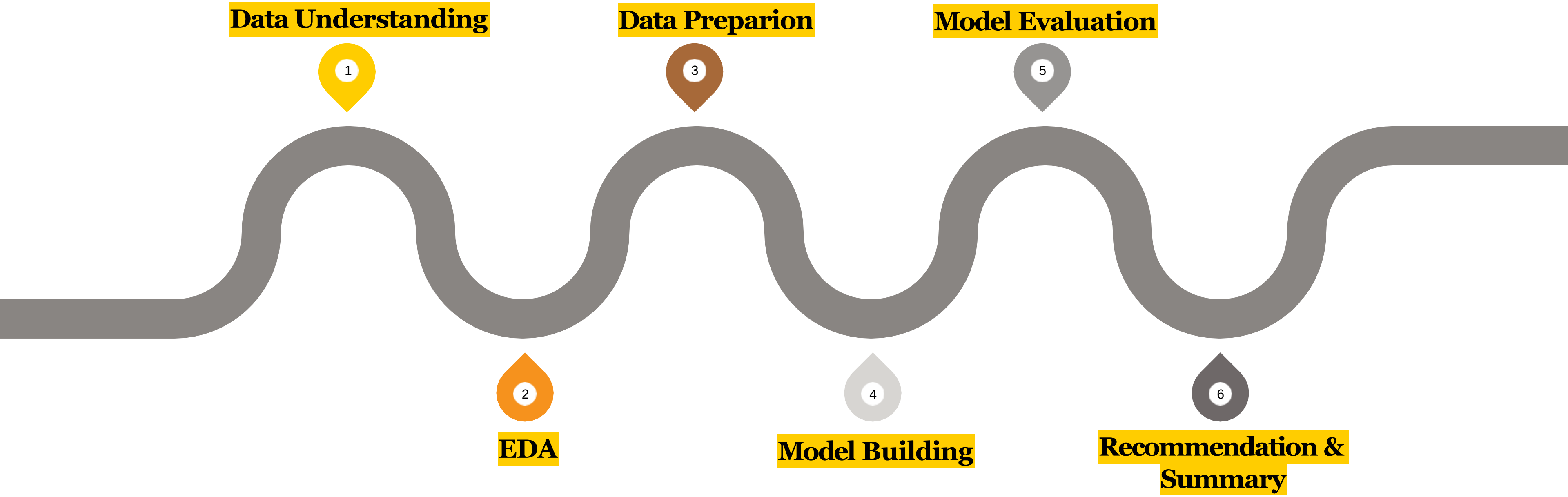
Problem :

- 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.

- Build a machine learning model for the X Education to categorized the customer with lead score to get the maximum conversion rate. The target to achieve a ballpark of the target lead conversion rate to be around 80%.

Objective: 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.

2 Roadmap


Data Understanding
1
EDA
2
Data Preparion
3
Model Building
4
Model Evaluation
5
Recommendation & Summary
6

3 This is a Data Under Standing and Approch

- The X Education provided data set with 9270 rows and 37 columns.
- The Dataset contains 9270 rows and 37 columns
- The Dataset contains 5 numerical and 27 categorical columns
- Some columns have null values which are to be handled during EDA
- The "Prospect ID" and "Lean Number" columns are not significant and to be drop during EDA
- The approach to the problem solving:
  - Data understanding with respect to objective of problem statement
  - Exploratory Data Analysis
  - Data Preparation
  - Model Building
  - Model Evaluation
  - Brief Summary

Exploratory Data Analysis

All the steps in Road Map performed in Jupyter Notebook with clear Markdown and Plots.

The insights and observations stated Markdown along with plots in Jupyter Notebook

4 Insights of EDA

Null Value Imputation and Dropped variables

- The column "TotalVisits" and "Page Views Per Visit" null values replaced by its mean
- The column "Lead Source" and "Last Activity" null values replaced by it's mode
- The variable India and Null values are approximate to 97% of total rows hence we decide to drop the Country column.
- The columns "Specialization" & "How did you hear about X Education" has a "Select" variable with 21% & 54% weighted. The "select" variable is completely different from other variables. Considering the variable "Select" and importance of columns, we decided to replace the "Select" variable with "Not declared" in "Specialization" and "Unknown" in "How did you hear about X Education".
- The null values in "Specialization" and "How did you hear about X Education" are replaced with Mode values.

4 Insights of EDA

Null Value Imputation and Dropped variables

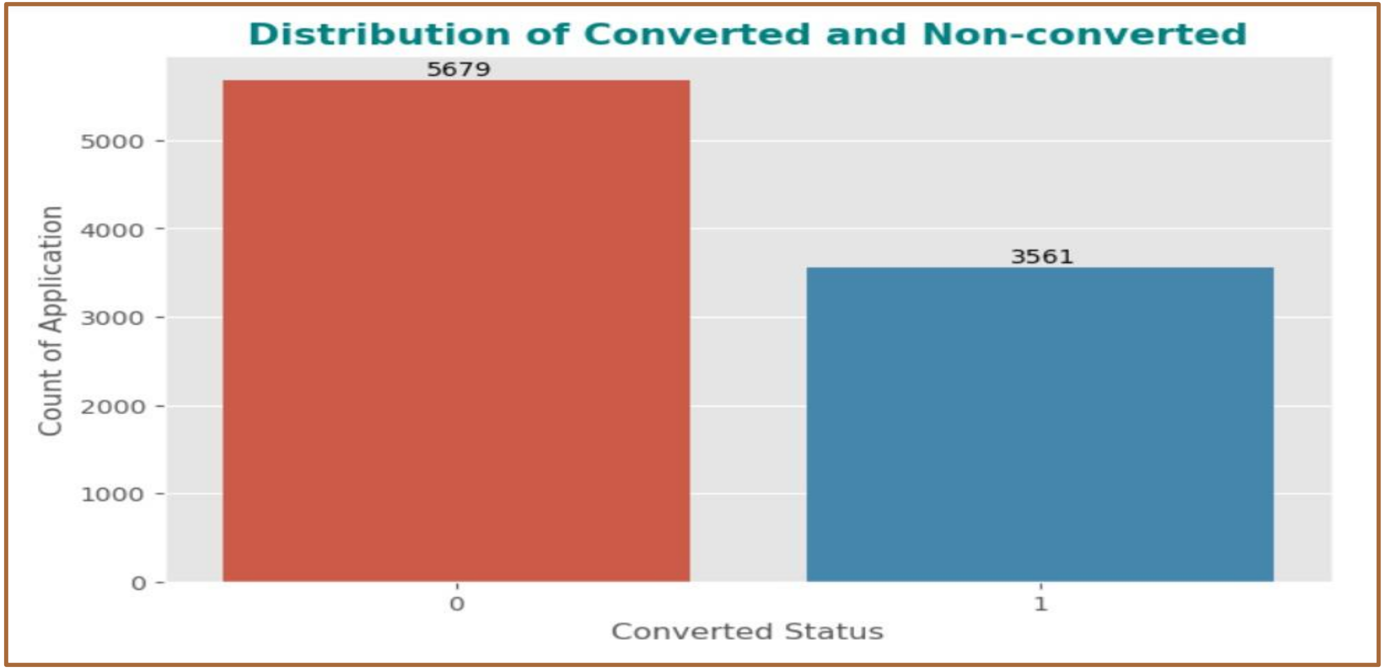
- The column "What is your current occupation" has 29.11% null values which are replaced by "Other" variable

- The count of "Housewife" and "Businessman" are not significantly high which are pushed to the variable "Others" and "Working Professional" respectevely.

- The columns 'Do Not Call','Tags','Update me on Supply Chain Content','Lead Profile', 'City','Get updates on DM Content','What matters most to you in choosing a course', 'Search', 'Magazine','Newspaper Article','X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations','Receive More Updates About Our Courses', 'I agree to pay the amount through cheque','Country' are not significant and does not provide sufficient information for problem statement, hense decided to drop the list of columns.

4 Insights of EDA

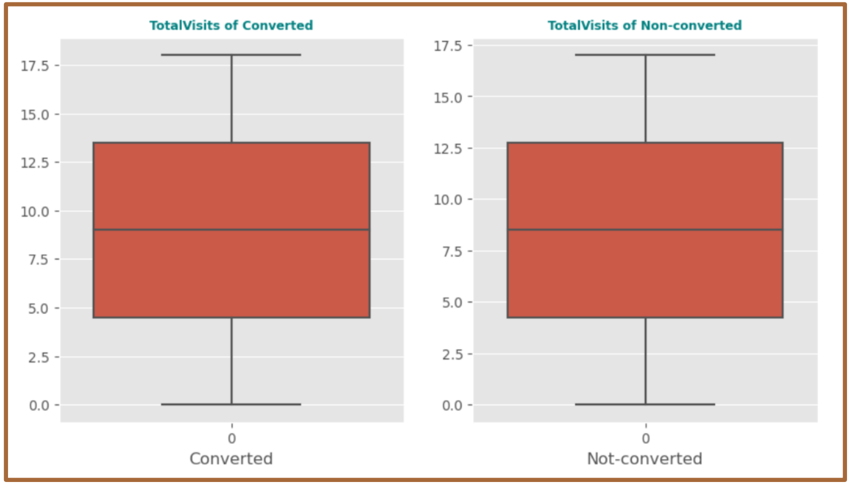
Insight from TARGET column “Converted”

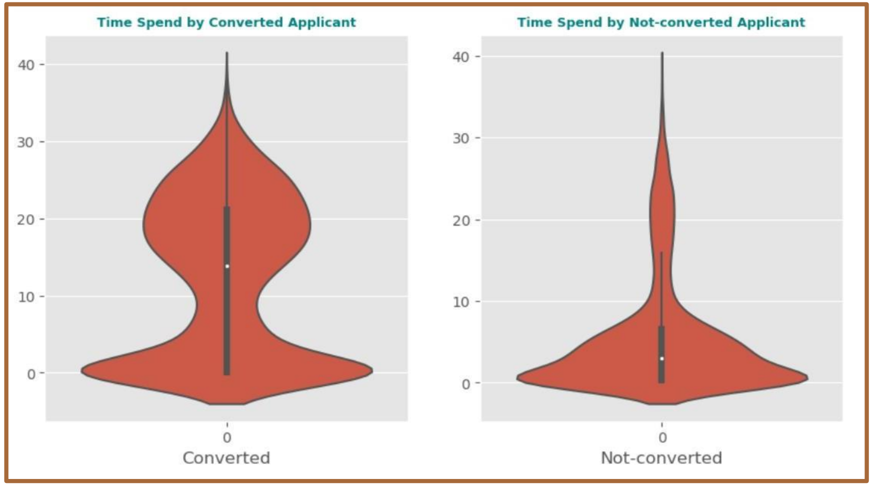


Insight: Out of 9270 application, 3561 are converted which leaves ~39% Converted rate.

4 Insights of EDA

Insight from Numterical column



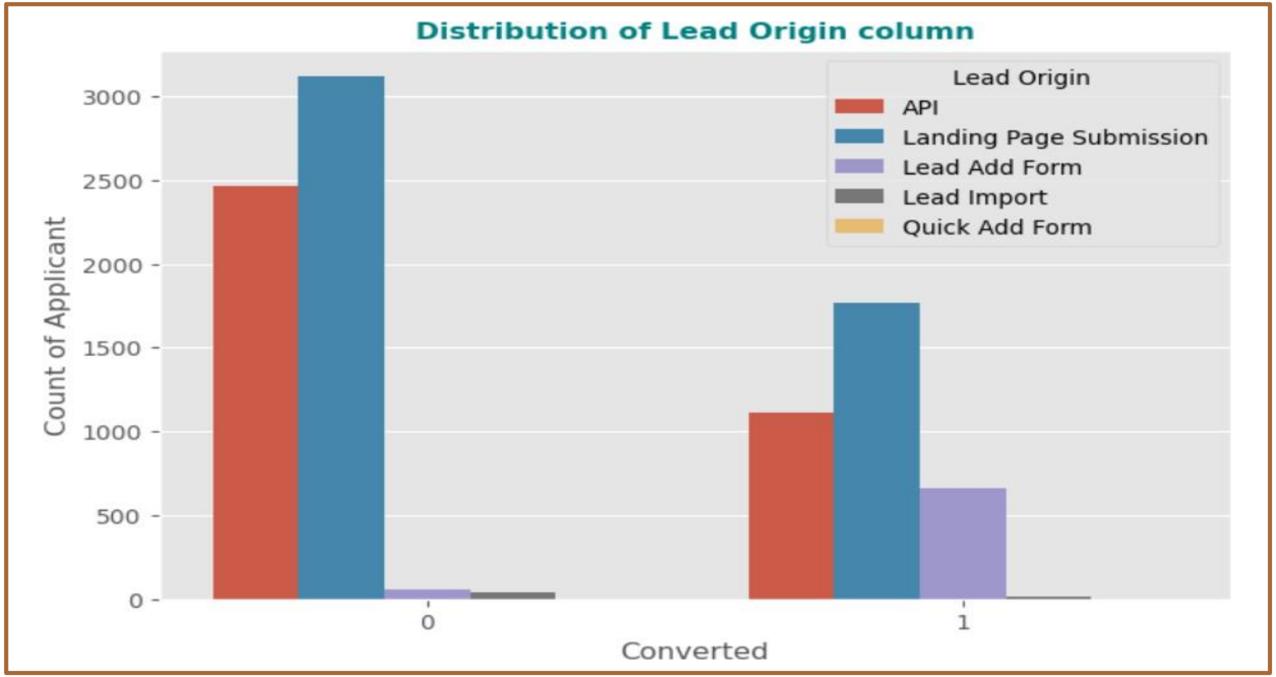


Insight: There are outlier in the "TotalVisits" but we are keeping the same since it is important data points.
Since the mean is 1 count, we can interpreted that the converted applicant visited at least once to the website.

Insight: The applicant who are not converted does not show interest in exploring course on website.
The applicant who are converted have spend average time of 15 min before enrolling.

4 Insights of EDA

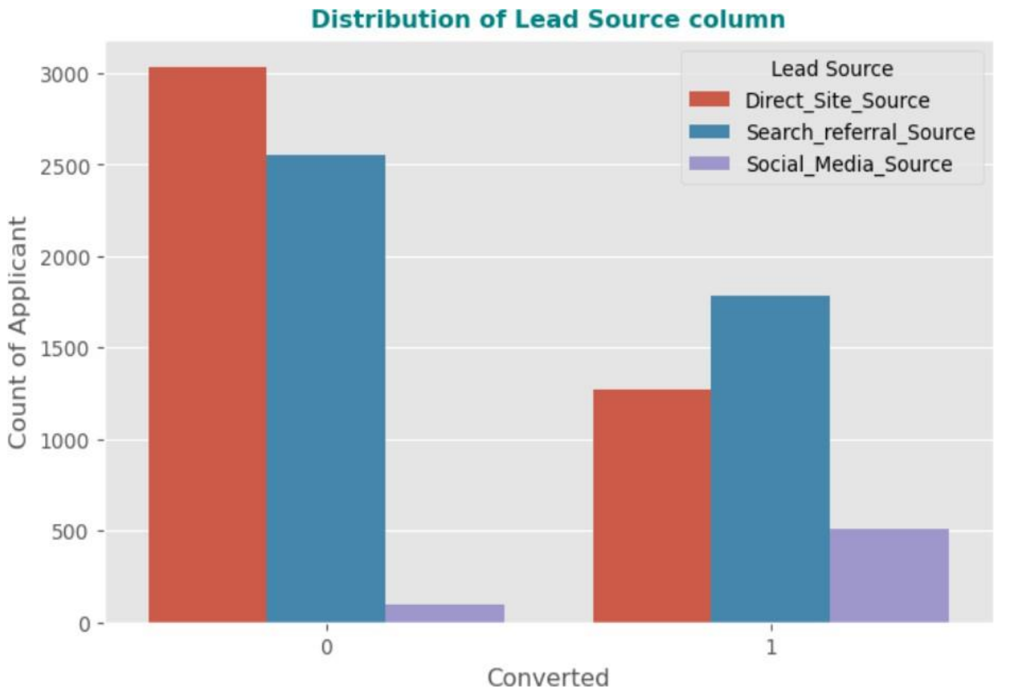
Insight from Lead Origin Catogorical columns



Insight: The "Lead Import" having highest conversion rate. API and "Landing Page Submission" conversion rate is lower compare to "Lead Import"

4 Insights of EDA

Insight from NAME_TYPE_SUITE Catogorical columns

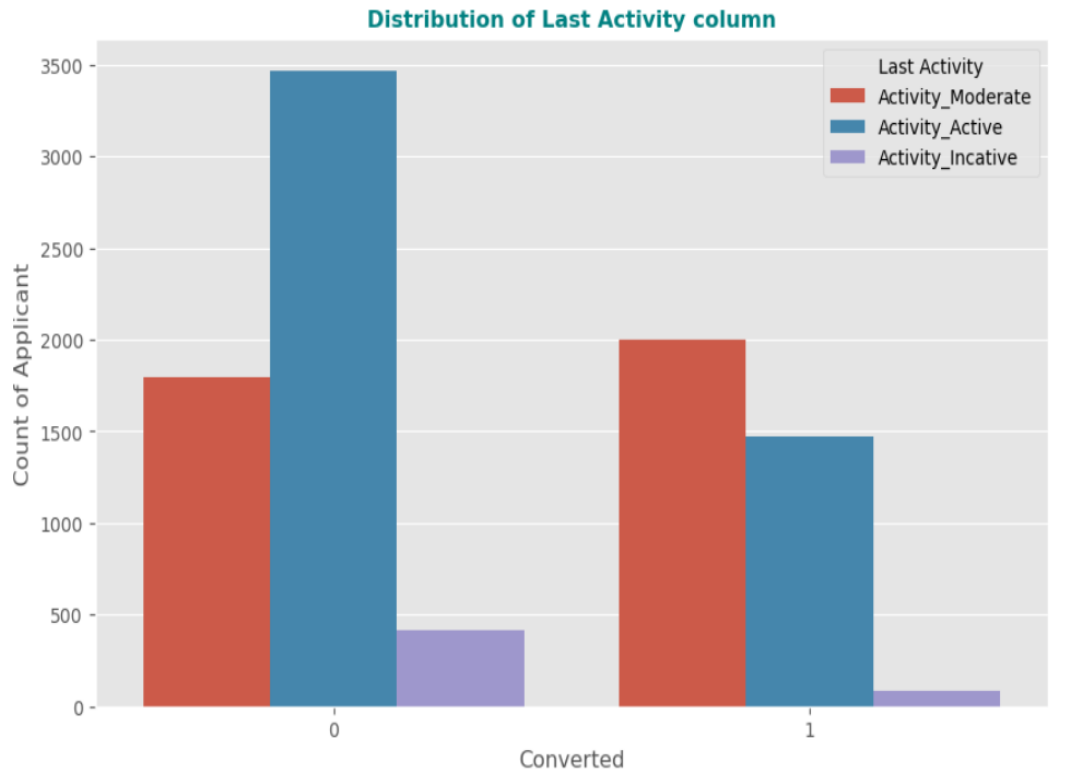


Assumption:

- The variables "Google","google", "bing","Welingak Website","Referral Sites","Organic Search" are considered as a "Search_referral_Source"
- The variable "Direct Traffic","Olark Chat","Live Chat" are considered as a "Direct_Site_Source"
- The variable "Facebook","Click2call","Social Media","blog","youtubechannel","Reference","Press_Release","Pay per Click Ads","WeLearn","welearnblog_Home","testone","NC_EDM" are considered as a "Social_Media_Source"

4 Insights of EDA

Insight from Last Activity Catogorical columns

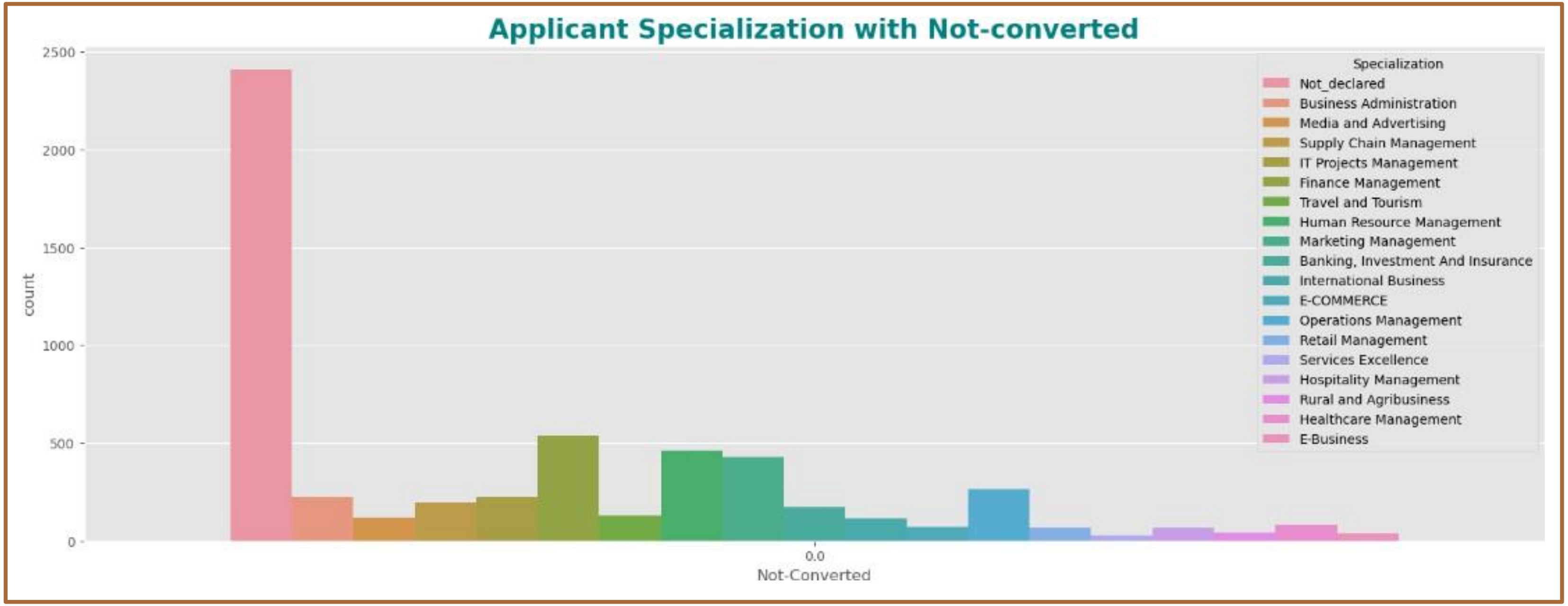


Assumption:

- The variable "Email Opened","Olark Chat Conversation","Converted to Lead" are mapped as "Activity_Active"
- The variable "SMS Sent","Page Visited on Website","Email Link Clicked","Form Submitted on Website","Had a Phone Conversation","Resubscribed to emails" are mapped as "Activity_Moderate"
- The variables "Email Bounced","Unreachable","Unsubscribed" ,"Approached upfront","View in browser link Clicked","Email Received","Email Marked Spam","Visited Booth in Tradeshow" are mapped as "Activity_Incative"

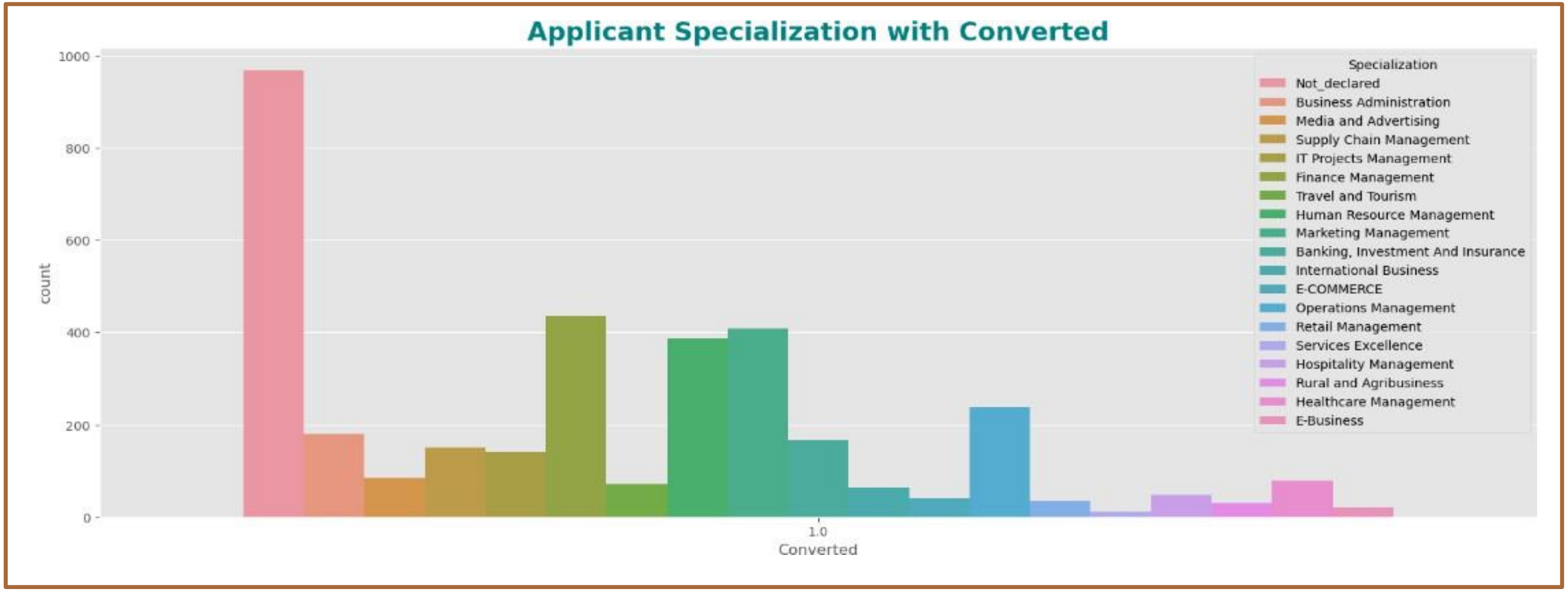
4 Insights of EDA

Insight from Specialization Catogorical columns



4 Insights of EDA

Insight from Specialization Catogorical columns

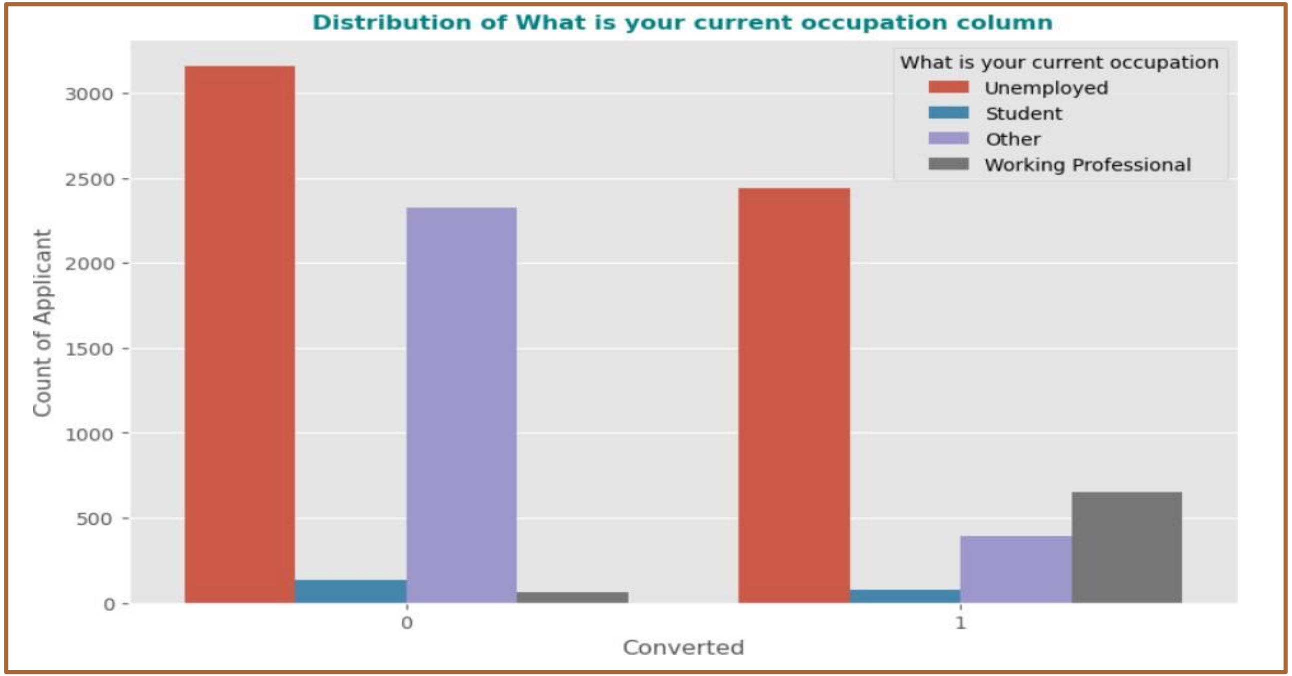


Insight:

- The null values and Select specializations are considered as Not declared
- Considering the important variables, we have not disturb the Specialization columns

4 Insights of EDA

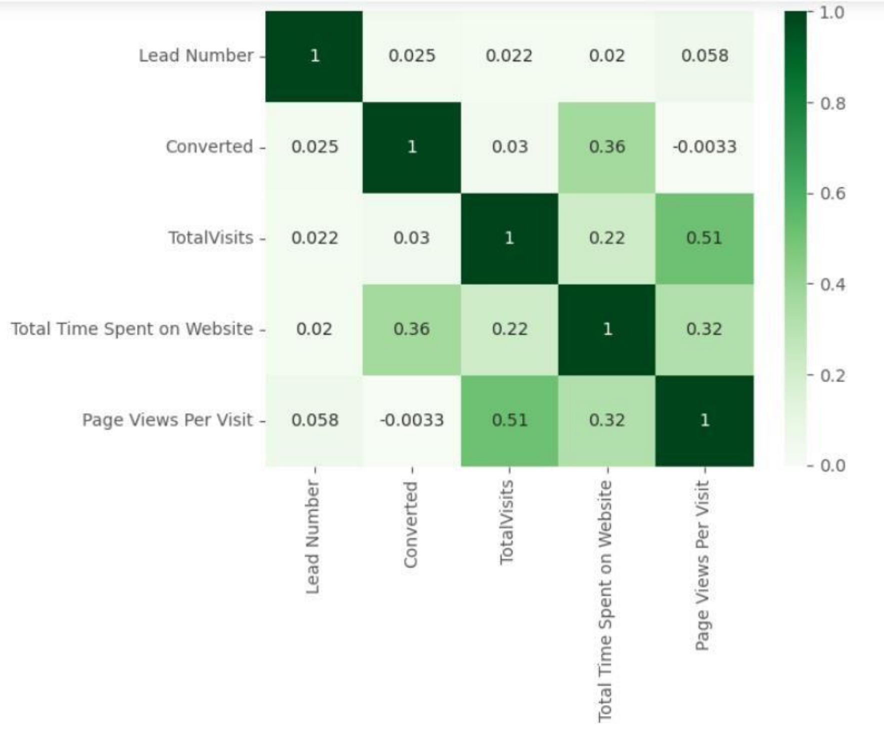
Insight from Current Occupation column



Insight: The applicant with Working Professional occupation have high conversion rate while the student have lowest.

4 Insights of EDA

Insight from Numeric to Numeric columns



Insight:
The correlation between the "Page views Per Visit" and "TotalVisits" are high (0.51)

Data Preparation

Data Preparation for Logistic Regression or any other machine learning algorithm is essential and crusial process.

5 Data Preparation

Insight from Amount Columns respect to TARGET

1) Prepared dummy variables for column “Lead Origin”, "Lead Source“, "Last Activity“, "Specialization“ and "What is your current occupation“

2) Get target/dependent variable on y and independent variables on X

3) Splitting the dataset for train and test by train_test_split

4) Scale the data set with MinMaxScaler

Model Building and Evaluation

LogisticRegression from Sklearn and Statemodel.api are used for model building

6 Model Building and Evaluation

Insight of Model-0 with Statemodel.api

In [3317]:

```
# Model_0 with all variables
lr0 = sm.GLM(y_train,(sm.add_constant(X_train)), family = sm.families.Binomial())
lr0.fit().summary()
```

Out[3317]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6435
Model Family:	Binomial	Df Model:	32
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2809.0
Date:	Sat, 12 Aug 2023	Deviance:	5617.9
Time:	20:14:39	Pearson chi2:	6.85e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3692
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7349	0.285	-2.582	0.010	-1.293	-0.177
Do Not Email	-1.3834	0.186	-7.428	0.000	-1.748	-1.018
TotalVisits	3.4279	1.827	1.877	0.061	-0.152	7.008
Total Time Spent on Website	4.1712	0.155	26.851	0.000	3.867	4.476
Page Views Per Visit	-4.1710	1.233	-3.382	0.001	-6.588	-1.754
A free copy of Mastering The Interview	-0.4239	0.094	-4.510	0.000	-0.608	-0.240
API	-3.5278	0.328	-10.744	0.000	-4.171	-2.884
Landing Page Submission	-4.6598	0.348	-13.391	0.000	-5.342	-3.978
Direct_Site_Source	1.6176	0.360	4.487	0.000	0.911	2.324
Search_referral_Source	1.4245	0.351	4.061	0.000	0.737	2.112
Activity_Active	-0.0687	0.226	-0.304	0.761	-0.511	0.374
Activity_Moderate	0.9243	0.222	4.171	0.000	0.490	1.359
Banking, Investment And Insurance	1.2738	0.205	6.214	0.000	0.872	1.676
Business Administration	0.9639	0.195	4.952	0.000	0.582	1.345
E-Business	0.9272	0.439	2.111	0.035	0.066	1.788
E-COMMERCE	1.3669	0.305	4.480	0.000	0.769	1.965
Finance Management	1.2569	0.152	8.281	0.000	0.959	1.554
Healthcare Management	1.4735	0.287	5.130	0.000	0.911	2.036
Hospitality Management	0.2204	0.327	0.673	0.501	-0.421	0.862
Human Resource Management	1.0319	0.155	6.640	0.000	0.727	1.336
IT Projects Management	1.2008	0.206	5.816	0.000	0.796	1.605
International Business	0.7880	0.269	2.933	0.003	0.261	1.315
Marketing Management	1.0134	0.150	6.760	0.000	0.720	1.307
Media and Advertising	1.2502	0.243	5.144	0.000	0.774	1.727
Operations Management	1.0800	0.178	6.067	0.000	0.731	1.429
Retail Management	0.7208	0.343	2.102	0.036	0.049	1.393
Rural and Agribusiness	1.6445	0.403	4.077	0.000	0.854	2.435
Services Excellence	0.8495	0.532	1.596	0.110	-0.194	1.893
Supply Chain Management	1.0195	0.204	5.009	0.000	0.621	1.418
Travel and Tourism	1.2277	0.256	4.788	0.000	0.725	1.730
Student	1.2191	0.237	5.145	0.000	0.755	1.684
Unemployed	1.1167	0.084	13.270	0.000	0.952	1.282
Working Professional	3.4286	0.192	17.880	0.000	3.053	3.804

Insight: The multiple variables have very high p-value, so decided to get best fit variable by Recursive Features Elimination(RFE)

6 Model Building and Evaluation

Apply Recursive Features Elimination

The RFE method gave 25 best fit columns as listed-

'Do Not Email', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit', 'API', 'Landing Page Submission', 'Direct_Site_Source', 'Search_referral_Source', 'Activity_Moderate', 'Banking, Investment And Insurance', 'Business Administration', 'E-COMMERCE', 'Finance Management', 'Healthcare Management', 'Human Resource Management', 'IT Projects Management', 'Marketing Management', 'Media and Advertising', 'Operations Management', 'Rural and Agribusiness', 'Supply Chain Management', 'Travel and Tourism', 'Student', 'Unemployed', 'Working Professional'

6 Model Building and Evaluation

Model Building Process

- Selected 25 best fit columns for Model Building Process
- Logistic Regression model build on Statemodel.api libraries with GLM
- Models performance evaluated by p-value and VIF
- Model evaluation process iterated with 13 various model converge to the p-value below 0.05 and VIF value below 5.0
- VIF of "Search_referral_Source" column is 5.11 which is very near to 5.0, hence decided to keep.

6 Model Building and Evaluation

Final Model Details

```python
# Build the final model for further testing and evaluation
X_train_sm = sm.add_constant(X_train[col])
final_model = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = final_model.fit()
res.summary()
```

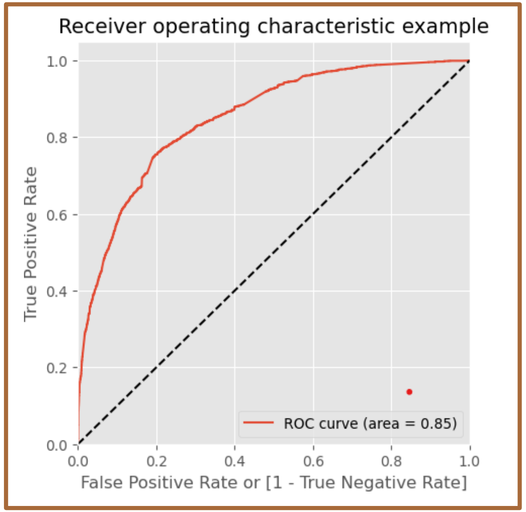
	Features	VIF
5	Search_referral_Source	5.11
4	Direct_Site_Source	3.55
2	Page Views Per Visit	3.22
11	Unemployed	2.70
3	API	2.24
1	Total Time Spent on Website	2.11
6	Activity_Moderate	1.75
12	Working Professional	1.22
8	Finance Management	1.18
0	Do Not Email	1.10
7	Banking, Investment And Insurance	1.06
10	Student	1.05
9	Healthcare Management	1.03

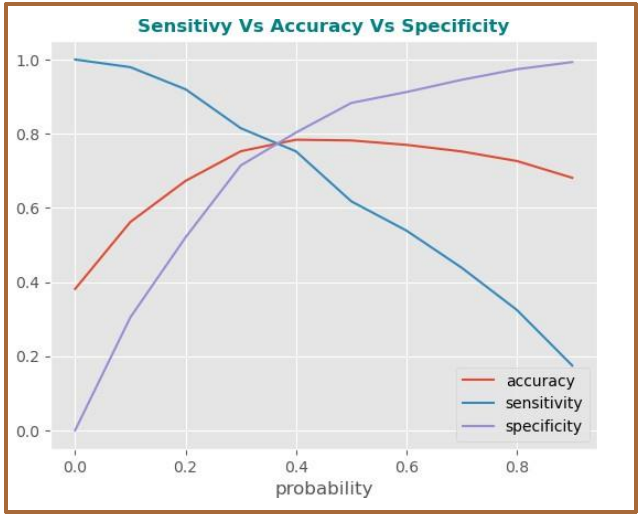
Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6454
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2983.9
Date:	Sat, 12 Aug 2023	Deviance:	5967.8
Time:	20:14:42	Pearson chi2:	6.48e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3341
Covariance Type:	nonrobust		

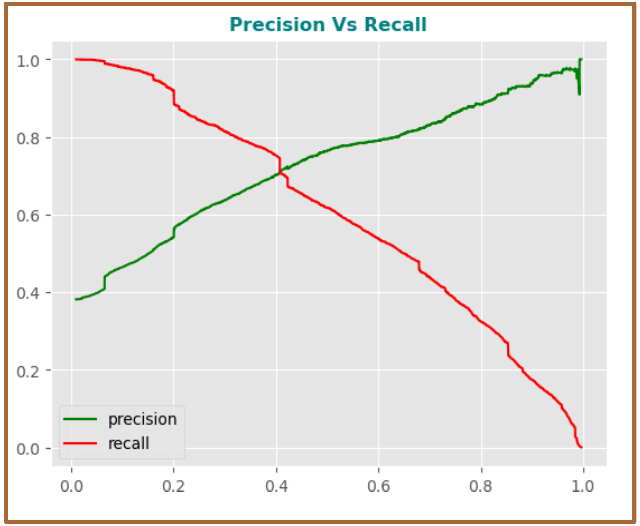
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5317	0.166	-3.202	0.001	-0.857	-0.206
Do Not Email	-1.1719	0.148	-7.904	0.000	-1.463	-0.881
Total Time Spent on Website	3.8418	0.148	25.995	0.000	3.552	4.131
Page Views Per Visit	-9.1773	1.135	-8.086	0.000	-11.402	-6.953
API	0.2005	0.078	2.579	0.010	0.048	0.353
Direct_Site_Source	-2.3353	0.159	-14.653	0.000	-2.648	-2.023
Search_referral_Source	-2.0718	0.166	-12.446	0.000	-2.398	-1.746
Activity_Moderate	1.0091	0.066	15.375	0.000	0.881	1.138
Banking, Investment And Insurance	0.3476	0.172	2.023	0.043	0.011	0.684
Finance Management	0.2435	0.106	2.294	0.022	0.035	0.452
Healthcare Management	0.5293	0.256	2.071	0.038	0.028	1.030
Student	1.2911	0.222	5.805	0.000	0.855	1.727
Unemployed	1.2823	0.082	15.706	0.000	1.122	1.442
Working Professional	3.6681	0.186	19.688	0.000	3.303	4.033

6 Model Building and Evaluation

Model Evaluation







Insight: T The trade of between the precision and recall at approximate at 0.4 which considered as threshold for predicting the dependent variable

6 Model Building and Evaluation

Evaluation Matrix

```
Classification Report of Train Set :

              precision    recall  f1-score   support

           0       0.84      0.80      0.82      4002
           1       0.70      0.75      0.73      2466

    accuracy                           0.78      6468
   macro avg       0.77      0.78      0.77      6468
weighted avg       0.79      0.78      0.79      6468



Classification Report of Test Set :

              precision    recall  f1-score   support

           0       0.83      0.81      0.82      1677
           1       0.72      0.75      0.73      1095

    accuracy                           0.79      2772
   macro avg       0.78      0.78      0.78      2772
weighted avg       0.79      0.79      0.79      2772
```

Reccomendation & Summary

7 Recommendation & Summary

Recommendations

1. Total Time Spent on Website: The feature "Total Time Spent on Website" has a positive coefficient in the model, indicating that leads who spend more time on the website are more likely to convert. This suggests that enhancing website engagement and content quality could be beneficial. Consider optimizing the user experience, providing valuable content, and ensuring clear calls to action to keep leads engaged and interested.

2. Direct Traffic Source: Leads coming from the "Direct_Site_Source" have a negative impact on conversion. Focus on improving the user experience for visitors arriving directly to website. Implement user-friendly navigation, intuitive design, and personalized content to increase their likelihood of converting.

3. Search and Referral Sources: Similar to direct traffic, leads from "Search_referral_Source" also show a negative impact on conversion. Optimize the search engine visibility and referral sources to ensure that the content aligns with the leads' interests. This could involve improving the strategy and fostering partnerships with relevant referral websites.

4. Working Professionals and Students: The categories "Working Professional" and "Student" have positive coefficients, indicating a higher likelihood of conversion for these groups. Tailor marketing efforts to resonate with their needs and preferences. Highlight how offerings can benefit them specifically, addressing their pain points and motivations.

5. Do Not Email: The "Do Not Email" feature negatively impacts conversion. This implies that leads who opt out of receiving emails are less likely to convert. While respecting privacy preferences, try to provide value through email communication. Send personalized and relevant content that nurtures leads and keeps them engaged with the brand.

6. Improve Page Views Per Visit: The negative coefficient for "Page Views Per Visit" suggests that too many page views might overwhelm or confuse leads. Work on streamlining the user journey and ensuring that each page visit provides clear and relevant information. Use data-driven insights to optimize the website's layout and content.

7 Recommendation & Summary

Summary

Lead Score: The "Lead Score" can serve as a useful tool for ranking and prioritizing leads. Allocate resources more efficiently by focusing on leads with higher scores. The lead score generated by 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.

The results indicate that the logistic regression model exhibited consistent performance across the training and test datasets. With an accuracy of approximately 78.37% on the training set and 78.64% on the test set, the model demonstrates a balanced ability to classify instances correctly, maintaining its efficacy when encountering new, unseen data.

Accuracy: The accuracy represents the percentage of correctly classified instances out of the total instances. In our case, the accuracy is around 78.37% for the training set and 78.64% for the test set. These accuracy values are quite similar, indicating that the model is generalizing reasonably well to unseen data.

Precision: Precision measures the proportion of true positive predictions among all positive predictions made by the model. Higher precision indicates fewer false positives. The model achieved a precision of around 70.19% on the training set and 72.16% on the test set for the positive class. These precision values are acceptable and show that when the model predicts the positive class, it's correct around 70-72% of the time.

Recall / Sensitivity: Recall (also known as sensitivity) measures the proportion of true positive predictions among all actual positive instances. It's an indicator of how well the model is capturing positive instances. The model achieved a recall of about 75.22% on the training set and 74.79% on the test set for the positive class. These values indicate that the model is capturing around 75% of the actual positive instances.

Specificity: Specificity measures the proportion of true negative predictions among all actual negative instances. It's an indicator of how well the model is identifying negative instances. The model achieved a specificity of 80.0% on the training set and 81.0% on the test set. These values suggest that the model is good at correctly identifying negative instances.

Overall, model seems to have balanced performance across various metrics on both the training and test sets. The small differences between the training and test set metrics indicate that the model is not overfitting.

Thanks!