Lead Score Assignment

Steps Followed for Analysis:

- Identify Problem Statement.
- Understanding Data.
- Cleaning Data.
- Visualization of Data and Identifying Data Pattern.
- Insights Derived from Univariate/Bivariate Analysis.
- Numeric Variable Co-Relation Analysis.
- Preparation of Data.
- Heatmap plot to derive co-relation of variables.
- Elimination of multi-collinear Features.
- Building Best Fit Model.
- Model Evaluation Metrics.
- Predictions from Test Dataset.
- Recommendation for Company.

Identify Problem Statement.

1. To **Identify most important factors** contributing to higher conversion rate.
2. Identify **focus areas** to increase probability of conversion.

Understanding Data.

Below are findings from initial analysis of Data:-

1. 1. Checked Data and have approximate 9000+ Lead records with total 37 features.
2. 2. Checked Data Types of Variables and there are total 30 categoric data types and 7 numeric data types.
3. 3. Checked Null/Missing values and there were few variables that have Null values.
4. 4. Few variables have label "**Select**," which stands for “**Null values**" in the context of the default choice that shows when a customer do not provide any details.

Cleaning Data.

Below steps taken to clean Dataset :-

1. Replaced ‘Select’ labels to null values as they are default values when customer does not provide any details and hence equivalent to Null values.
2. Features with more than 40% null values is removed as they provide very few valuable information for model building.
3. Features having only one label values like ‘Magazine’ , ‘Search’, ‘Newspaper’ have only label values ‘No’ throughout the data and they were removed as they do not add any value and will impact in model building process.
4. There were many label values very few in numbers and merged them to 'Other' label value.
5. Features with less than 40% null values were replaced with ‘Missing’ label and features having less than 5% were imputed with mode values for categoric variables and with mean value for numeric variable.

Visualization of Data and Identifying Data Pattern.

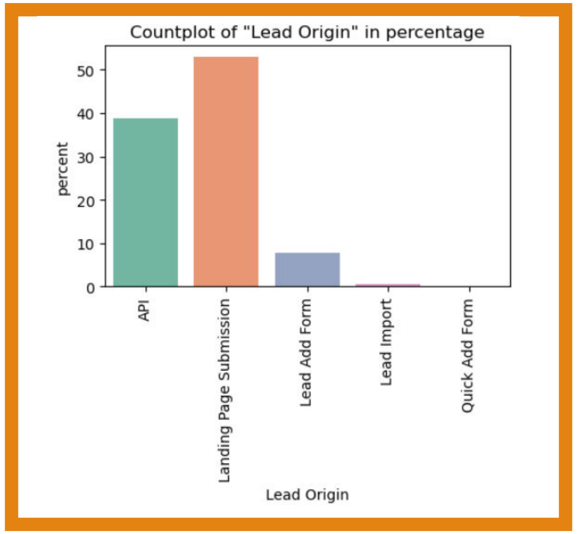
Below seps taken for Analyzing data:

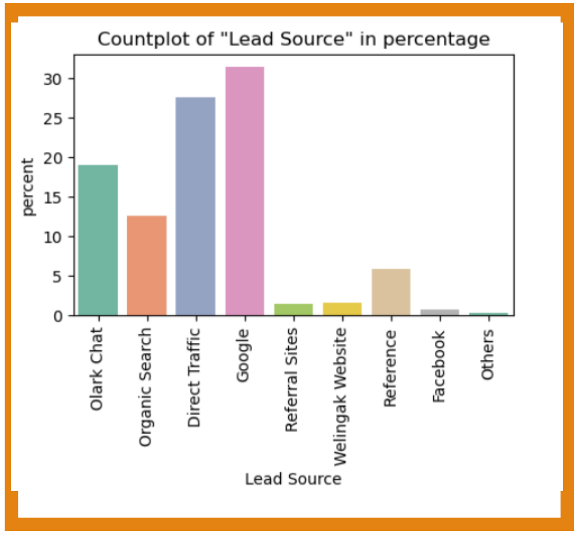
1. Univariate Analysis of Categoric Variables.
2. Univariate Analysis Numeric Variables.
3. Bivariate and Multivariate Analysis of Categoric and Numeric Variable.
4. Count plot and Box plot of variables to derive insights.

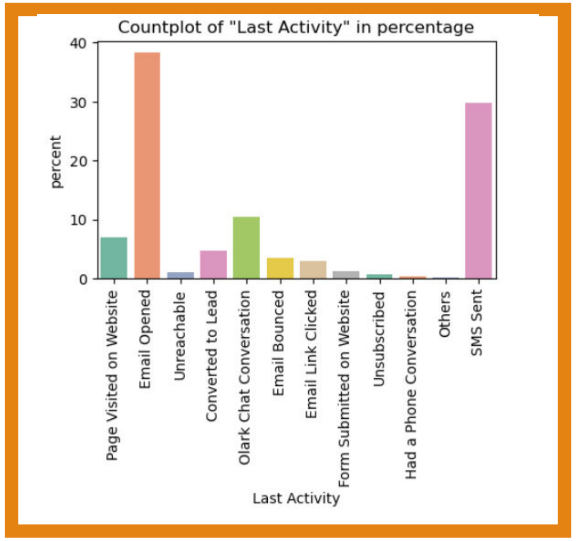
Visualisation of Data and Identifying Data Pattern

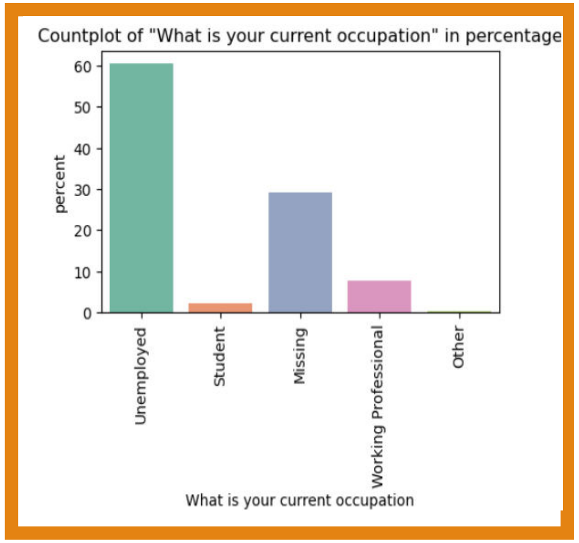
Below are observation from Data Visuals:-

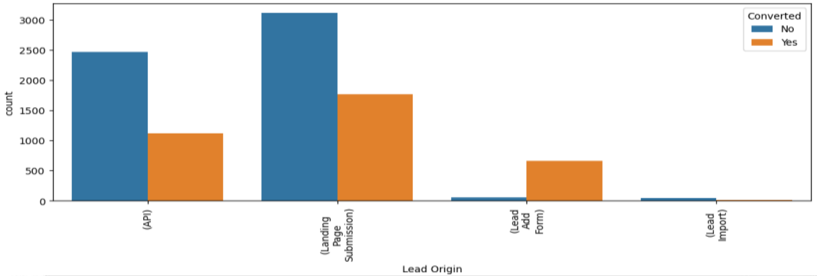
1. 52% of leads have their origin identified at 'Landing Page Submission'.
2. Highest Leads got the source of company from 'Google' (32%)
3. 39% of Leads last Activity was 'Email Openend' which is highest. 30% of Leads sent SMS as their Last activity.
4. Around 9% of Leads are currently Working. Maximum Leads (60%) are unemployed

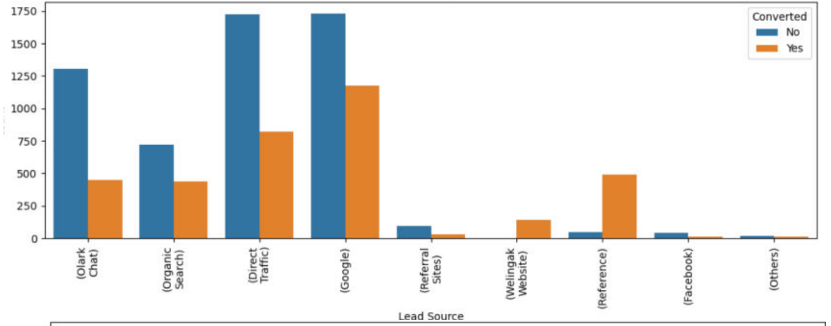


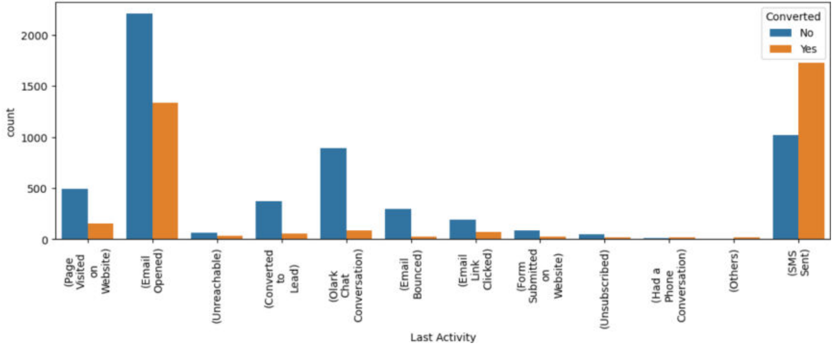


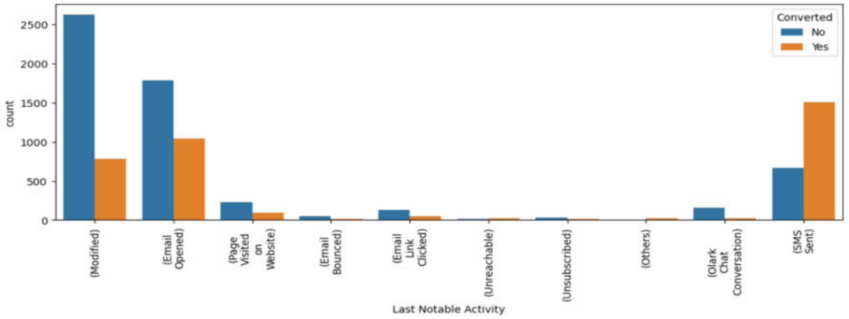












Visualisation of Data and Identifying Data Pattern

Below are observation from Data Visuals:-

1. lead conversion rate are highest for 'Lead Add Form' label of Lead Origin.
2. Successful conversion rate are highest from 'Reference' source and second highest successful conversion rate are from 'Welingak Website'.
3. leads with last activity 'Sms sent' have Successful conversion rate.
4. there is higher chance of successful conversion rate in 'Marketing Management' and 'Operations Management' specialization.
5. Higher conversion rate are observed in 'Working Professional'
6. Conversion Rate is highest in Leads with last notable activity as 'SMS Sent'.

Insights Derived from Univariate/Bivariate Analysis.

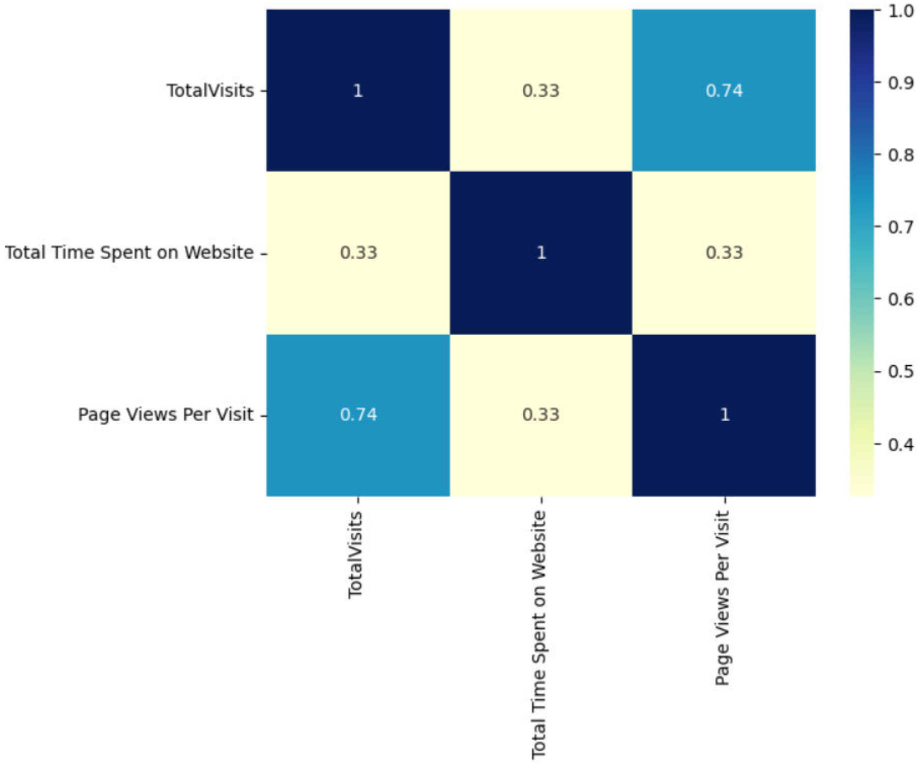
1. To maximize lead conversion rate of 'Olark chat', 'Organic Search', 'Direct Traffic' and 'google' source. We need to focus more on generating more leads from 'Reference' and 'Wellingak Website' as these sources show promising successful conversion.

2. We should focus more on increasing leads in 'Business Administartion', 'Finance Management', 'HR Management', 'Marketing Management' and 'Operations Management'. To look for ways to increase lead conversation rate 'Banking, Investment and Insurance', 'Healthcare Management', 'Marketing Management' and 'Operations Management’.

3. To look for ways to increase lead conversation rate for Students and Unemployed.

4. Converted Leads tend to spend more time on Website and also they view more pages and visit more websites as compared to non-converted customers.

5. We should focus more on Leads who spends greater amount of time on Website, visits more websites and visits more pages as they tend to become converted customers.



Numeric Variable Co-Relation Analysis.

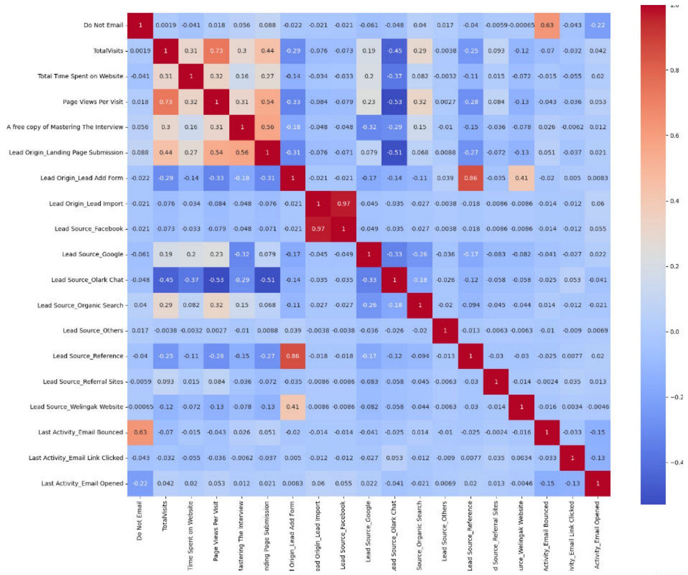
Below are observation from Data Visuals:-

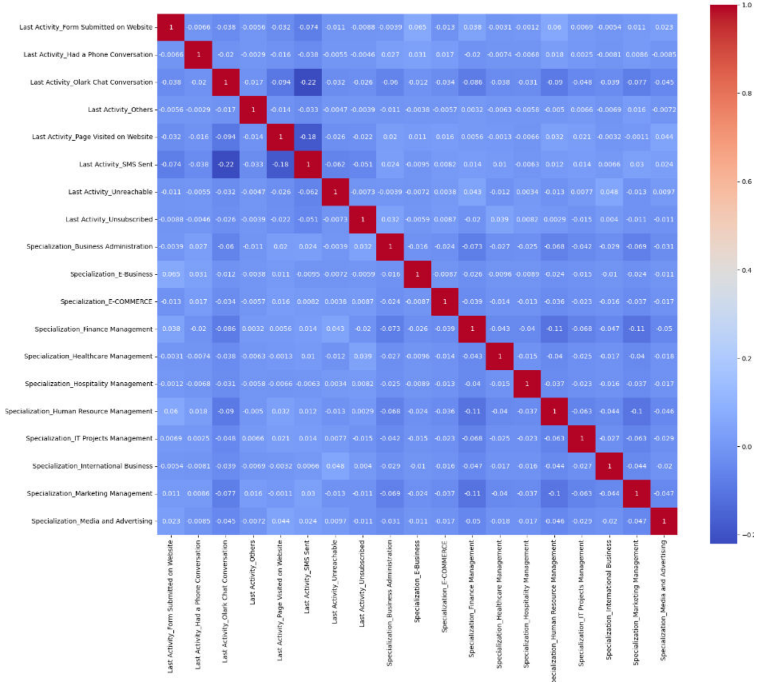
1. Total Visits and Page Views Per Visit are strongly positively co-related.
2. Total Time Spent on Websites and Page Views Per Visit are weakly positively co-related.

Preparation of Data.

Below steps taken for Data preparation:-

1. Mapped categoric variables having labels ‘Yes’ and ‘No’ to binary values.
2. Created dummy variables of object type variables.
3. Converted all variables to float type for easy interpretation of data.
4. Splitting of dataset into Train and Test dataset of model building and evaluation
5. Standardized numerical features using StandardScaler method.





Heatmap plot to derive co-relation of variables.

Below are observation from Heatmap graphs:

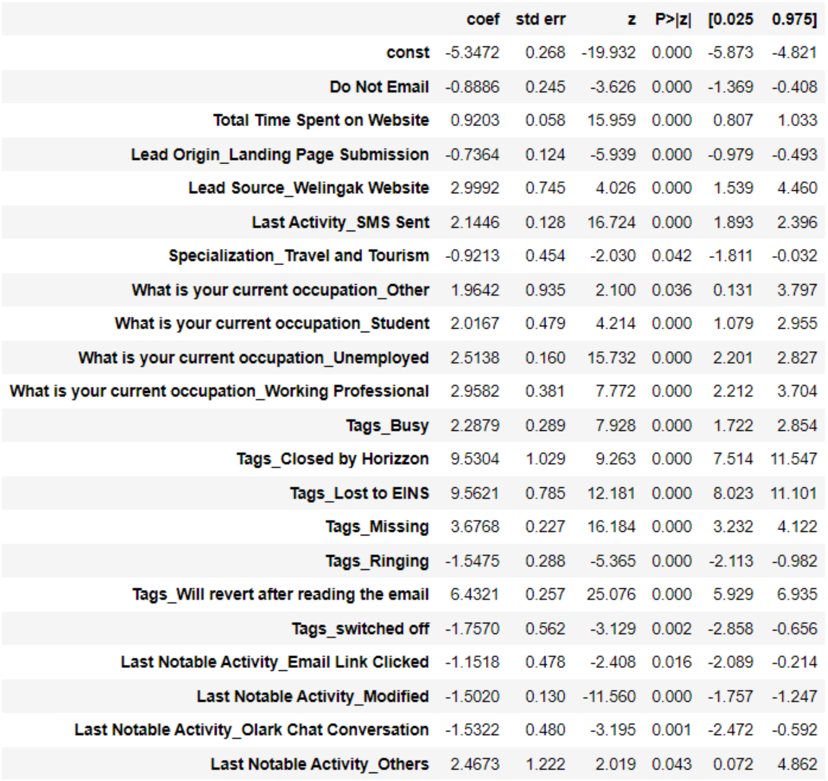
1. Lead_origin_Lead_import is highly positively corelated with Lead_Source_Facebbok.
2. Lead_Source_reference is highly positively corelated with Lead_Origin_Lead Add Form.
3. Total_Visits is having high positive correlation with Page Views Per Visit.
4. Lead_source_Welingak_Website is having weak positive correlation with Lead_Origin_Lead Add Form.
5. Lead_source_Olark_chart is strongly negatively corelated to Page Views Per Visit.
6. ead_source_Olark_chart is strongly negatively corelated to Lead_Origin_Landing Page Submission.
7. Last_Notable_Activity_Email_Opened is having strong negative correlation with Last_Notable_Activity_Modified.

Elimination of multi-collinear Features.

1. By using automated method of RFE, retained top 30 features from total 77 dummy features.
2. Used manual method of calculating VIF of each variables and eliminated variable having VIF greater than 5 and it’s significance value greater than 0.05 during every model building step.

Building Best Fit Model.

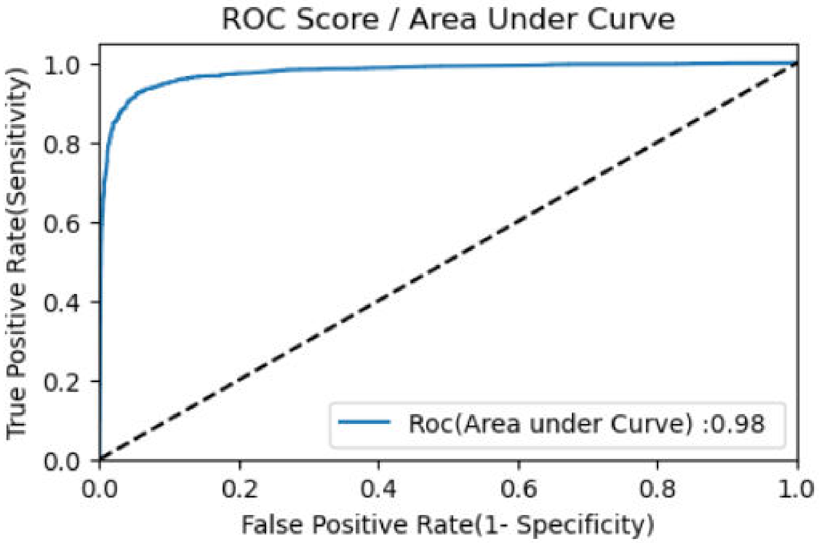
Iterated model building process ten time and obtained last best fit model that have features VIF value less than 5 and significance value less than 0.05.



	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.3472	0.268	-19.932	0.000	-5.873	-4.821
Do Not Email	-0.8886	0.245	-3.626	0.000	-1.369	-0.408
Total Time Spent on Website	0.9203	0.058	15.959	0.000	0.807	1.033
Lead Origin_Landing Page Submission	-0.7364	0.124	-5.939	0.000	-0.979	-0.493
Lead Source_Welingak Website	2.9992	0.745	4.026	0.000	1.539	4.460
Last Activity_SMS Sent	2.1446	0.128	16.724	0.000	1.893	2.396
Specialization_Travel and Tourism	-0.9213	0.454	-2.030	0.042	-1.811	-0.032
What is your current occupation_Other	1.9642	0.935	2.100	0.036	0.131	3.797
What is your current occupation_Student	2.0167	0.479	4.214	0.000	1.079	2.955
What is your current occupation_Unemployed	2.5138	0.160	15.732	0.000	2.201	2.827
What is your current occupation_Working Professional	2.9582	0.381	7.772	0.000	2.212	3.704
Tags_Busy	2.2879	0.289	7.928	0.000	1.722	2.854
Tags_Closed by Horizzon	9.5304	1.029	9.263	0.000	7.514	11.547
Tags_Lost to EINS	9.5621	0.785	12.181	0.000	8.023	11.101
Tags_Missing	3.6768	0.227	16.184	0.000	3.232	4.122
Tags_Ringing	-1.5475	0.288	-5.365	0.000	-2.113	-0.982
Tags_Will revert after reading the email	6.4321	0.257	25.076	0.000	5.929	6.935
Tags_switched off	-1.7570	0.562	-3.129	0.002	-2.858	-0.656
Last Notable Activity_Email Link Clicked	-1.1518	0.478	-2.408	0.016	-2.089	-0.214
Last Notable Activity_Modified	-1.5020	0.130	-11.560	0.000	-1.757	-1.247
Last Notable Activity_Olark Chat Conversation	-1.5322	0.480	-3.195	0.001	-2.472	-0.592
Last Notable Activity_Others	2.4673	1.222	2.019	0.043	0.072	4.862

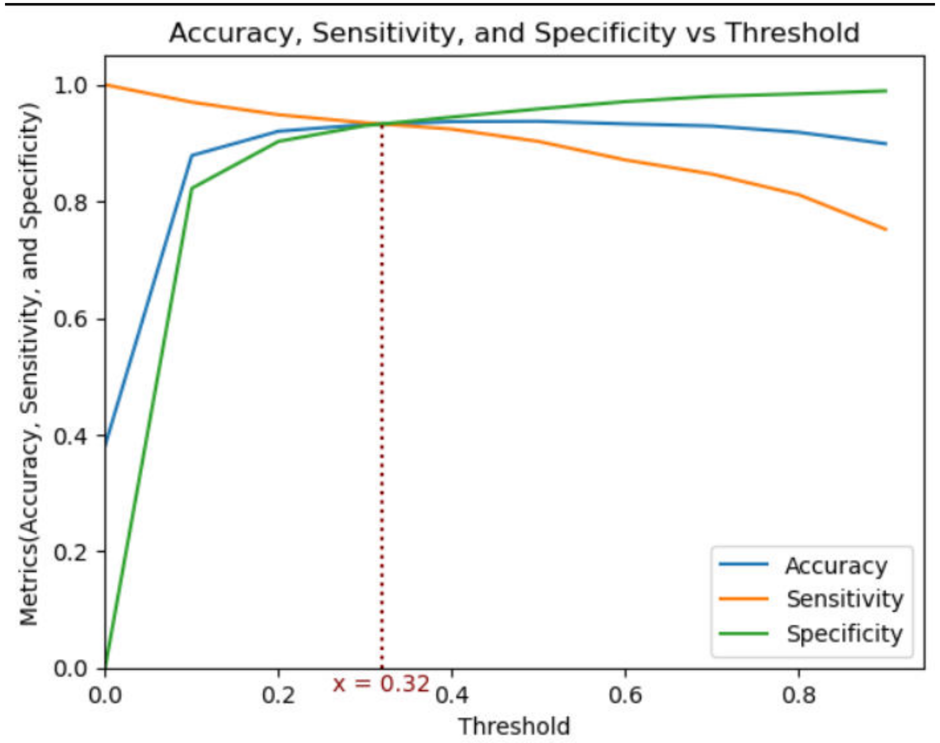
Model Evaluation Metrics.

1. **Finalised best fit model** with **Accuracy** score of **94%**, **Sensitivity** score of **90%** and **Specificity** score of **96%.**
2. ROC score obtained is 0.98 (98%) of graph area which is very good model and it shows good balance between Sensitivity and specificity score.



Model Evaluation Metrics.

'0.32' is optimal threshold value obtained where Sensitivity and Specificity and Accuracy meets together having metrics % values ranging from 90-94%.



Predictions from Test Dataset.

1. Used optimal threshold values as 0.32 obtained from Sensitivity-Specificity trade offs to predict test data set.
2. Last best fit model predicted Converted values of Test data with an Accuracy score of 94%, Sensitivity score of 95% and Specificity score of 93% which is very good measure.

Recommendation for Company.

Below are recommendations for X Education company to increase probability of Lead Conversion:

1. Leads who learned about company via source ‘Welingak Website’ have more probability to become customers and hence company should **focus on broadcasting feedback and ratings from previous clients on Welingak Website** that will have an impact on increase of conversion rate.

2. Leads who have ‘SMS sent’ as their last activity recorded have more chance in getting converted. Company can **increase publishing attractive marketing campaigns via SMS** for these categories of leads so that they have high chance of getting converted to customers.

3. Working Professional Leads have higher chance of buying online courses. Company can focus on **providing offers or rewards for Working Professionals Leads** so that they can buy and avail those offers which will ultimately increase probability of conversion.

4. **Below are the top three variables that contribute most towards probability of lead getting converted: -**

- Total time spent on Website.
- Page Views Per Visit.
- Lead Source.

Thank You.