Lead Scoring Assignment
Group Case Study

GROUP MEMBERS :- VIỆT HOÀNG NGUYỄN, TEJAS P, KHOZEMA GOODLUCK

Problem Statement

- X Education sells online courses to industry professionals. Company gets leads from various sources like social media marketing, references etc.

- Once these leads are acquired, employees from the sales team get in touch with the lead to convince them for enrolling in course. However, conversion rate for leads is very low (about 30%)

- Company wants to identify most potential leads (“Hot Leads”) from all leads received. Sales team can approach only to “Hot Leads” to convince them for enrolling on course. This way sales team can devote more time to potential leads resulting in higher lead conversion rate

Objective

- Develop a machine learning model to calculate “Lead Score” based on various paraments and classify positive leads based on cut-off criteria.
- Identify important parameters and their effect on “Lead Score”
- As this is a classification problem “Logistic Regression Model” is suitable machine learning model

Analysis Approach

1 • Understanding Data Structure

2 • Handling Null Values

3 • Exploratory Data Analysis

4 • Data Preprocessing

5 • Automatic Feature Selection

6. • Manual Model Training

7. • Find Probability Threshold

8. • Model Evaluation

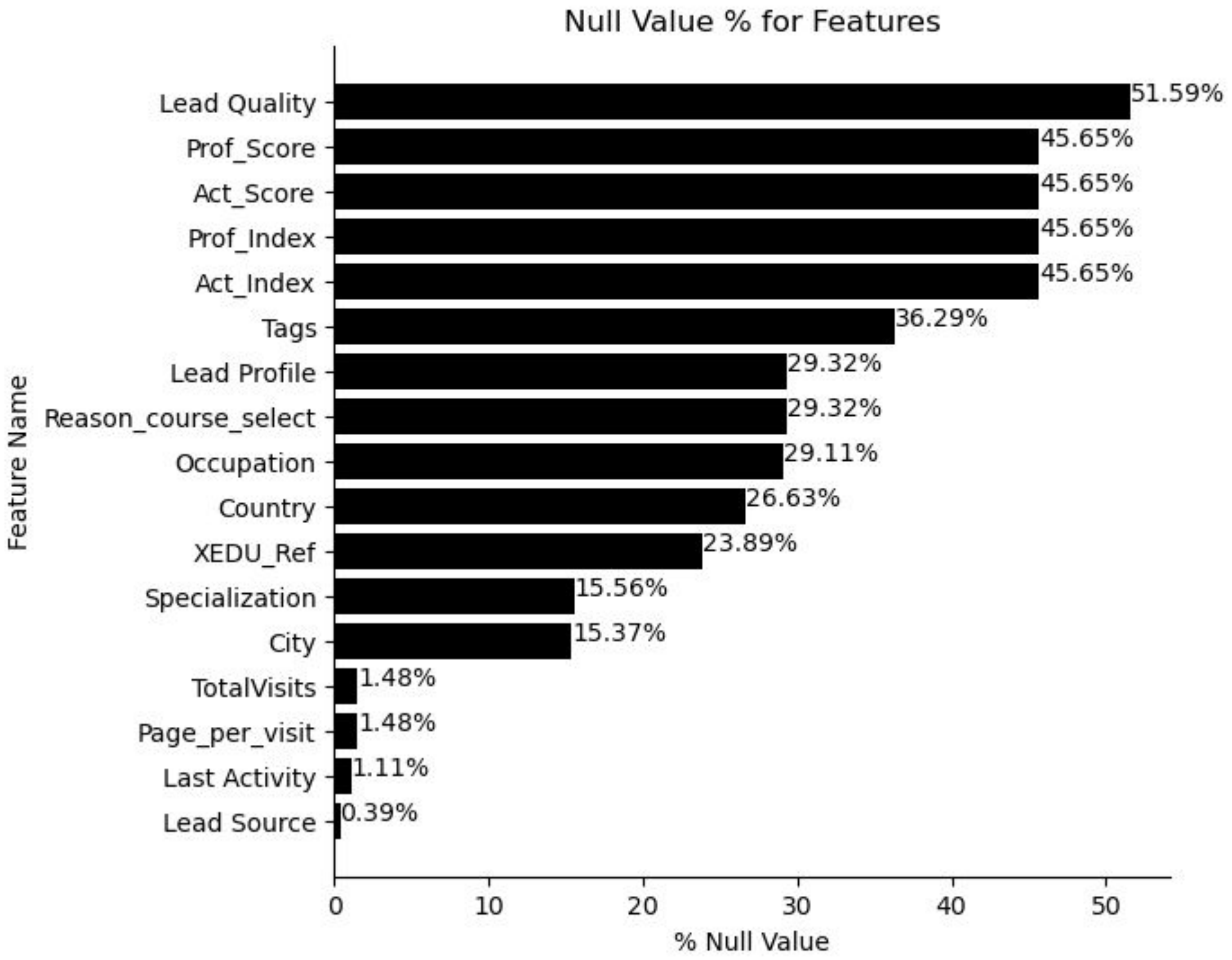
9. • Model Interpretation

Understanding Data Structure

- Dataset is having 9240 Records with 37 Features
- Target Variable is “Converted”
- Some of the features are having null values
- Many Columns are having very long descriptive names . Which Needs to be converted to short names for easy handling
- Prospect ID" and "Lead Number" are unique identifier for each record . These column are not useful for analysis and can be dropped

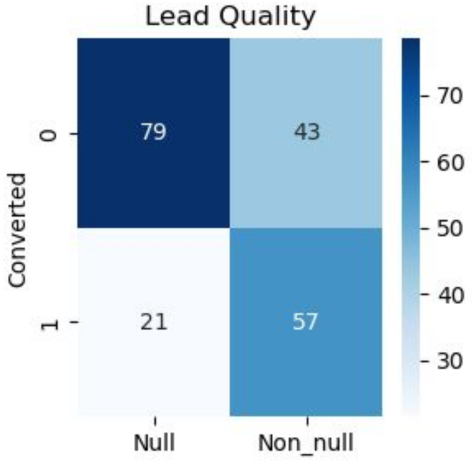
Column Name	Revised Colum Name
Do Not Email	DN_Email
Do Not Call	DN_Call
Total Time Spent on Website	Web_Time
Page Views Per Visit	Page_per_visit
How did you hear about X Education	XEDU_Ref
What is your current occupation	Occupation
What matters most to you in choosing this course	Reason_course_select
Newspaper Article	News_Article
X Education Forums	X_Forum
Digital Advertisement	Digi_Adv
Through Recommendations	Recommendation
Receive More Updates About Our Courses	updt_require
Update me on Supply Chain Content	updt_SCcontent
Get updates on DM Content	updt_DM
Asymmetrique Activity Index	Act_Index
Asymmetrique Profile Index	Prof_Index
Asymmetrique Activity Score	Act_Score
Asymmetrique Profile Score	Prof_Score
I agree to pay the amount through cheque	Chq_pymnt
a free copy of Mastering The Interview	Free_Copy
Last Notable Activity	Notable_Activity

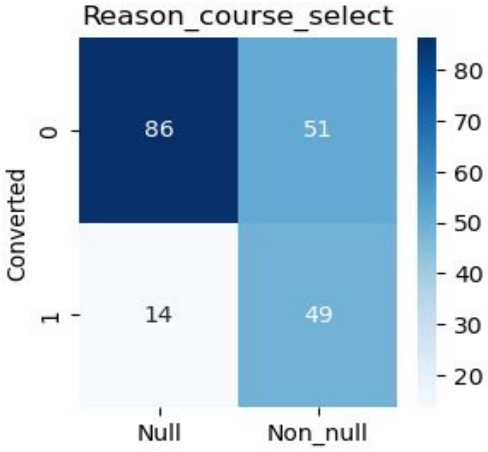
Handling Null Values

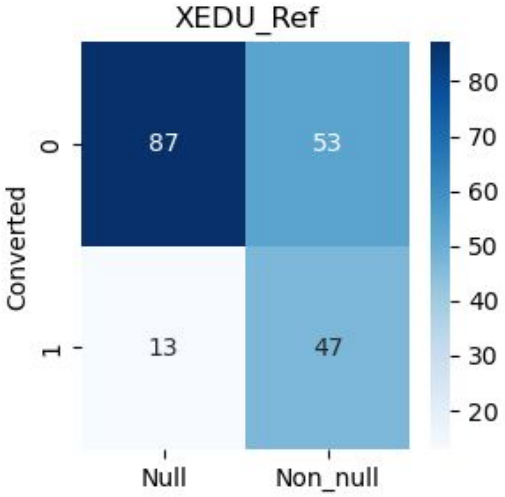


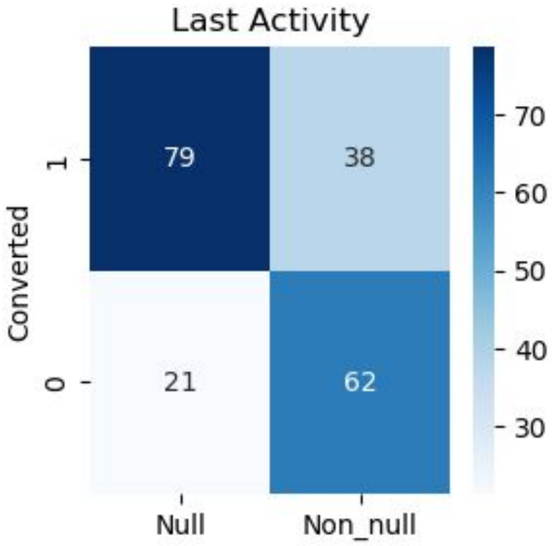
Handling Missing Values:- Impute Null Values with “Select” to indicate value is missing

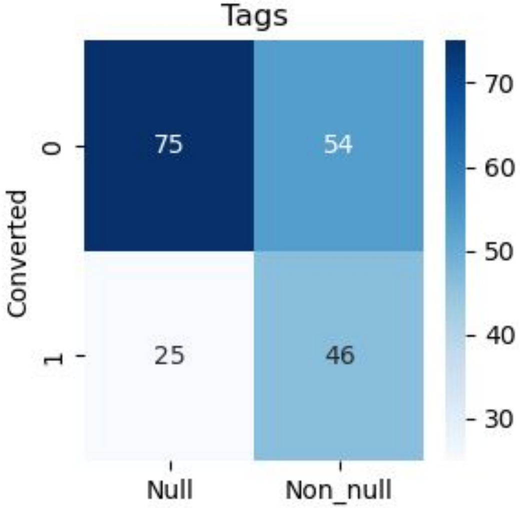
- For all columns below, Lead Conversion to 0 is very high for missing values. If value is missing then the chances of lead conversion to negative is very high.
- To indicate missing value, missing values are replaced with “Select”

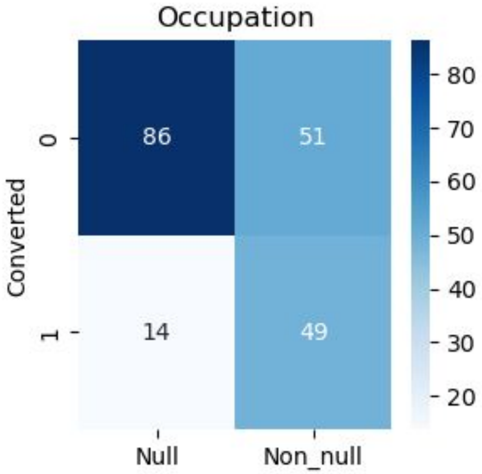


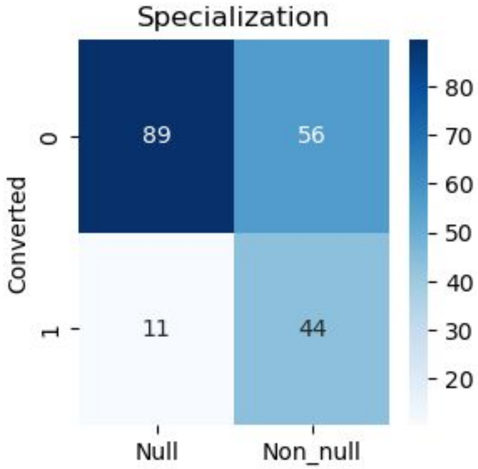






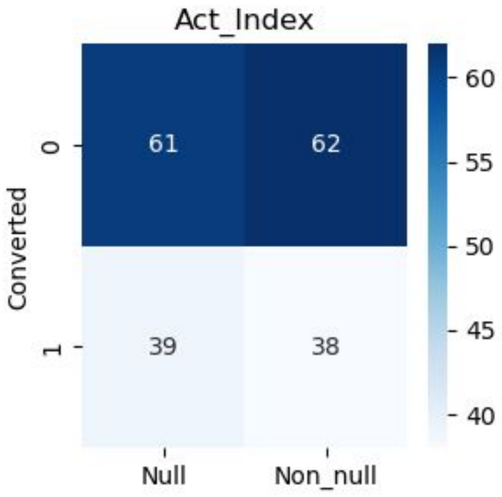


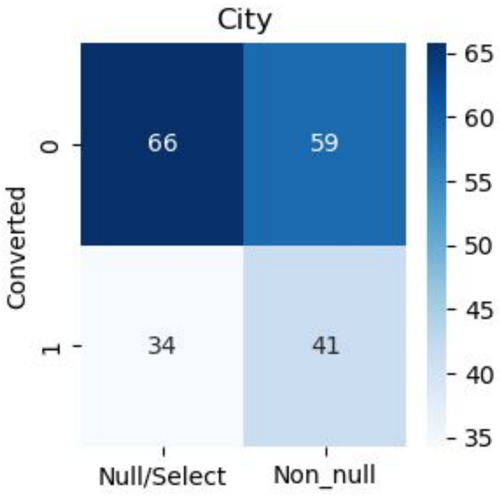


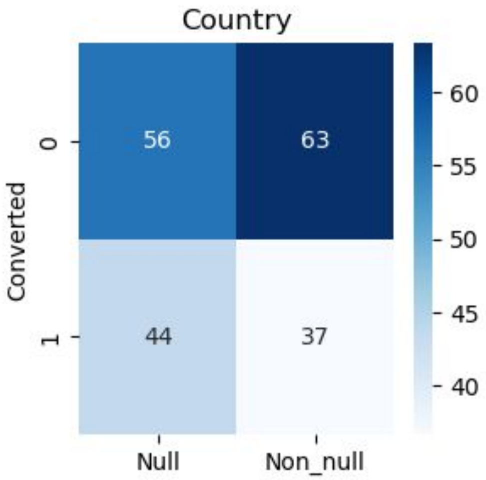


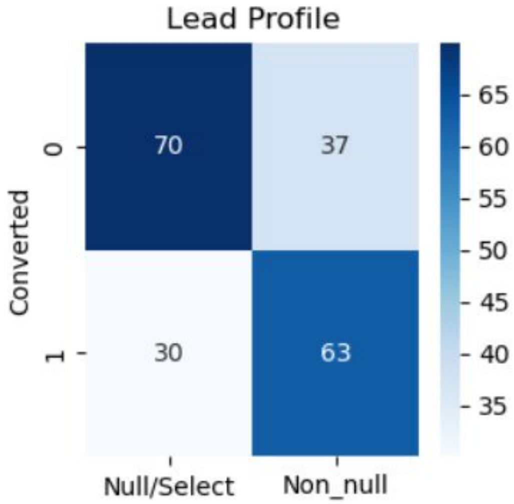
Handling Missing Values:- Drop columns

- For all columns missing values are very high.
- % Distribution for missing and non missing value is similar so these features does not seems important from analysis
- These columns can be **dropped** from analysis
- Act_Profile,Act_Score,Profile_Score column have missing values for same record where Act_Index values is missing . These columns can also be dropped similar to Act_Index



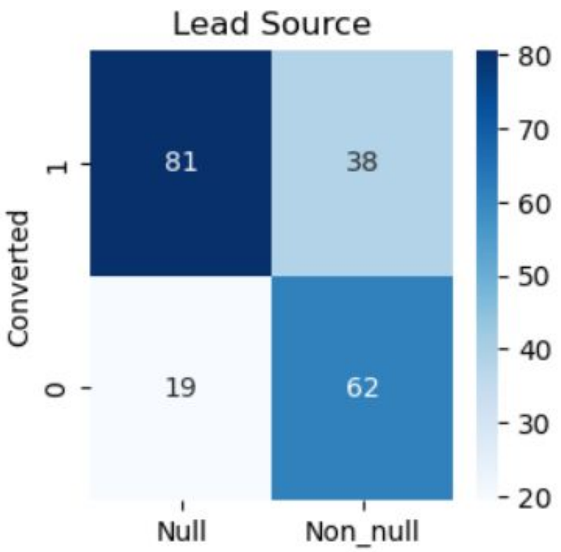






Handling Missing Values:- Replace Missing Values with Statistical Measure

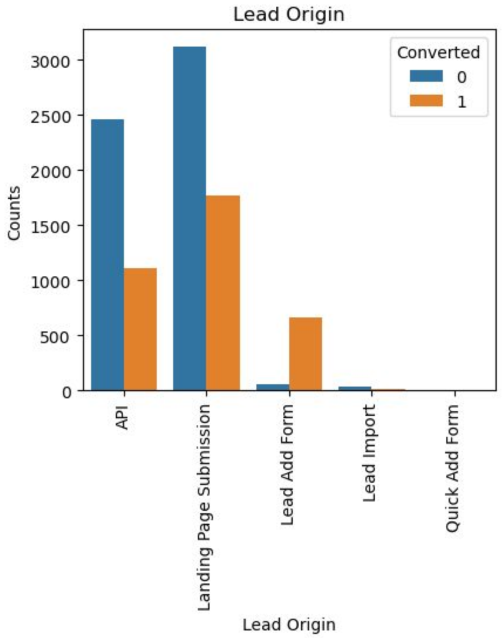
- Missing Values in “Total Visit” and “Page_per_visit” are less than 1.5% . Missing values in both these numerical features to be **imputed with median value**
- Lead Source :-
  - Imbalace of target variable in “null” and “non null” values is high so presence of null value can be important factor for identifying trends.
  - “Lead Source” and “Lead Origin” seems to be related features. Most frequent value in “Lead Origin” is “Landing Page Submission” . So values in Leads Souce can be **imputed with mode values** from all records where “Lead Origin: is “Landing Page Submission”

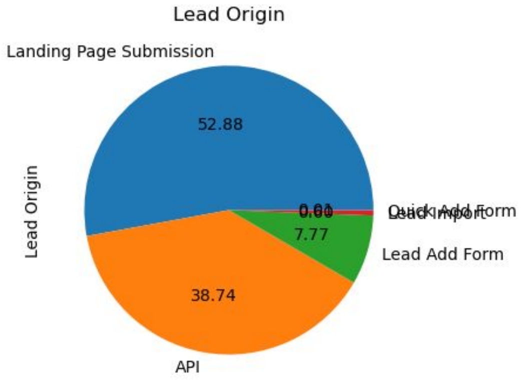


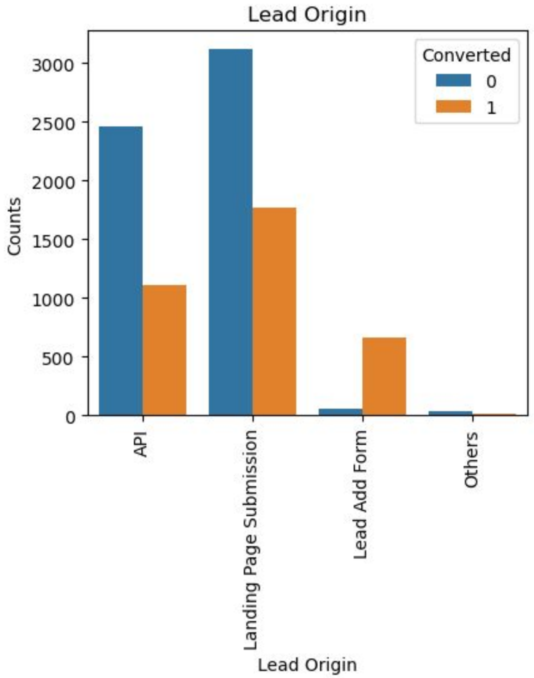
Exploratory Data Analysis

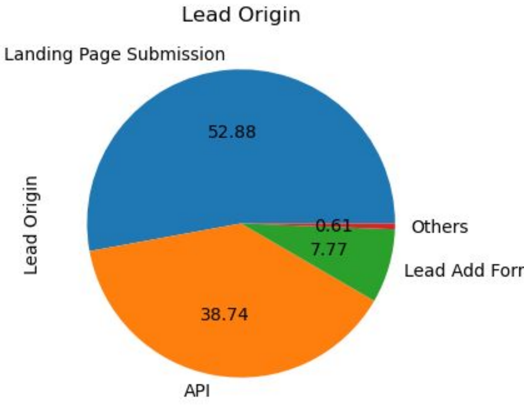
- Exploratory Data Analysis :
  - Handling irregular Values / Data Format
  - Handling Outliers
  - Segmented Analysis based on target variable (“Converted”)
  - Data Visualization and understanding data pattern

Lead Origin





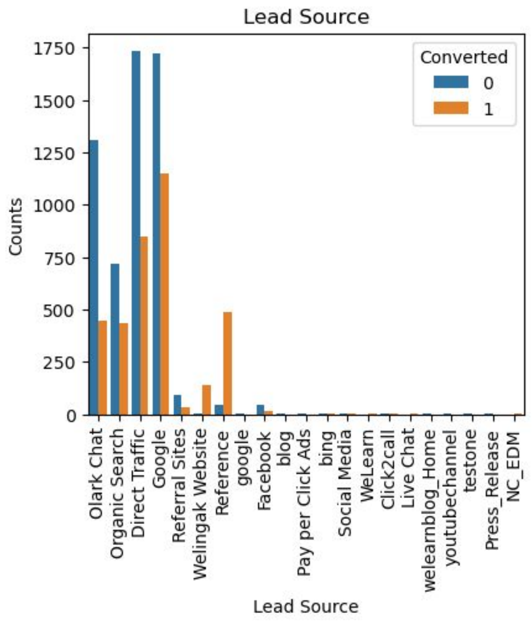


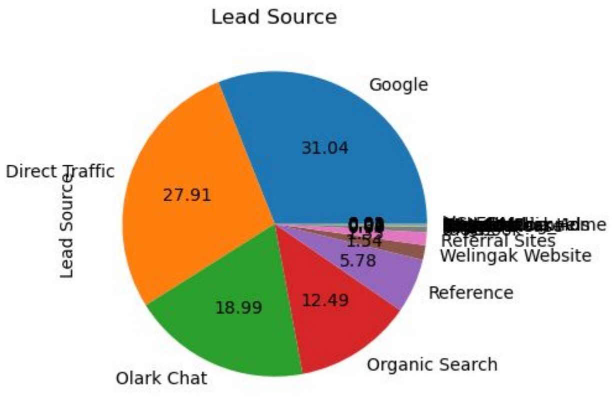


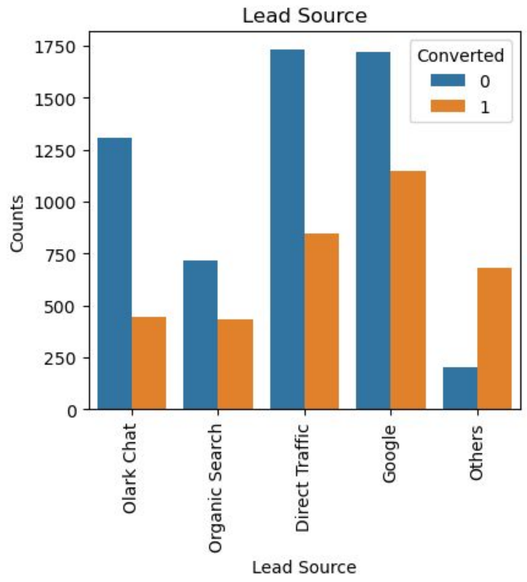
Observations:-

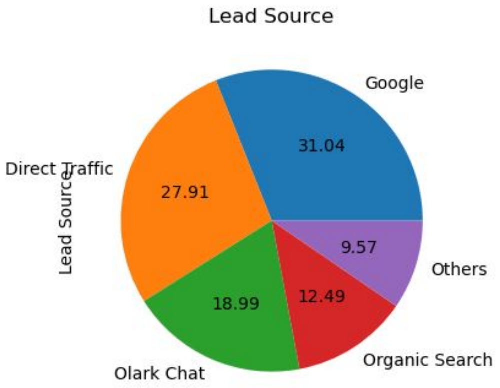
- Majority of Leads are from "API" and "Landing Page Submission“
- "Lead Import" and "Quick Add Form" are very low counts . These categories can be clubbed under other categories
- For Category **"Lead Add Form"** probability of lead **turning positive is high**

Lead Source





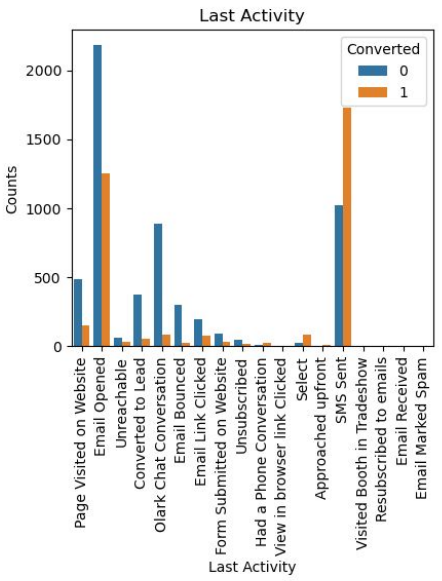


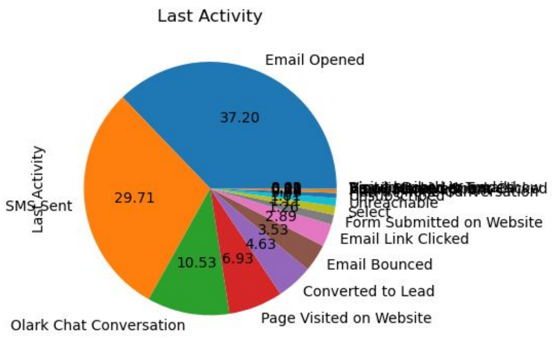


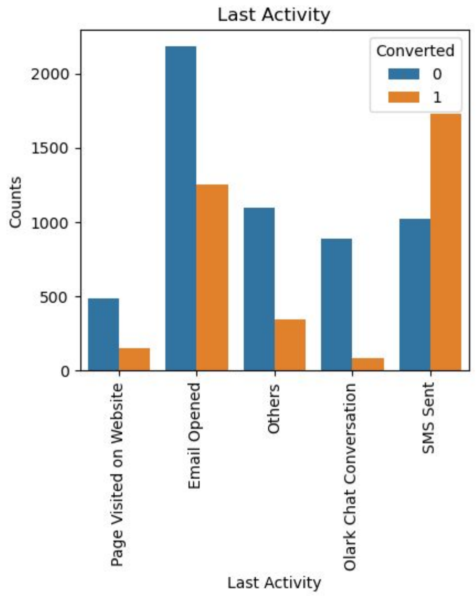
Observation:-

- There are many values with count less than 6% . These values with value count <6% can be clubbed to category "Others“
- If **Lead Source** is **"Reference"** probability of lead turning positive is very high

Last_Activity



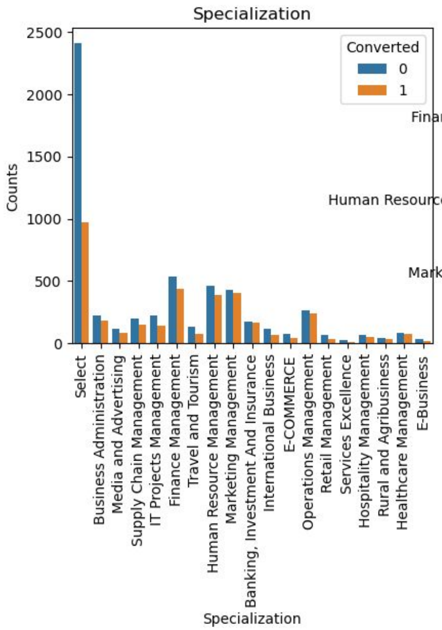


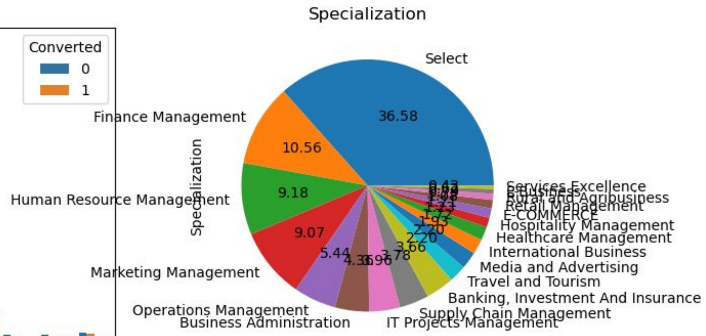


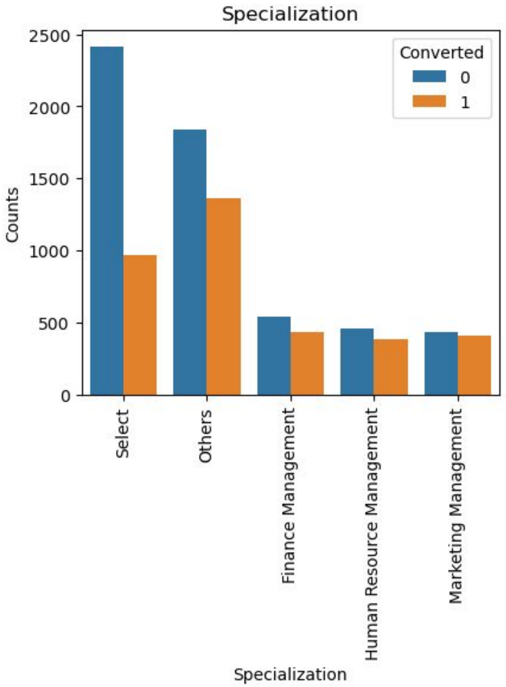
Observation:-

- There are many categories with low values counts. Thes categories can be clubbed to “others” categories
- For category “SMS Sent” , chance of turning lead to positive is high

Specialization



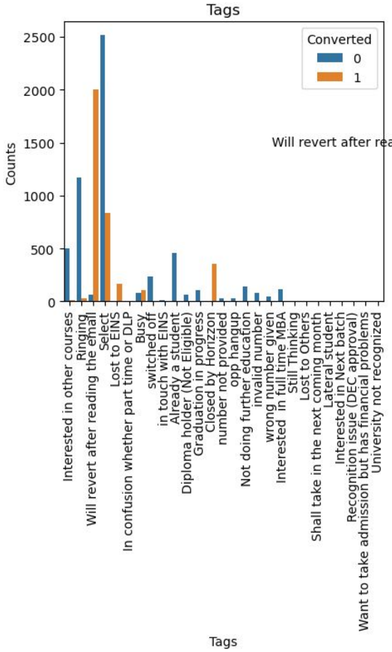


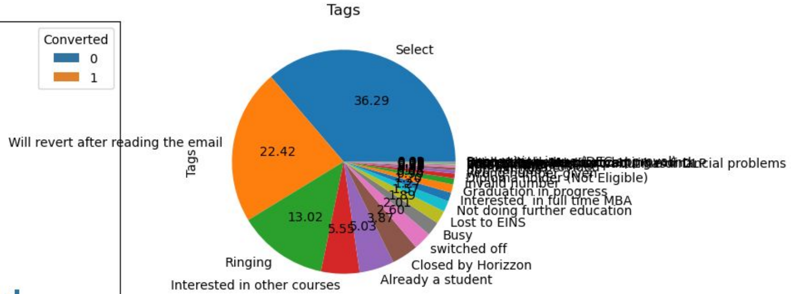


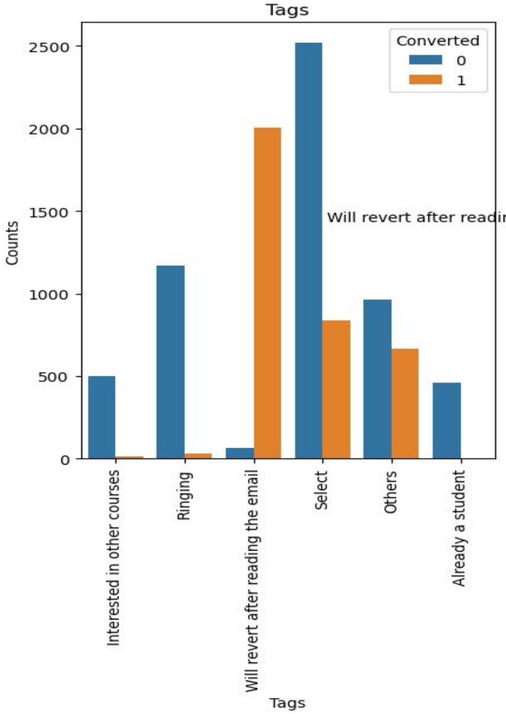
Observation:-

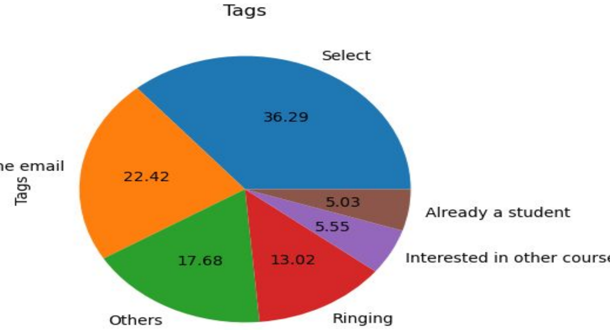
- There are many values with count less than 5% . These values with value count <6% can be clubbed to category "Others“
- Category "Select" is very high compared to other categories . This category has highest negative rate compared to others.
- Detail **missing** in category ("Select") can be significant factor to indicate that **lead will turn negative**

Tags





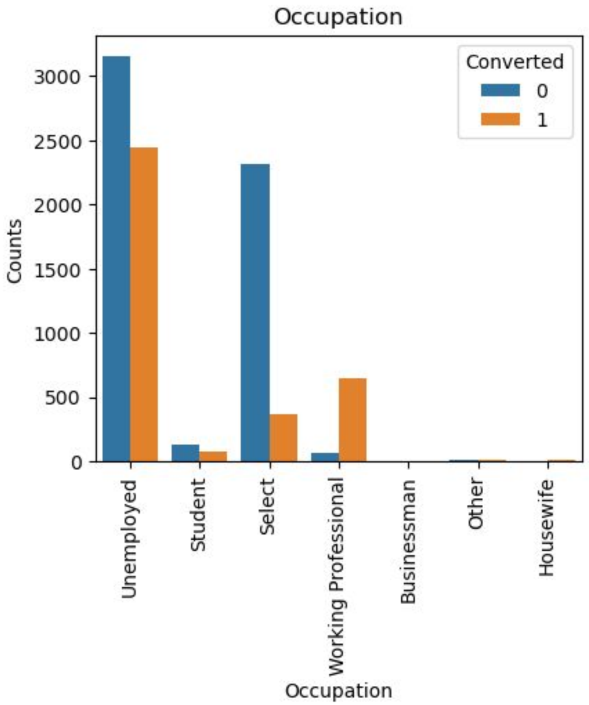


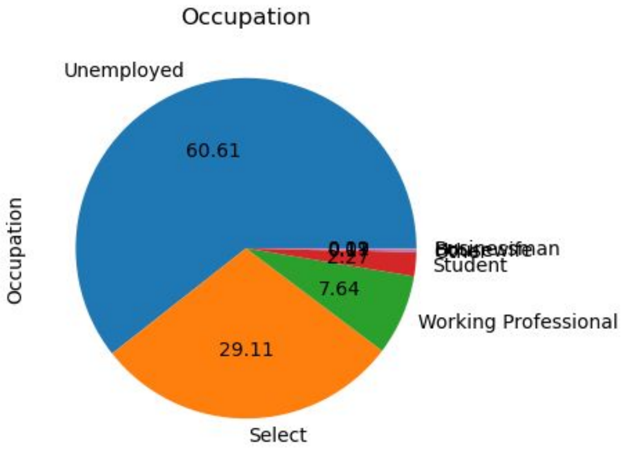


Observation:-

- Category **"Select"** i.e. missing is high compared to other categories . Probability of lead **turning negative is high**
- Category **"Closed by Horizon" or "Will revert after reading the email"** - probability of lead **turning positive is high**
- Category **"Ringing"** , probability of lead **turning negative is very high**
- Other Categories with count values less than 5% can be clubbed into "Others" category

Occupation

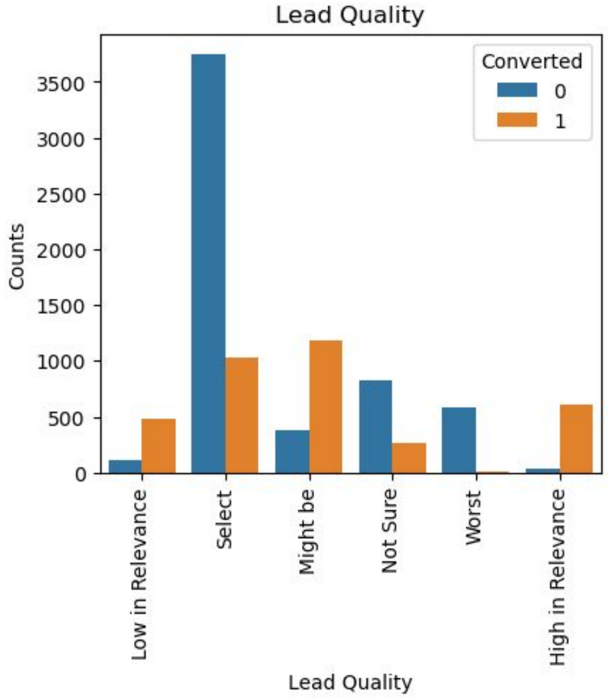


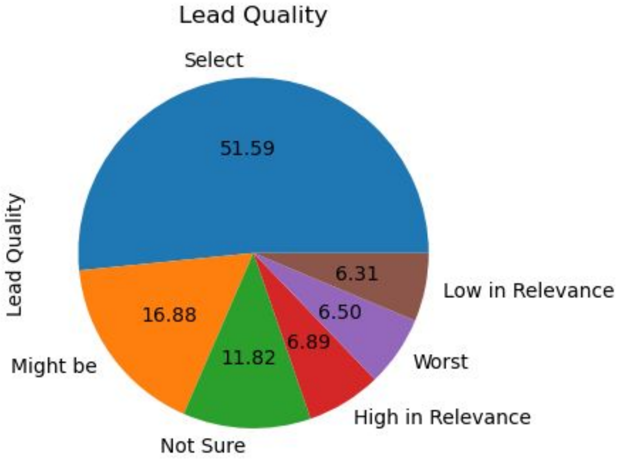


Observation:-

- For “Working Professional” probability of lead turning to positive is very high

Lead Quality

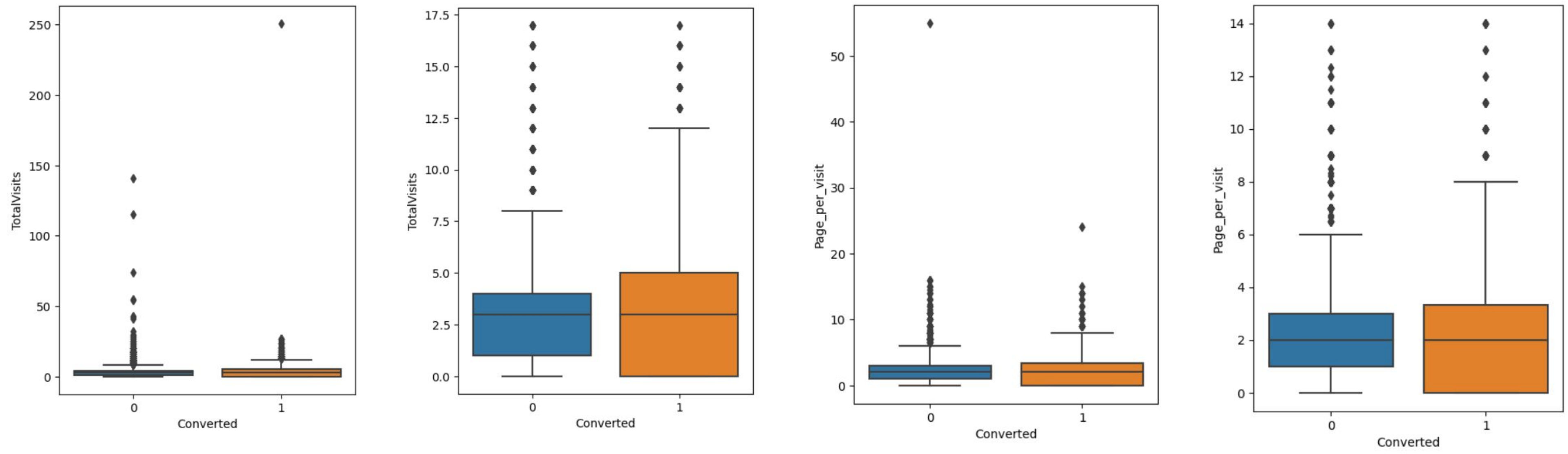


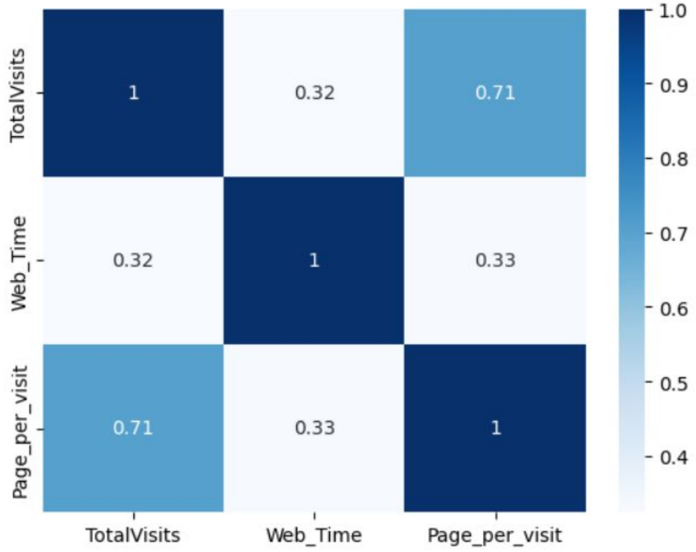


Observation:-

- There are High Number of records with "Lead Quality" **category "Select" i.e. value is missing** . Probability of lead **turning negative is high** if "Lead Quality" value is "Select" i.e. missing
- For "Lead Quality" category **"Not Sure" or "Worst"** ,probability of lead **turning negative are high**
- If "Lead Quality" values are **"High in Relevance" ,"Low in Relevance", "Might be"** - probability of lead **turning positive is high**

Total Visits, Page_per_visit

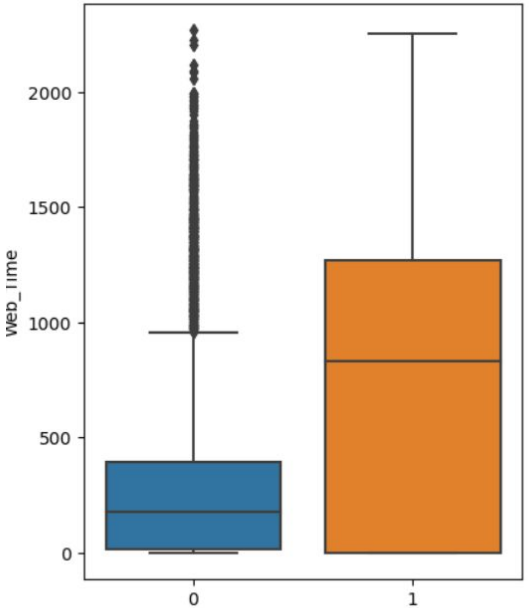


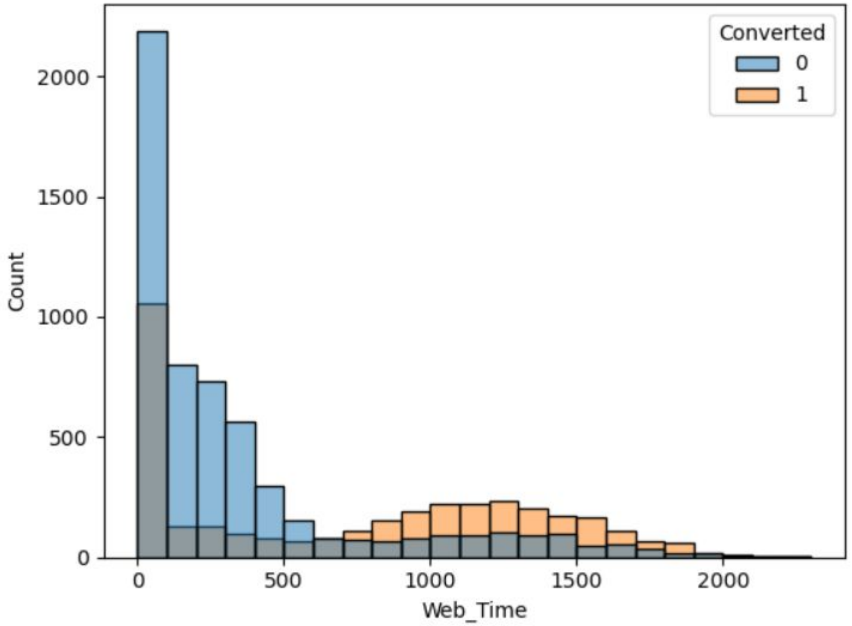


Observation:-

- “Total Visits” and “Page_per_visit” have some outliers with very high value. To reduce effect of outliers on model training ; outlier values above 99% percentile values replaced with 99% percentile values.
- There seems high correlatin between “Total Visits” and “Page_per_visit” . “Page_per_visit” column is dropped to avoid multicolinearity and optimum performace of model

Web_Time





Observation:-

- Web_Time for positive leads are vey high compared to negative leads. **For Web_Time >500** chances of lead B

Summary_Exploratory Data Analysis

Sr.No.	Shortlisted Columns	Insights
1	Lead Origin	**"Lead Add Form"** probability of lead **turning positive is high**
2	Lead Source	**"Reference"** probability of lead **turning positive is very high**
3	Converted	**Target Variable**
4	TotalVisits	**Higher "Total Visit**" probability of lead **turning positive is high.**
5	Web_Time	Higher **"Web_Time"** probability of lead **turning positive is high**
6	Last Activity	**"SMS Sent"** , chance of lead **tuning to positive is high**
7	Specialization	**"Missing Value"** indicates chance of lead **turning to negatie is high**
8	Occupation	**"Working Professional"** probability of lead **turning to positive is very high**
9	Tags	Category **"Missing Values" or "Ringing" probability** of lead **turning negative is very high**. While for Categories **"Closed by Horizzon" or "Will revert after reading the email"** - probability of lead **turning positive is high**
10	Lead Quality	For categories **"Lead Quality" category "Not Sure" or "Worst" ,probabitlity** of lead **turning negative are high.** For categories **"Lead Quality" values are "High in Relevance" ,"Low in Relevance", "Might be"** - probability of lead **turning positive is high**
11	DN_Email	Further to be analyzed during model building
12	XEDU_Ref	Further to be analyzed during model building
13	Reason_course_select	Further to be analyzed during model building
14	Free_Copy	Further to be analyzed during model building

Data Preprocessing

- Data Preprocessing is done to convert textual features to numerical features and scaling features to similar scale for feeding to machine learning model
- Techniques Employed :
  - Categorical Columns with Binary Values (Yes/No) converted to (1/0)
  - Categerorical Columns with multiple categories :- One Hot Encoding / Dummifying
  - Min-Max Scaling :- Data in numerical columns are highly skewed so min-max scaling is applied to retain distribution probperties.

Sr.No.	Shortlisted Columns	Variable Type	Data Preprocessing
1	Lead Origin	Caegorical	One Hot Encoding / Dummifying
2	Lead Source	Categorical	One Hot Encoding / Dummifying
3	Converted	Boolean	Convert to 1/0
4	TotalVisits	Numerical	Min / Max Scaling
5	Web_Time	Numerical	Min / Max Scaling
6	Last Activity	Categorical	One Hot Encoding / Dummifying
7	Specialization	Categorical	One Hot Encoding / Dummifying
8	Occupation	Categorical	One Hot Encoding / Dummifying
9	Tags	Categorical	One Hot Encoding / Dummifying
10	Lead Quality	Categorical	One Hot Encoding / Dummifying
11	DN_Email	Boolean	Convert to 1/0
12	XEDU_Ref	Categorical	One Hot Encoding / Dummifying
13	Reason_course_select	Categorical	One Hot Encoding / Dummifying
14	Free_Copy	Boolean	Convert to 1/0

Automatic Feature Selection

- After Preprocessing of data ; dataset is having 45 features which is very high for feeding to model and shortlisting manually one by one.
- Automatic Recursive Feastue Elemination Techinique is applied to shorlist Top-15 features
- These 15 features will be utilized for further fine tuning model manually

Sr.No.	Shortlisted Columns By RFE
1	Occupation_Working Professional
2	Tags_Ringing
3	Tags_Select
4	Tags_Will revert after reading the email
5	Occupation_Housewife
6	Last Activity_SMS Sent
7	Reason_course_select_Select
8	cat_Web_Time_(248.0, 936.0]
9	Last Activity_Olark Chat Conversation
10	DN_Email_1
11	Lead Quality_Might be
12	Lead Quality_Not Sure
13	Lead Quality_Select
14	Lead Quality_Worst
15	Lead Origin_Others
16	Lead Origin_Lead Add Form
17	cat_Web_Time_(12.0, 248.0]
18	Web_Time
19	Tags_Others
20	Tags_Interested in other courses

Manual Model Tuning – First Model

- Manual Model Turning is done to reduce number of featues from 20 features selected by RFE for optimal model performance
- Features with high p values (> 0.05) and high Variance Inflation Fatctor –VIF (>5)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8911	0.751	-3.849	0.000	-4.363	-1.419
Occupation_Working Professional	0.8009	0.272	2.943	0.003	0.267	1.334
Tags_Ringing	-0.2040	0.757	-0.270	0.788	-1.687	1.279
Tags_Select	4.2684	0.734	5.815	0.000	2.830	5.707
Tags_Will revert after reading the email	6.4890	0.749	8.665	0.000	5.021	7.957
Occupation_Housewife	20.6137	1.39e+04	0.001	0.999	-2.73e+04	2.73e+04
Last Activity_SMS Sent	1.4821	0.106	14.035	0.000	1.275	1.689
Reason_course_select_Select	-2.0580	0.137	-15.010	0.000	-2.327	-1.789
cat_Web_Time_(248.0, 936.0]	-0.6954	0.115	-6.053	0.000	-0.921	-0.470
Last Activity_Olark Chat Conversation	-1.1269	0.202	-5.586	0.000	-1.522	-0.732
DN_Email_1	-1.4376	0.212	-6.774	0.000	-1.854	-1.022
Lead Quality_Might be	-1.5991	0.230	-6.955	0.000	-2.050	-1.148
Lead Quality_Not Sure	-1.8850	0.228	-8.276	0.000	-2.331	-1.439
Lead Quality_Select	-1.4736	0.214	-6.883	0.000	-1.893	-1.054
Lead Quality_Worst	-3.4768	0.408	-8.520	0.000	-4.277	-2.677
Lead Origin_Others	-1.1375	0.626	-1.818	0.069	-2.364	0.089
Lead Origin_Lead Add Form	1.9682	0.236	8.336	0.000	1.505	2.431
cat_Web_Time_(12.0, 248.0]	-1.2491	0.147	-8.506	0.000	-1.537	-0.961
Web_Time	2.6471	0.218	12.161	0.000	2.221	3.074
Tags_Others	3.2827	0.727	4.514	0.000	1.857	4.708
Tags_Interested in other courses	0.4543	0.798	0.569	0.569	-1.110	2.019

	Features	VIF
0	Occupation_Working Professional	1.35
1	Tags_Ringing	2.76
2	Tags_Select	10.26
3	Tags_Will revert after reading the email	3.28
4	Occupation_Housewife	1.01
5	Last Activity_SMS Sent	1.84
6	Reason_course_select_Select	4.45
7	cat_Web_Time_(248.0, 936.0]	1.67
8	Last Activity_Olark Chat Conversation	1.36
9	DN_Email_1	1.12
10	Lead Quality_Might be	2.37
11	Lead Quality_Not Sure	2.21
12	Lead Quality_Select	8.15
13	Lead Quality_Worst	1.33
14	Lead Origin_Others	1.02
15	Lead Origin_Lead Add Form	1.40
16	cat_Web_Time_(12.0, 248.0]	2.09
17	Web_Time	2.51
18	Tags_Others	2.67
19	Tags_Interested in other courses	1.61

Manual Model Tuning – Final Model

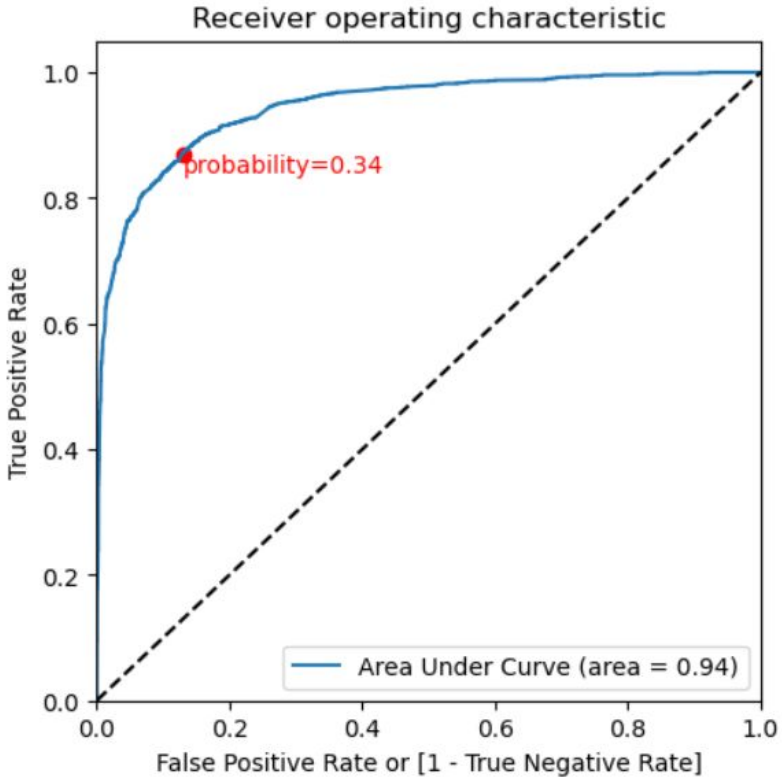
- After removing features based on p-value and VIF in mutliple steps , final model is derived which has all p values within range (< 0.05) and VIF (<5)
- Finally model has 14 features

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6703	0.117	-14.290	0.000	-1.899	-1.441
Occupation_Working Professional	1.0937	0.260	4.199	0.000	0.583	1.604
Tags_Will revert after reading the email	5.1621	0.207	24.942	0.000	4.756	5.568
Last Activity_SMS Sent	1.5451	0.095	16.195	0.000	1.358	1.732
Reason_course_select_Select	-0.5113	0.107	-4.772	0.000	-0.721	-0.301
cat_Web_Time_(248.0, 936.0]	-0.7212	0.105	-6.893	0.000	-0.926	-0.516
Last Activity_Olark Chat Conversation	-0.9238	0.193	-4.798	0.000	-1.301	-0.546
DN_Email_1	-1.1797	0.203	-5.821	0.000	-1.577	-0.783
Lead Quality_Might be	-1.4893	0.168	-8.849	0.000	-1.819	-1.159
Lead Quality_Not Sure	-1.6300	0.148	-11.026	0.000	-1.920	-1.340
Lead Quality_Worst	-3.0150	0.364	-8.272	0.000	-3.729	-2.301
Lead Origin_Lead Add Form	2.2291	0.206	10.796	0.000	1.824	2.634
cat_Web_Time_(12.0, 248.0]	-1.4257	0.138	-10.328	0.000	-1.696	-1.155
Web_Time	2.2899	0.194	11.781	0.000	1.909	2.671
Tags_Others	1.4268	0.110	12.957	0.000	1.211	1.643

	Features	VIF
0	Occupation_Working Professional	1.35
1	Tags_Will revert after reading the email	2.36
2	Last Activity_SMS Sent	1.69
3	Reason_course_select_Select	1.71
4	cat_Web_Time_(248.0, 936.0]	1.35
5	Last Activity_Olark Chat Conversation	1.20
6	DN_Email_1	1.12
7	Lead Quality_Might be	1.97
8	Lead Quality_Not Sure	1.39
9	Lead Quality_Worst	1.19
10	Lead Origin_Lead Add Form	1.24
11	cat_Web_Time_(12.0, 248.0]	1.46
12	Web_Time	1.89
13	Tags_Others	1.48

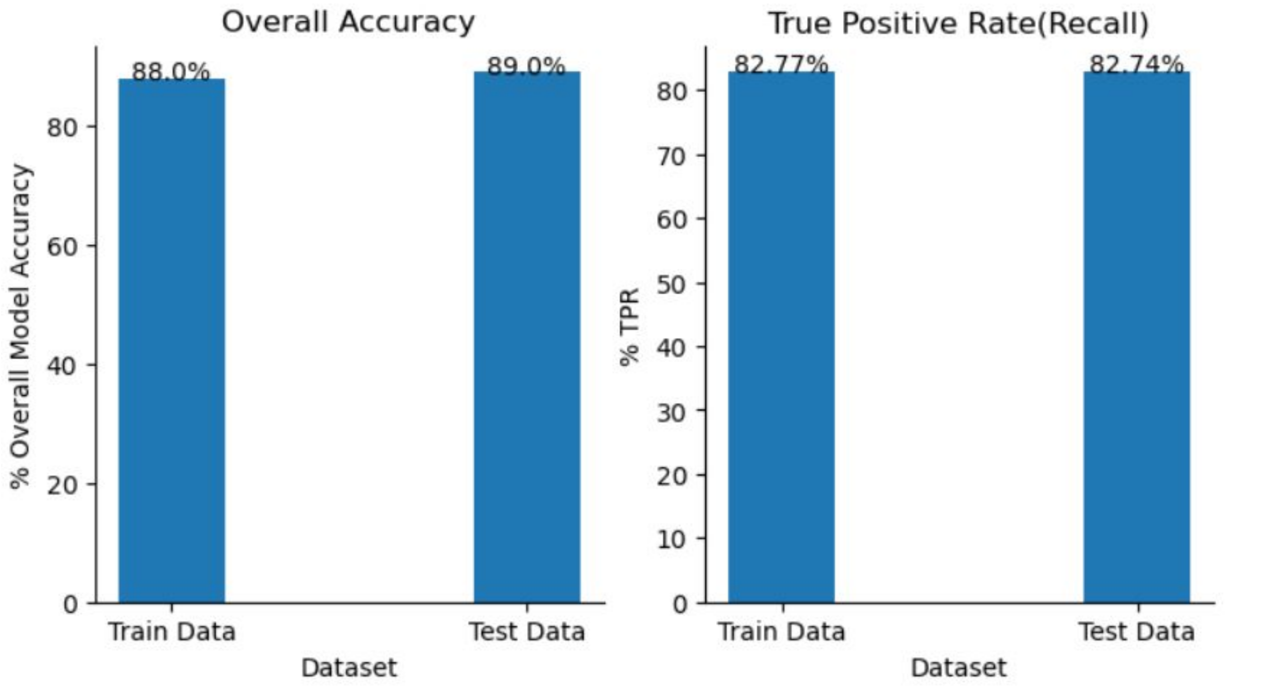
Find Probability Threshold Values

- Logistic Regression Model provides output interms of probaility for positive conversion of target variable
- We need to find out optimum probability value above which datapoint can be classifed positive.
- This is done by plotting Receiver Operating Characteristic Curve (ROC Curve) for different threshold values and selecting optimuum threshold where True Positive Rate **(TPR) is maximum** and **False Positve Rate (FPR)** is minimum
- **Optimum Probability Threshold** from ROC found to be **0.34**
- **Area Under Curve (AUC)** which important factor is **0.96** which denotes that **model is highly reliable**



Evaluation of Model

- Evaluation of model is done on both train and test data set to ensure consistent performance of model on both train and test data
- Company has given target of at least 80% lead turning Positive i.e. Out of Total Leads identified as Hot Leads (Potential Positive) , 80% should turn actual positive.
- In addition to overall accurcacy of the model, TPR(Recall / Sensitivity) = True Positive / Total Actual Positive = TP/(TP+FM) >=80% is important criteria for analysis
- **Model Accuracy is high(>88%)** and **True positive rate is >82%** also both matrics are consistent on both data set. Also AUC is 94%. **The performance of model is acceptable and can be considered as final model**



Model Interpretation

Rank	Feature Name	Weights
1	Tags_Will revert after reading the email	5.162069
2	Lead Quality_Worst	-3.01495
3	Web_Time	2.289906
4	Lead Origin_Lead Add Form	2.229077
5	Lead Quality_Not Sure	-1.63
6	Last Activity_SMS Sent	1.54511
7	Lead Quality_Might be	-1.4893
8	Tags_Others	1.426799
9	cat_Web_Time_(12.0, 248.0]	-1.42573
10	DN_Email_1	-1.17971
11	Occupation_Working Professional	1.093747
12	Last Activity_Olark Chat Conversation	-0.9238
13	cat_Web_Time_(248.0, 936.0]	-0.72122
14	Reason_course_select_Select	-0.51127

Interpretation :-

- Tag :- Tag is most significant parameter for probability of lead turning positive or negative. If Tag is "Will Revert After Reading Email" or "Others" - Chances of Lead Turning to Positive is very high
- Lead Quality :- Lead Quality is significant parameter for probaility of lead turning to positive or negative. If Lead Quality has Not Sure, Might be or Worst, the probability of lead turning negative is very high as they have negative correlation coefficients
- Web Time :- if Web time <936 then probability of lead turning negative increases significantly. In General as Web_Time increases probaility of lead turning positive increases
- Lead Origin :- If Lead Oringin is from “Lead Add Form” probailiy of lead turning positive is hight
- Lead Activity : - If Lead Activity is “SMS Sent” probabiliy of lead turning positive is high
- Occuption :- If Occupation is “working professional” probability of lead turning positive is high
- Last Activity :- If last activity is “Olark Chat Conversation” chances of lead conversion to negative is very high
- Reason_course_select :- If Value is missing ; chances of lead turning negative is very high.b

Summary :-

- Tags, Lead Quality and Web_time are top 3 parameters for evaluating probaility of lead
- Cut off for Lead Score is 34. If Lead score is above 34 then lead can be considered as “Hot Lead”
- Overall accuracy of model is 88%
- Chances for hot lead turning to positive lead is 82%

THANK YOU