Lead Score Case Study

Boosting Conversions of X Education with ML

Submitted By
- Kritika Sinha
- Kundan Dombale
- Komal Patil

Index

Business Objective

X Education offers online courses to professionals seeking to upskill, pursue higher education, or transition into new careers.

Education company gets leads from Search engine like Google and several websites.

Despite being getting large number of Lead, their lead conversion is 30%.

To increase the Lead conversion, Sales to team has to know hot leads that needs to be targeted and focused.

To achieve 80% conversion target set by CEO, It has been decided to use Predictive Model that helps team to identify correct leads with high potential

Problem Statement

Lead conversion Rate :

Currently, Conversion Rate of X education Company is around 30%. This shows that lead Generation is much higher than Conversion rate. Leads are getting generated by various sources like google, websites, videos, forms, past referrals.

Lead Scoring system:

Scoring system is one of the technique which will enable company to score leads based on their probability to get converted into customer. This will help to put sale efforts in directive and effective way. This system is currently not used by the Company which are leading to sales team effort loss.

Lead prioritization:

There are cold leads which less likely be converted into customers. Efforts put on that customer can result it revenue and effort loss. This can also increase probability of hot leads losses because of time sensitivity.

Target Conversion rate & Goal

Our target is to increase conversion rate from 30% to 80%. To achieve this, building a logistic regression model will be helpful. Model can help us to give probability of conversion which is nothing but Lead scoring and predict hot lead which is likely to convert

Assumptions

Feature reduction:

To avoid Model to get too complex by creating a lot of dummy variables, we have binned the options into "Other" if their total contribution in that column is less than 10%.

Approach & Methodology -1

1. Importing libraries
2. Sourcing data from Leads.csv
3. Checking columns, metadata, shape, structure
4. Identifying columns that contain “Select” data and replacing it with Null
5. Removing Highly Skewed data columns.
6. Calculating missing value percentage for every column for both files using isnull.mean() function
7. Replacing missing values with below approach
    a. If missing value percentage is more than 40 % then Drop the columns
    b. Dividing columns into categorical and numerical columns. If number of unique values> 30 then Categorical columns else its Numerical column.
    c. Fill Categorical missing values with mode and Numerical missing values with Median.

Approach & Methodology -2

8. Outlier check using boxplot for every numerical column
9. Handling Outlier of data using capping (upper Whisker) and flooring (lower Whisker)
    a. If outliers are at lower side of box plot then flooring i.e. replacing it with lower whisker value.
    b. If outliers are at upper side of box plot then capping i.e. replacing it with upper whisker value.
10. Performing univariate Analysis
    a. Numerical Variable : Histogram
    b. Categorical Variable : Countplot
11. Performing Bivariate Analysis using Bar Plot , Scatter plot, Boxplot
    a. Numerical Variable : Scatter Plot
    b. Categorical Variable : Box Plot, Bar Plot
12. Performing multivariate Analysis
    a. Numerical Variable : Heatmap

Exploratory Data Analysis (EDA)

Dataset contains 37 columns and 9240 rows.

All the columns are categorical columns except Prospect ID, Lead Number, TotalVisits, Total Time Spent on Website, Page Views Per Visit

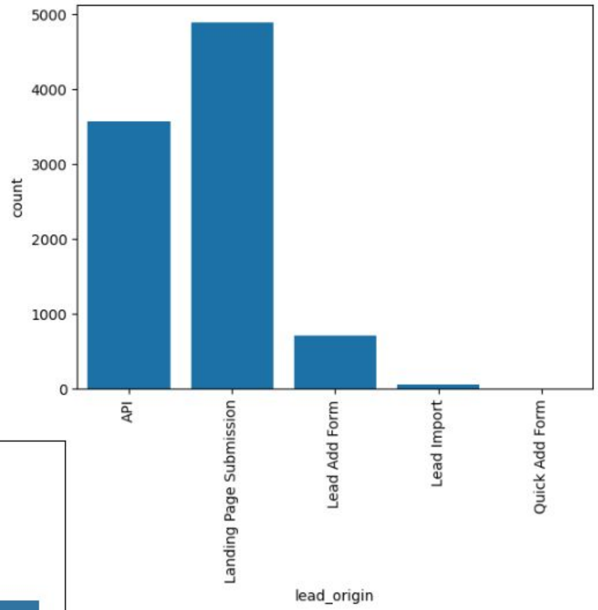
```
[ ] data_df.shape
(9240, 37)
```

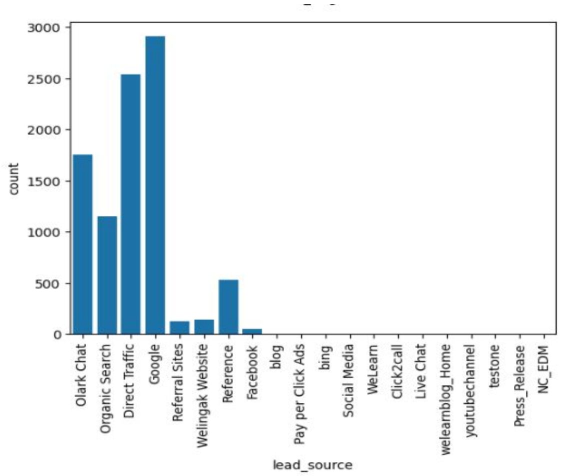
```
Data columns (total 37 columns):
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   Prospect ID                                9240 non-null   object
 1   Lead Number                                9240 non-null   int64
 2   Lead Origin                                9240 non-null   object
 3   Lead Source                                9204 non-null   object
 4   Do Not Email                               9240 non-null   object
 5   Do Not Call                                9240 non-null   object
 6   Converted                                  9240 non-null   int64
 7   TotalVisits                                9103 non-null   float64
 8   Total Time Spent on Website                9240 non-null   int64
 9   Page Views Per Visit                       9103 non-null   float64
 10  Last Activity                              9137 non-null   object
 11  Country                                    6779 non-null   object
 12  Specialization                             7802 non-null   object
 13  How did you hear about X Education         7033 non-null   object
 14  What is your current occupation            6550 non-null   object
 15  What matters most to you in choosing a course  6531 non-null   object
 16  Search                                     9240 non-null   object
 17  Magazine                                   9240 non-null   object
 18  Newspaper Article                          9240 non-null   object
 19  X Education Forums                         9240 non-null   object
 20  Newspaper                                  9240 non-null   object
 21  Digital Advertisement                      9240 non-null   object
 22  Through Recommendations                    9240 non-null   object
 23  Receive More Updates About Our Courses     9240 non-null   object
 24  Tags                                       5887 non-null   object
 25  Lead Quality                               4473 non-null   object
 26  Update me on Supply Chain Content          9240 non-null   object
 27  Get updates on DM Content                  9240 non-null   object
 28  Lead Profile                               6531 non-null   object
 29  City                                       7820 non-null   object
 30  Asymmetrique Activity Index                5022 non-null   object
 31  Asymmetrique Profile Index                 5022 non-null   object
 32  Asymmetrique Activity Score                5022 non-null   float64
 33  Asymmetrique Profile Score                 5022 non-null   float64
 34  I agree to pay the amount through cheque   9240 non-null   object
 35  A free copy of Mastering The Interview     9240 non-null   object
 36  Last Notable Activity                      9240 non-null   object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

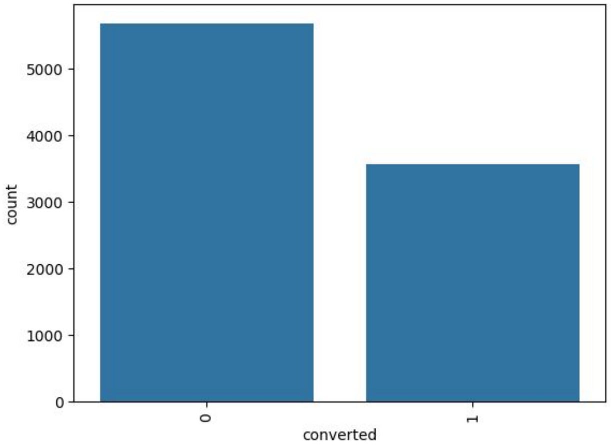
EDA - Univariate Analysis

Insights:

1. Most of the Leads are origin from API and Landing Page Submission.
2. Source of the maximum Leads are Olark Chat, Organix Search. Direct Traffic , google
3. Most of the lead preferred not to call and email
4. less then 40 % leads are converted into customer



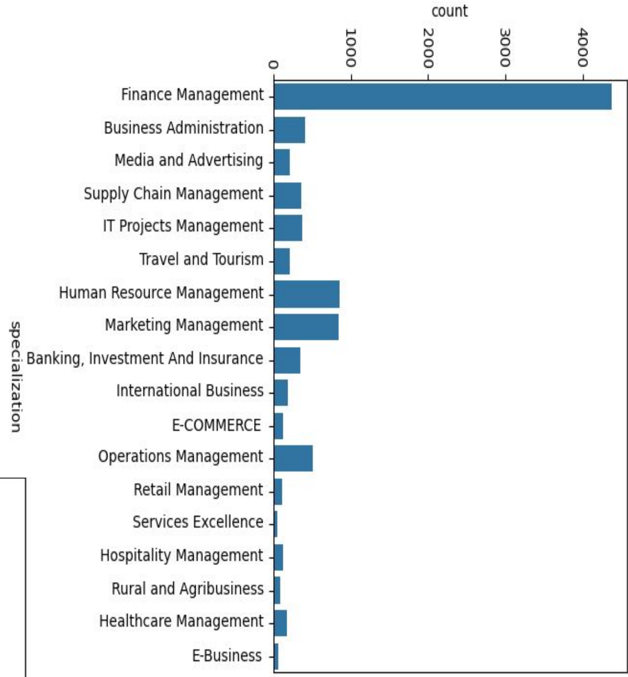


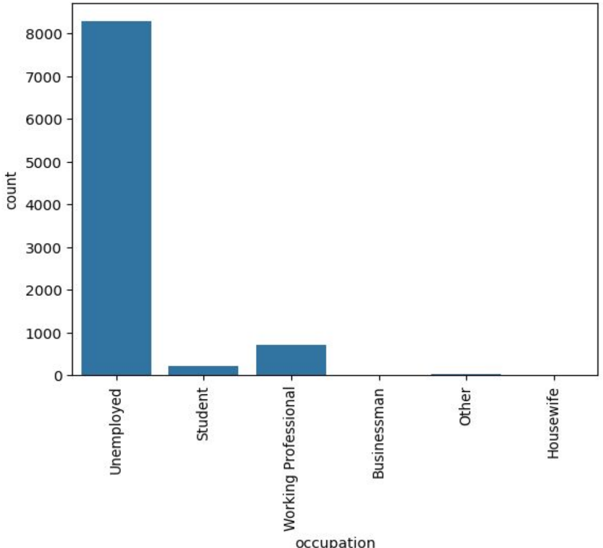


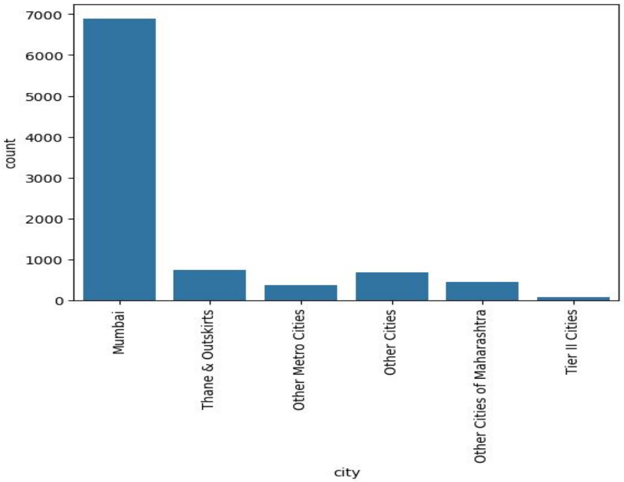
EDA - Univariate Analysis

Insights:

1. Most of the customers worked in finance management before or unemployed
2. Most of the customers are from Mumbai.
3. Most of the customers current status is they will revert after reading the email



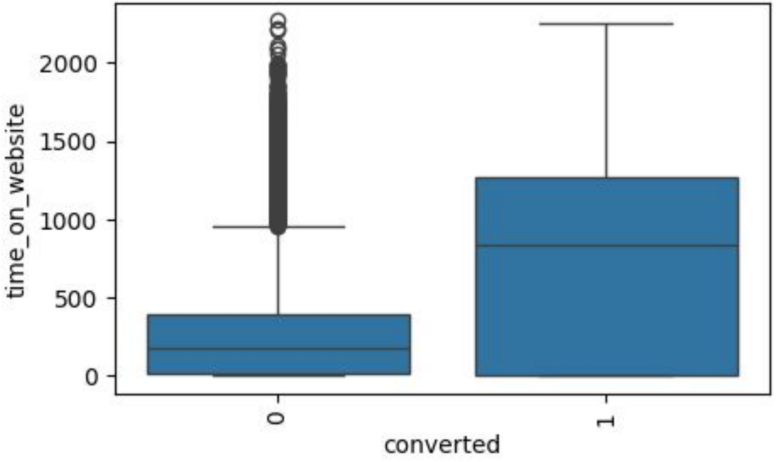


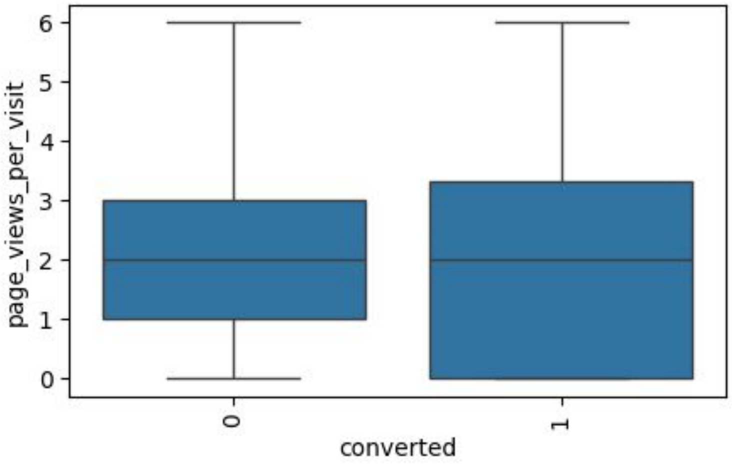


EDA - Bivariate Analysis

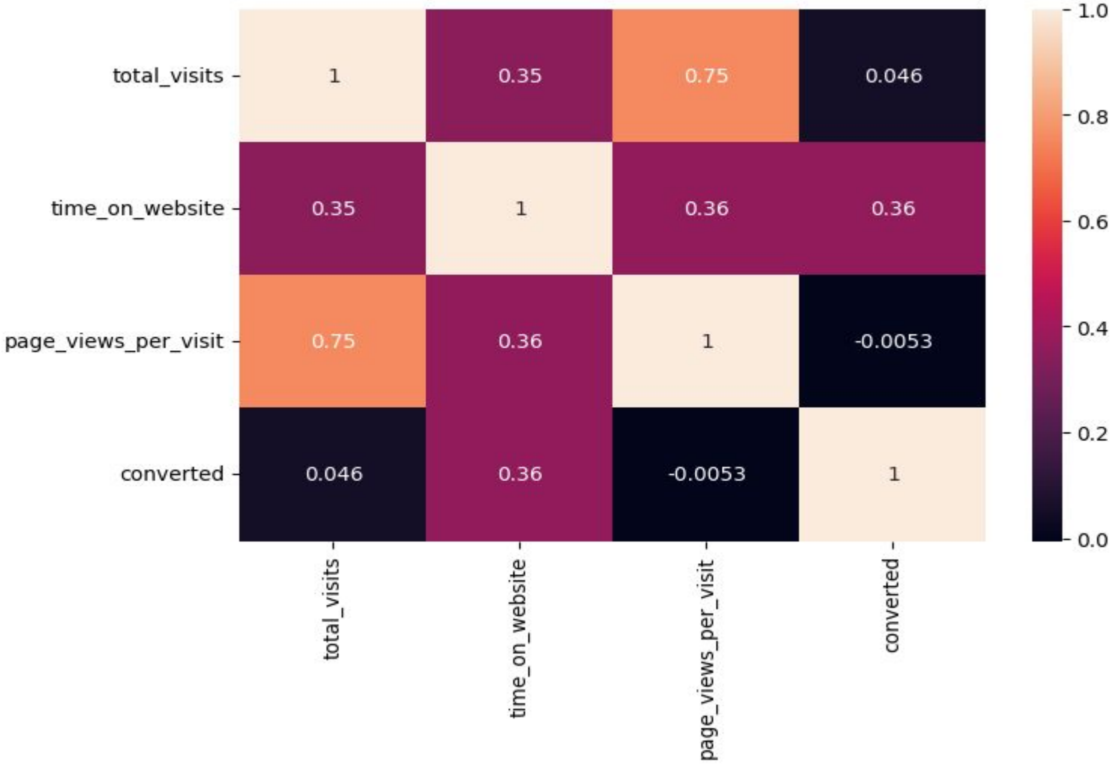
Insights:

1. The total time spent by the customer on the website is high for Converted leads than others.
2. Page view per visit is more in converted Leads than others.





EDA - Multivariate Analysis



Insights:

1. Total visits and page views per visit are correlated
2. Total visits and page views per visit, both of them are not correlated with Converted column.

Machine Learning

1. Splitted Data into train & Test Sets with 70-30 Ratio.
2. Scale the Numerical features with standard scaling
3. Model 1: Created with all the columns (r^2 = 0.5229)
4. Model 2: Using RFE , Selected significant Feature by eliminating Less significant one. 15 Selected Features are in Screenshot.
5. Model 3: Removed tags_Ringing which has high P Value = 0.902
6. Model 4: Removed tags_Interested in other courses which has high P Value = 0.350
7. Model 5: Removed tags_switched off which has high P Value = 0.187
8. Model 6: Removed last_activity_SMS Sent which has high VIF Value = 6.11
9. Model 7: Removed occupation_Unemployed' which has high VIF Value = 5.28

Final Model: Machine learning model is created with less than 0.05 P value and less than 5 VIF value

```
# Features selected
col = X_train.columns[rfe.support_]
col
```

```
Index(['time_on_website', 'lead_origin_Other', 'last_activity_Email Opened',
       'last_activity_Olark Chat Conversation', 'last_activity_SMS Sent',
       'occupation_Unemployed', 'occupation_Working Professional', 'tags_Busy',
       'tags_Closed by Horizzon', 'tags_Interested in other courses',
       'tags_Other', 'tags_Ringing',
       'tags_Will revert after reading the email', 'tags_switched off',
       'last_notable_activity_SMS Sent'],
      dtype='object')
```

```
# Features eliminated
X_train.columns[~rfe.support_]
```

```
Index(['total_visits', 'page_views_per_visit', 'mastering_interview',
       'lead_origin_Landing Page Submission', 'lead_source_Google',
       'lead_source_Olark Chat', 'lead_source_Organic Search',
       'lead_source_Other', 'last_activity_Other',
       'last_activity_Page Visited on Website',
       'specialization_Business Administration',
       'specialization_Finance Management',
       'specialization_Human Resource Management',
       'specialization_IT Projects Management',
       'specialization_Marketing Management',
       'specialization_Media and Advertising',
       'specialization_Operations Management', 'specialization_Other',
       'specialization_Supply Chain Management',
       'specialization_Travel and Tourism', 'city_Others',
       'city_Thane & Outskirts', 'last_notable_activity_Modified',
       'last_notable_activity_Other'],
      dtype='object')
```

Model Evaluation

Generalized Linear Model Regression Results

Dep. Variable:	converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6457
Model Family:	Binomial	**Df Model:**	10
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2011.6
Date:	Tue, 18 Feb 2025	**Deviance:**	4023.3
Time:	08:12:42	**Pearson chi2:**	9.03e+03
No. Iterations:	9	**Pseudo R-squ. (CS):**	0.5090
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.6743	0.207	-27.417	0.000	-6.080	-5.269
time_on_website	0.9884	0.042	23.268	0.000	0.905	1.072
lead_origin_Other	2.8312	0.189	14.999	0.000	2.461	3.201
last_activity_Email Opened	0.7829	0.095	8.230	0.000	0.596	0.969
last_activity_Olark Chat Conversation	-0.7823	0.190	-4.123	0.000	-1.154	-0.410
occupation_Working Professional	2.6305	0.229	11.466	0.000	2.181	3.080
3 **tags_Busy**	4.0232	0.286	14.089	0.000	3.464	4.583
1 **tags_Closed by Horizzon**	9.7764	1.031	9.478	0.000	7.755	11.798
tags_Other	3.4316	0.220	15.610	0.000	3.001	3.862
2 **tags_Will revert after reading the email**	4.4932	0.192	23.360	0.000	4.116	4.870
last_notable_activity_SMS Sent	2.7976	0.122	23.001	0.000	2.559	3.036

Top Feature which highly contributing to the model

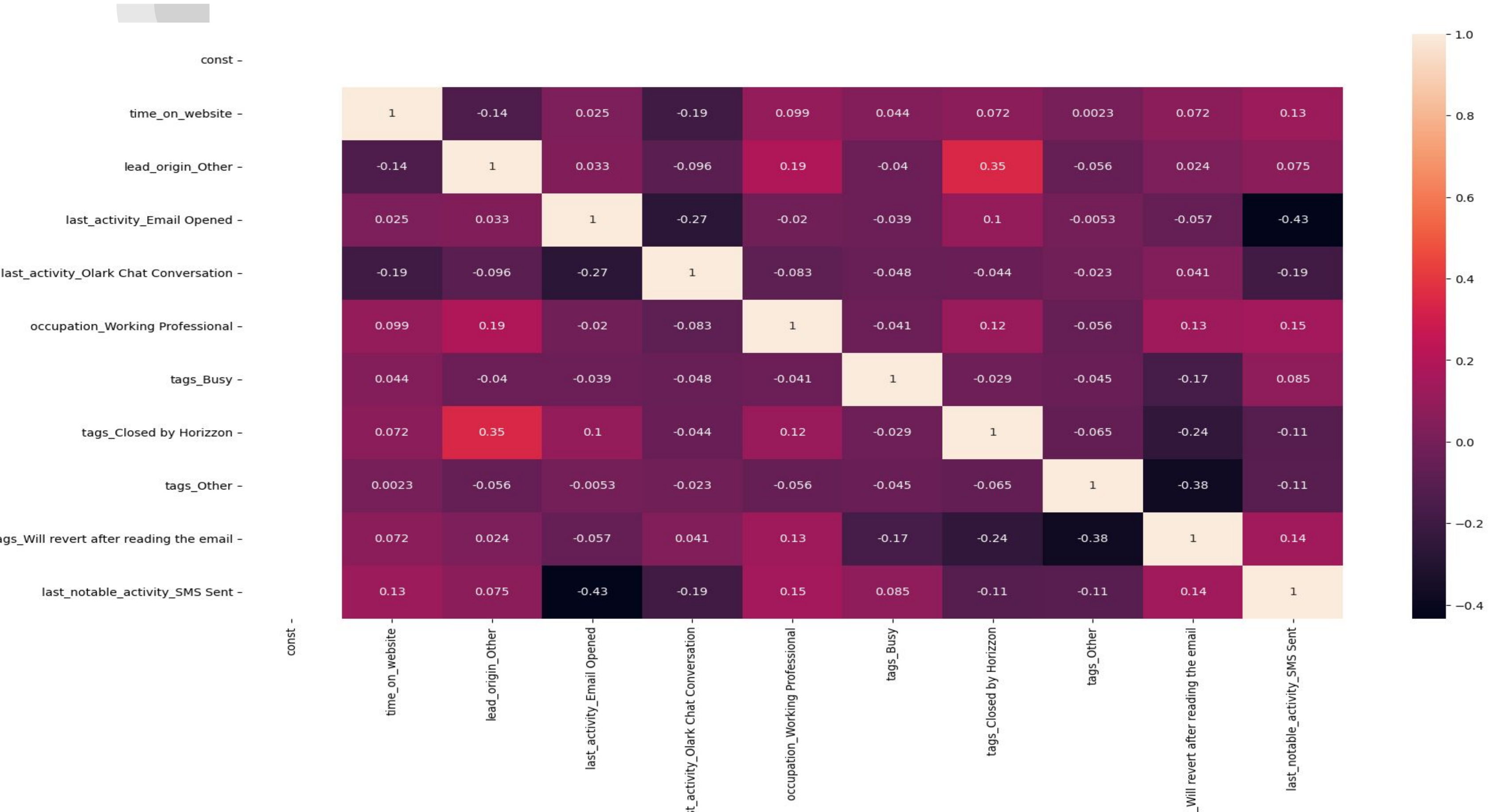
Variance Inflation Factor (VIF) Analysis:

	Features	VIF
0	tags_Will revert after reading the email	2.14
1	last_activity_Email Opened	1.59
2	last_notable_activity_SMS Sent	1.52
3	lead_origin_Other	1.38
4	tags_Closed by Horizzon	1.30
5	last_activity_Olark Chat Conversation	1.23
6	occupation_Working Professional	1.19
7	time_on_website	1.12
8	tags_Other	1.09
9	tags_Busy	1.04

Hence, the final model is as below:

Converted = - 5.6743 + 0.9884 * Total Time Spent on Website + 2.83 * lead_origin_Other + 0.7829 * last_activity_Email Opened - 0.7823 * last_activity_Olark Chat Conversation + 2.6305 * occupation_Working Professional + 4.0232 * tags_Busy + 9.7764 *tags_Closed by Horizzon + 3.4316 * tags_Other + 4.4932 * tags_Will_revert after reading the email + 2.7976 * last_notable_activity_SMS Sent

Selected Feature Correlation



Model Metrics

Training Set Confusion Matrix and other metrics

	Predicted 0	Predicted 1
Actual 0	3636	339
Actual 1	483	2010

```
metrics_df
```

	Metric	Value
0	Accuracy	0.872913
1	Precision	0.855683
2	Recall	0.806258
3	F1 Score	0.830235

```
Sensitivity: 0.8063
Specificity: 0.9147
False Positive Rate: 0.0853
Positive Predictive Value: 0.8557
Negative Predictive Value: 0.8827
```

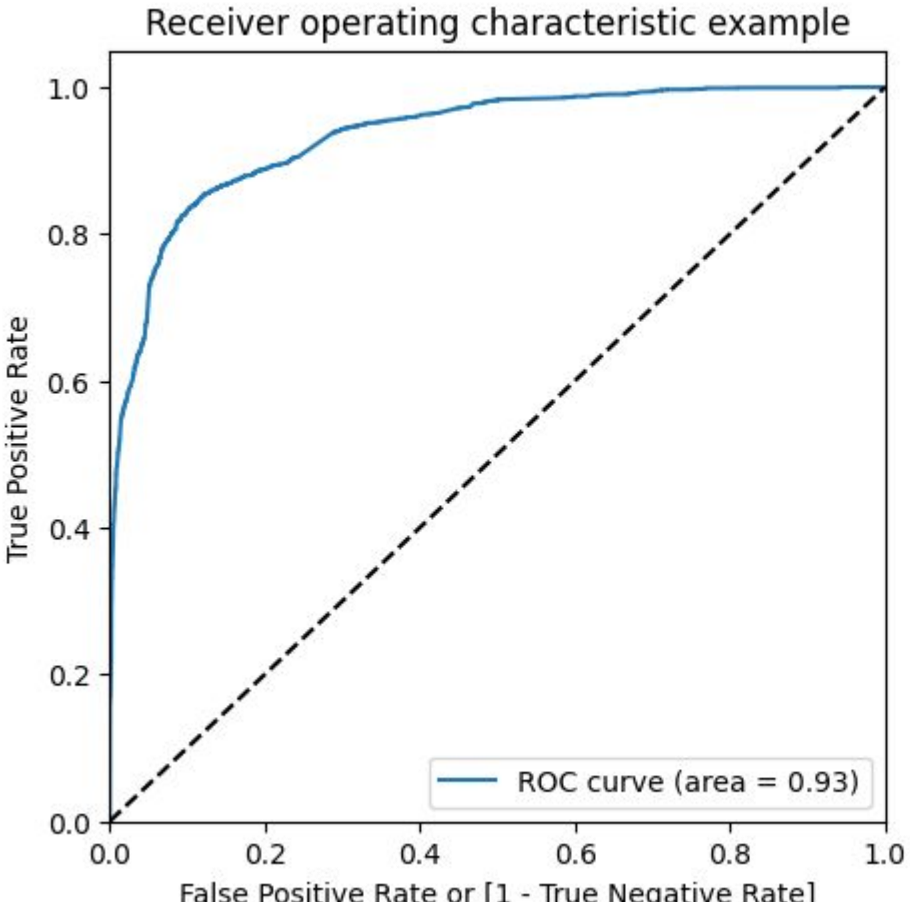
Testing Set Confusion Matrix and other metrics

```
array([[1437,  267],
       [ 146,  922]])
```

	Metric	Value
0	Accuracy	0.851010
1	Precision	0.775442
2	Recall	0.863296
3	F1 Score	0.817014

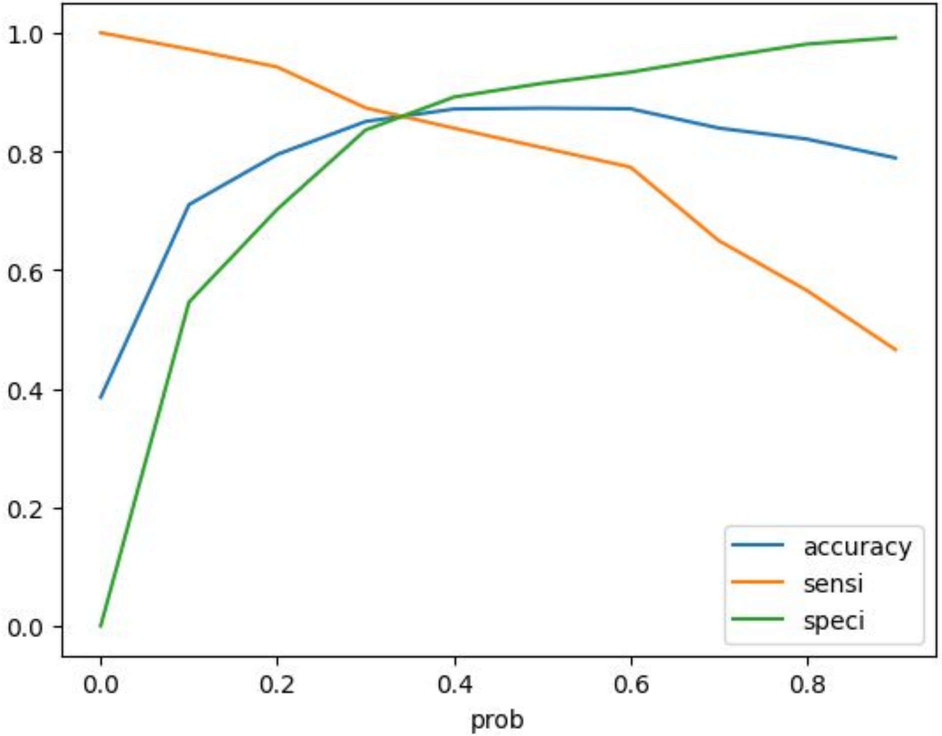
```
Sensitivity: 0.8633
Specificity: 0.8433
False Positive Rate: 0.1567
Positive Predictive Value: 0.7754
Negative Predictive Value: 0.9078
```

Model Evaluation - ROC AUC Train Set



The ROC curve with AUC of 0.93 indicating Model is performing exceptionally well in distinguishing between positive and negative classes.

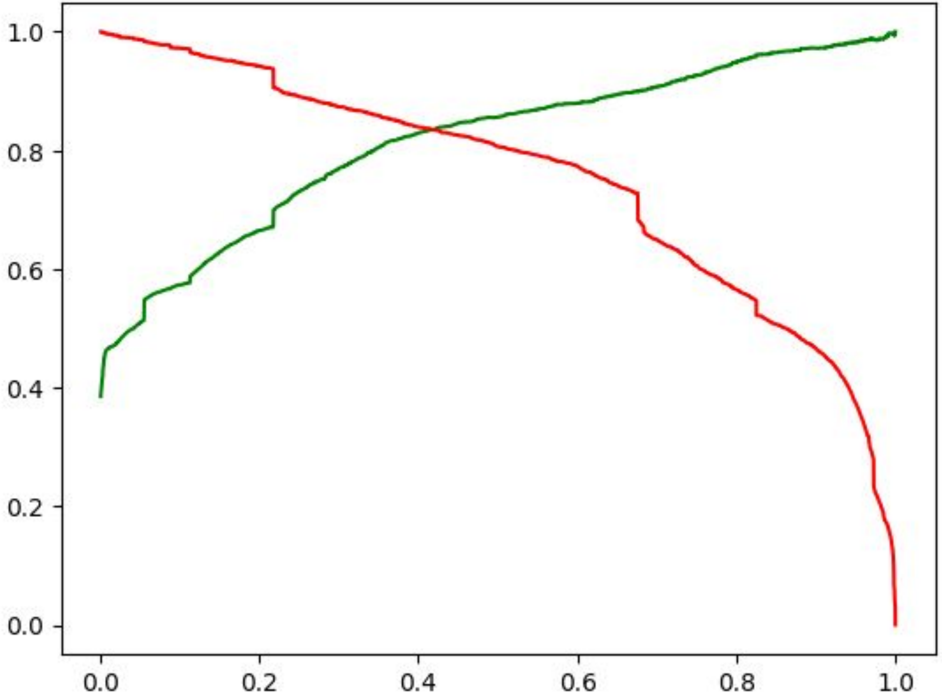
Model Evaluation - Accuracy - Sensitivity - Specificity



All 3 curves are intersecting at 0.3.

Accuracy is 0.8503 at this threshold.

Model Evaluation - Precision Recall Curve



Precision Recall Curve is intersecting at 0.4

Accuracy at that point is 0.8713

Key Findings & Insights

Lead Scores is probability of the lead to convert multiplied by 100
- Accuracy (87.29%): The model is correct 87.29% of the time.
- Precision (85.56%): High precision means fewer false positives. Important for businesses where false positives are costly
- Recall (80.63%): Model is catching more true hot leads.
- F1 Score (83.02%): The model is fairly balanced between avoiding false positives (precision) and not missing true positives (recall).
- Sensitivity: 0.8063 The model correctly identifies 80.63% of all actual hot leads.
- Specificity: 0.9147: 91.47% of actual cold leads were correctly classified as cold.
- False Positive Rate: 0.0853 : This means sales reps will spend minimal time on leads that won't convert.
- Positive Predictive Value: 0.8557: If the model predicts a lead is hot, there's an 85.57% chance it is actually hot.
- Negative Predictive Value: 0.8827: If the model predicts a lead as cold, there's an 88.27% chance it is actually cold.

Top three Features contributing to Model: Tags_Closed By Horizon, Tags_Will_revert after reading the email, tags_Busy are highly contributing feature with positive correlation.

Recommendation / Conclusion

1. **Usage of Lead Score** to enable Sales teams to Focus on Hot leads and save efforts on any random Cold leads. Here we are considering if lead score is more than 30 then it can be Hot lead and chance of getting converted is high.
2. **Optimize sales team efforts:** By categorizing Leads, if Lead Score >70 then High priority Leads (Time sensitive) then Immediate actions like Mailing, calling, Messaging. If score is between 30-70 then send promotional mails , discounted offers. If Score less than 50 then minimal efforts and remarketing.
3. **Sales strategy:** Leads with Tags like “Closed by Horizzon” and “Will revert after reading email” have a strong positive correlation with conversion.can be prioritized for follow-ups.Leads tagged as “Busy” → Follow up at different times/days to increase response rates.
4. **Variable Threshold based on Lead Sources:** Threshold can be varied based on leads acquired Source to get higher accuracy.
5. **Rebuilding and recalibration of model:** Customer behaviour changes over time which will impact coefficient and features of the model. Additional features can be required to get same recall and accuracy in future or we may need to drop some features or updating coefficient, model rebuilding or recalibration should be consider after every few years.