Lead Scoring Case Study
- Presentation -

Submitted by :

Hetal Khanapure

INTRODUCTION

Problem Statement :

- To Build a **Logistic Regression Model** to predict whether a lead for online courses for an education company i.e. X Education would be successfully converted as potential customer or not.
- Creating a model such that customers with high lead score have higher chances for conversion to potential customers can be considered as **HOT LEADS**. The ball park of the target lead conversion rate is around **80%.**
- The model should be built in such a way that it can be adjusted if the company's requirement changes in near future.

Business Objective:

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into **paying customers**.
- To build a Logistic Regression Model to assign lead score between **0 to 100**, so that company can target potential customers.
- Have to decide on threshold of the probability as cut-off above which lead can be predicted as converted and below not converted.
- **Lead Probability** is multiplied by 100 to get the **Lead Score** between 0 to 100.

METHODOLOGY

- Understanding data for analysis
- Data Pre-processing which include cleaning and preperation of the data
- Exploratory Data Analysis.
- Feature Scaling
- Splitting the data into Test and Train dataset.
- Using **RFE (Recursive Feature Elimination)** to identify best features.
- Building a logistic Regression model with feature selected by RFE.
- Eliminate all features with **high p-value** and **VIF** values and finalizing the model.
- Based on threshold probability will take it as cut-off value and predict the dependent variable.
- Calculate Lead Score.
- Performing Model Evaluation on various metrics like:
  - **Accuracy**
  - **Sensitivity**
  - **Specificity**
  - **Precision**
  - **Recall**

FEATURE ENGINEERING

Removing Columns Having Only One Unique Value

- **Magazine**
- **Receive More Updates About Our Courses**
- **Update me on Supply Chain Content**
- **Get updates on DM Content**
- **I agree to pay the amount through cheque**

These variables have only one value so we will drop these columns as not responsible in predicting a successful lead case

Removing Rows Having High Missing Value

- **Lead Source**

We will combine lead source which are less in count as other which resulted as following value counts:

- Google 2904
- Direct Traffic 2543
- Olark Chat 1755
- Organic Search 1154
- Reference 534
- Welingak Website 142
- Referral Sites 125
- Other 83

Removing Sales Related Columns

- **Last Activity**
- **Tags**
- **Last Notable Activity**

We need to identify hot leads from the data generated directly from source. Sales people added this sales related columns.. But we want to help them to decide which are hot leads before they approach them. So our analysis should be based on data generated by source.. Not by sales people

Imputing Null Values With Mode

- **Total Visits**
- **Page Views Per Visit**

Are continuous variables with outliers. Hence null values of these columns are imputed with mode.

EXPLORATORY DATA ANALYSIS

Lead Converted – 38.5%

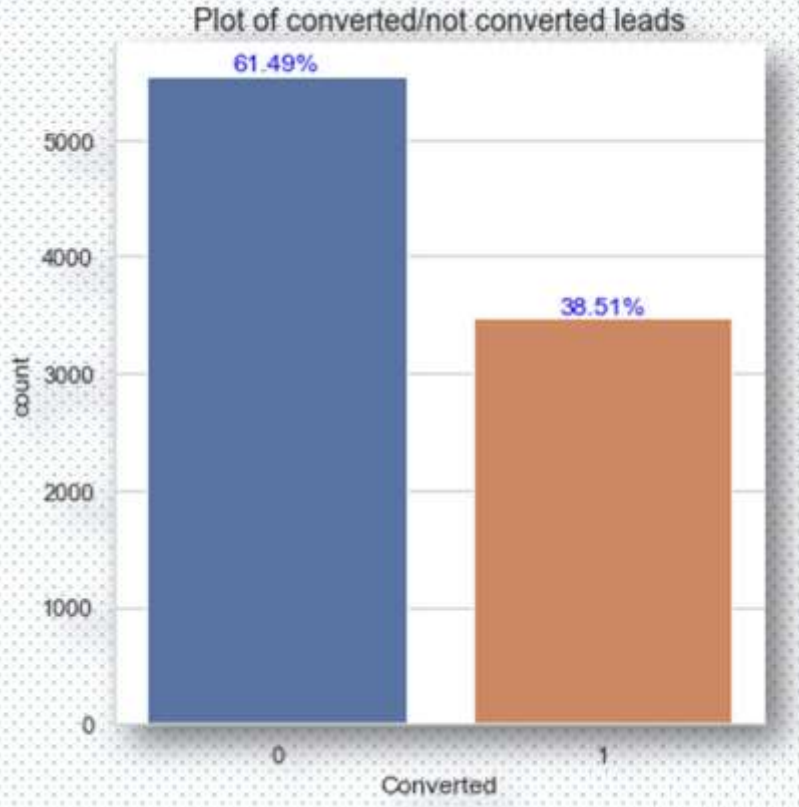
Lead Not – Converted – 61.5%

No. of converted leads: 3477

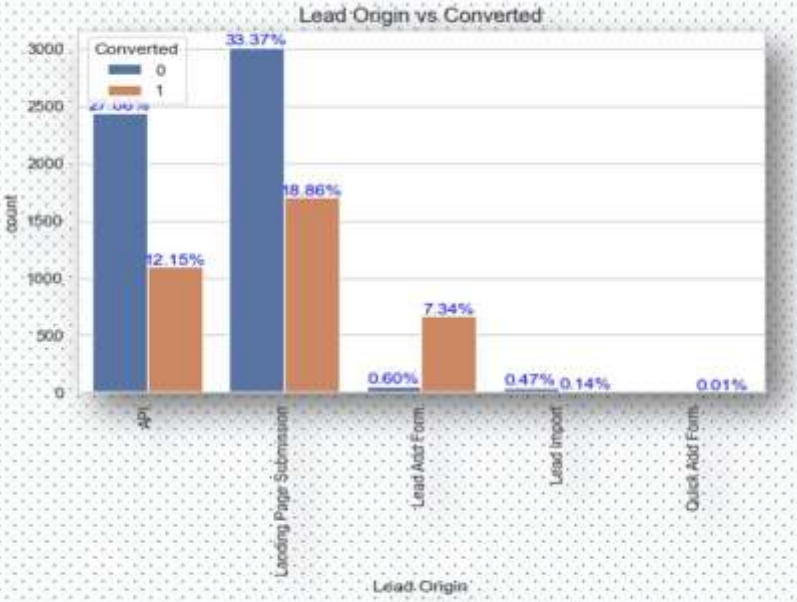
No. of not - converted leads: 5552

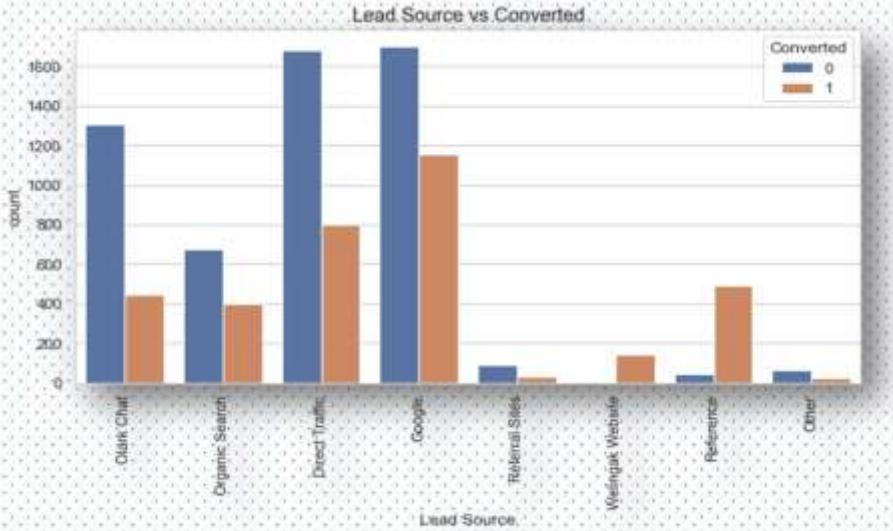
Percentage of converted leads: 38.50

.



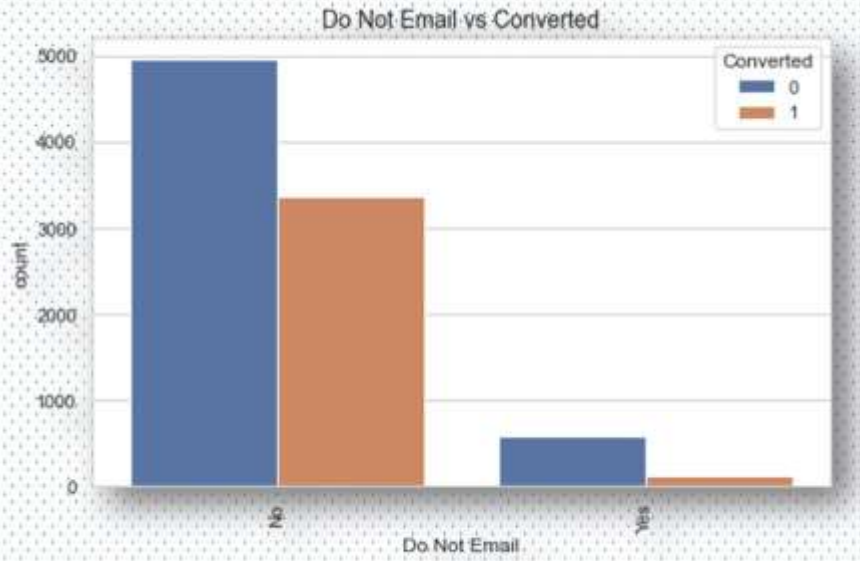
EXPLORATORY DATA ANALYSIS

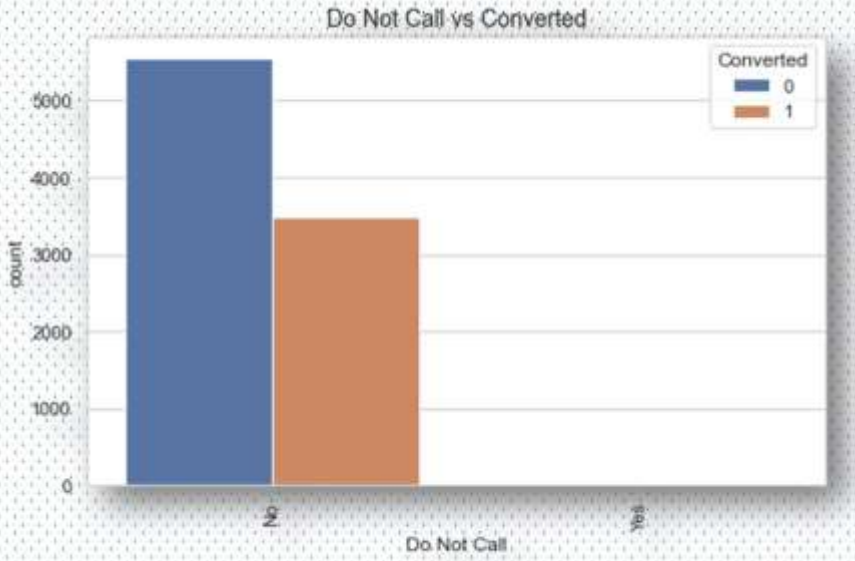




- Lead Conversion rate is **HIGH** when Lead Source is **Google**
- Lead Conversion rate is **HIGH** when Lead Source is **Welingak Website**
- Lead Conversion rate is **HIGH** when Lead Origin is **Landing Page Submission**

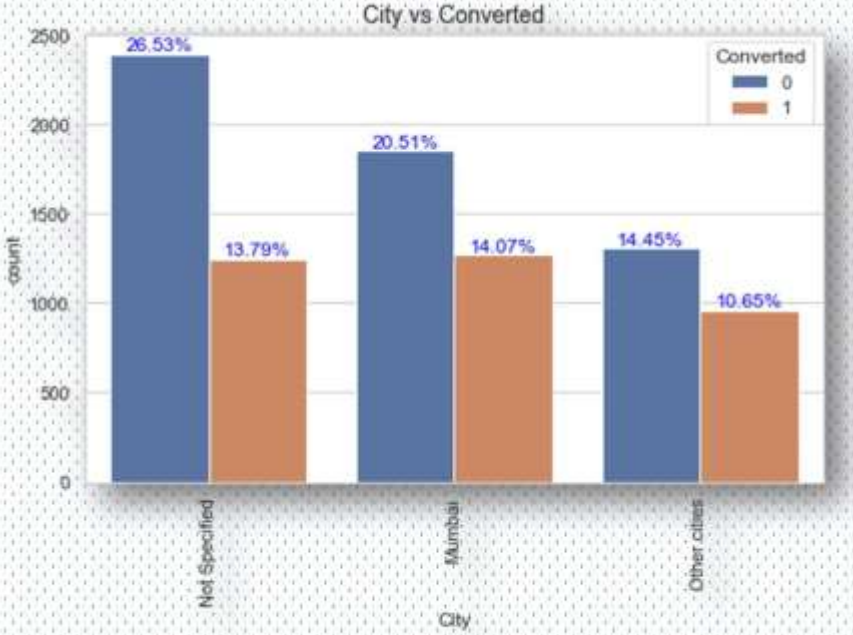
EXPLORATORY DATA ANALYSIS

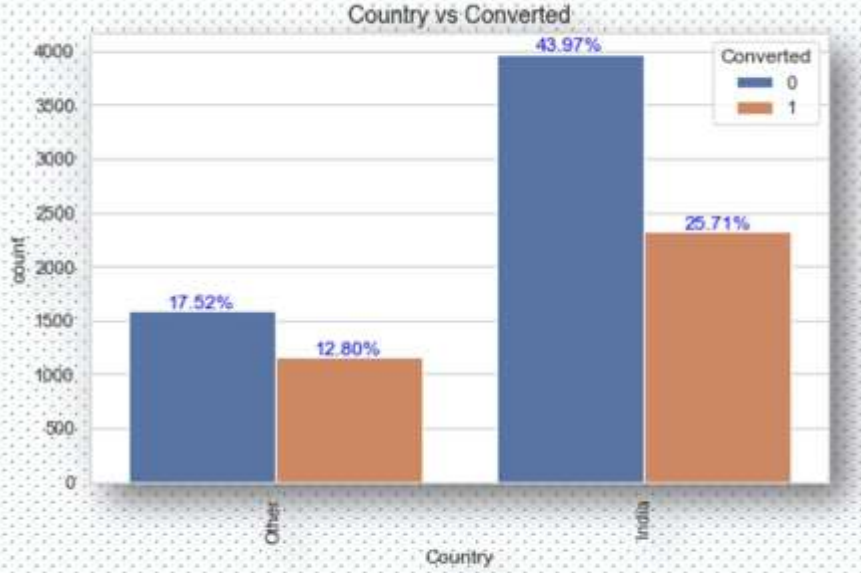




- Lead Conversion rate is **HIGH** when people want to get **Mail**
- Lead Conversion rate is **HIGH** when people want to get **Call**

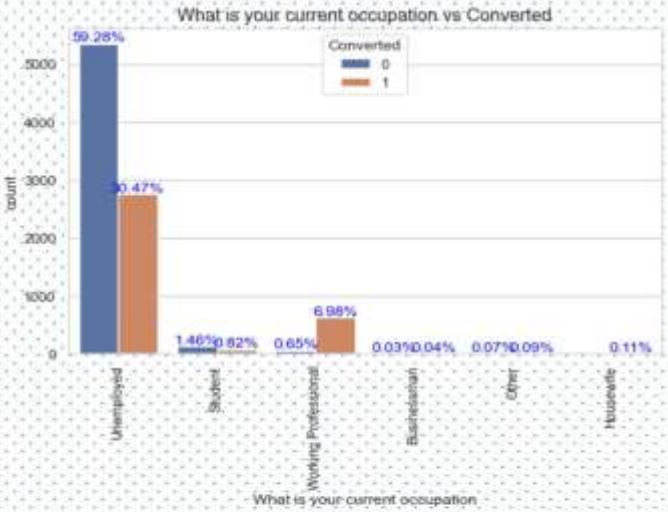
EXPLORATORY DATA ANALYSIS



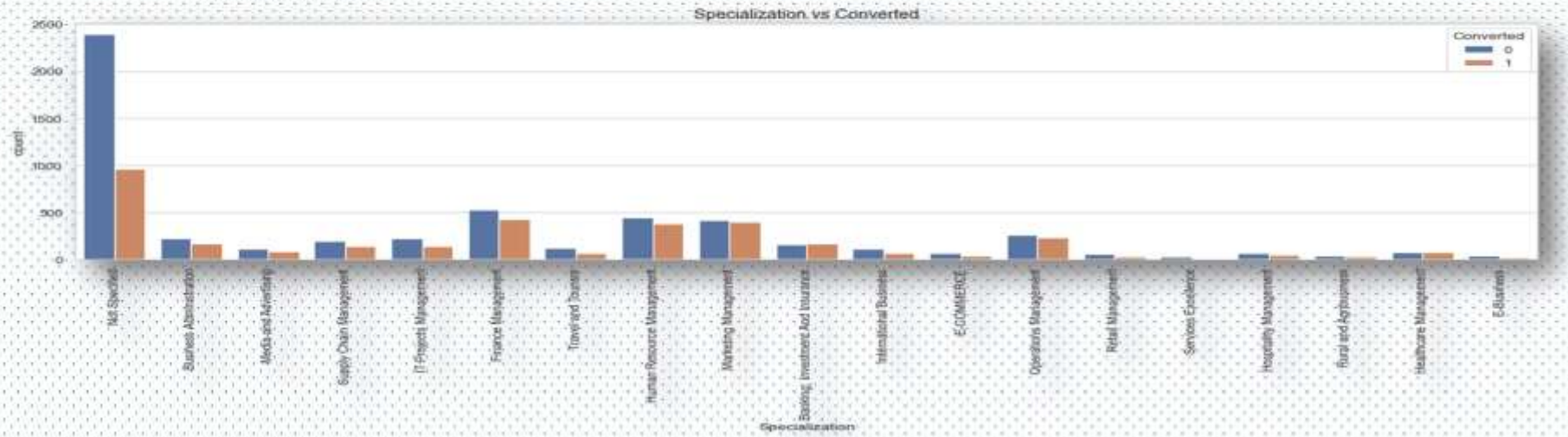


- Lead Conversion rate is **HIGH** in **Mumbai** and when City is **Not-Specified**
- Lead Conversion rate is **VERY LOW** in **Tier II Cities**

EXPLORATORY DATA ANALYSIS



- Lead Conversion rate is **HIGH** in **Specialization** and when user have **Not-Specified**
- Lead Conversion rate is **HIGH** in **Unemployed and 100%** conversion rate in **Housewife**
- **Working Professional** occupation leads are **likely** to be converted



DATA PREPARATION

DUMMY ENCODING

- **Lead Origin**
- **Lead Source**
- **Specialization**
- **What is your current occupation**
- **City**
- **Country**

Dummy Features For categorical variables with multiple levels(one-hot encoded)

FEATURE SCALING

We will Rescale following variables for better model evaluation:

- `TotalVisits` : The total number of visits made by the customer on the website.
- `Total Time Spent on Website` : The total time spent by the customer on the website.
- `Page Views Per Visit`: Average number of pages on the website viewed during the visits

Scaling helps in interpretation. Good to have all the variables on the same scale for the model to be easily interpretable

TEST – TRAIN SPLIT

- Origin data frame split into **train and test dataset.**
- Train dataset: **to train the model**
- Test dataset: **to evaluate the model**

RFE
(Recursive Feature Elimination)

- The predictor variables are more therefore will use **RFE.**
- Will pass **arbitory number** of variable to select, in our case will take **15**

```
logreg = LogisticRegression()

# running RFE with 15 variables as output
rfe = RFE(logreg, 15)
rfe = rfe.fit(X_train, y_train)

# create datafrme of the variable with rfe rank
rfe_df = pd.DataFrame({'variables': X_train.columns, 'Support': rfe.support_, 'Ranking': rfe.ranking_})
rfe_df.sort_values(by='Ranking')

# Rfe supported variable
col = X_train.columns[rfe.support_]
col
```

OUTPUT

```
Index(['Do Not Email', 'Total Time Spent on Website', 'LeadOrigin_Landing Page
Submission', 'LeadOrigin_Lead Add Form', 'LeadOrigin_Lead Import', 'LeadSource_Olark
Chat', 'LeadSource_Welingak Website', 'Specialization_Hospitality Management',
'Specialization_Not Specified', 'Current_Occupation_Housewife',
'Current_Occupation_Other', 'Current_Occupation_Student',
'Current_Occupation_Unemployed', 'Current_Occupation_Working Professional', 'City_Not
Specified'], dtype='object')
```

BUILDING THE MODEL

Generalized Linear Model Regression Result

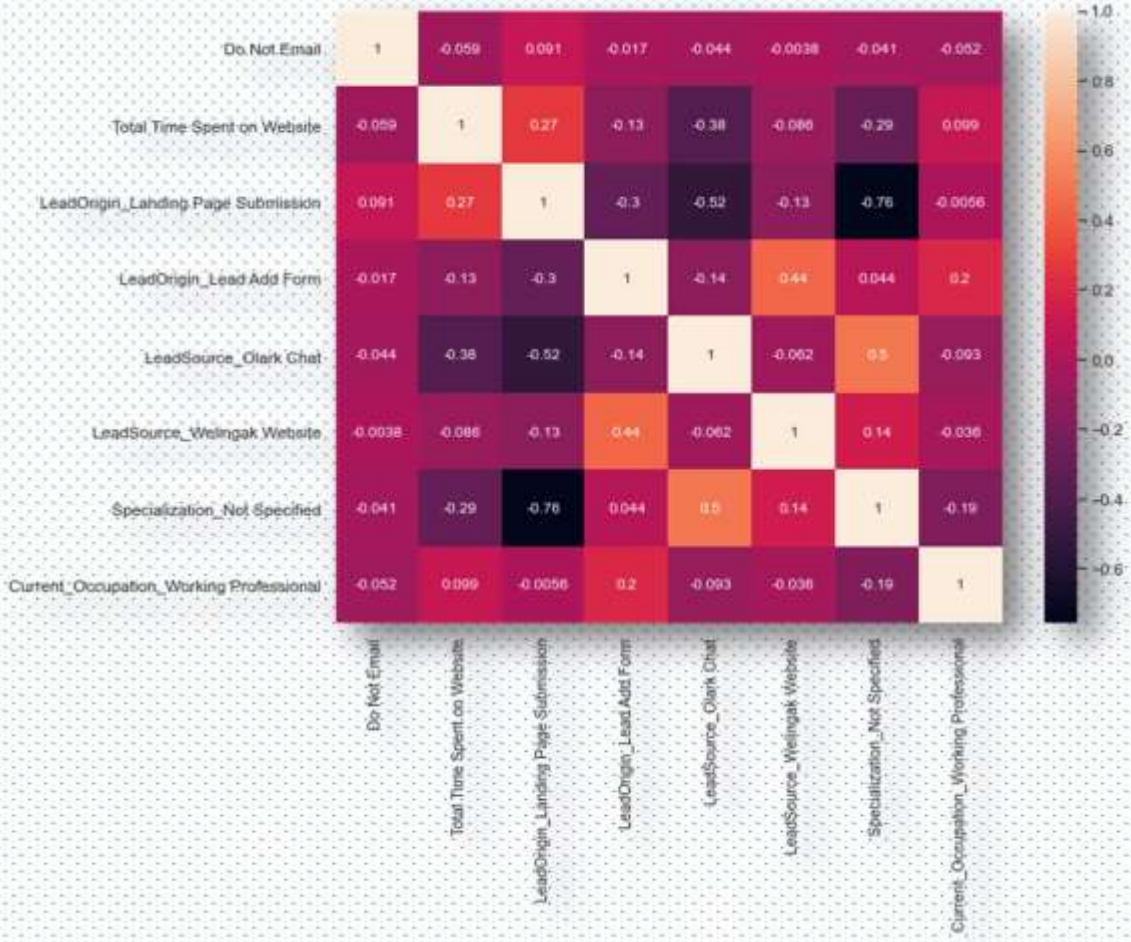
Dep. Variable:	Converted	**No. Observations:**	6320
Model:	GLM	**Df Residuals:**	6311
Model Family:	Binomial	**Df Model:**	8
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2947.3
Date:	Sat, 15 May 2021	**Deviance:**	5894.6
Time:	20:19:04	**Pearson chi2:**	8.53e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0533	0.116	-0.459	0.647	-0.281	0.175
Do Not Email	-1.1785	0.154	-7.637	0.000	-1.481	-0.876
Total Time Spent on Website	1.0719	0.038	28.394	0.000	0.998	1.146
LeadOrigin_Landing Page Submission	-0.9749	0.121	-8.031	0.000	-1.213	-0.737
LeadOrigin_Lead Add Form	3.4501	0.210	16.404	0.000	3.038	3.862
LeadSource_Olark Chat	0.8074	0.111	7.262	0.000	0.590	1.025
LeadSource_Welingak Website	2.6791	0.744	3.603	0.000	1.222	4.137
Specialization_Not Specified	-1.2464	0.116	-10.708	0.000	-1.475	-1.018
Current_Occupation_Working Professional	2.5361	0.179	14.183	0.000	2.186	2.887

Features	VIF	
4	LeadSource_Olark Chat	1.86
6	Specialization_Not Specified	1.82
3	LeadOrigin_Lead Add Form	1.46
5	LeadSource_Welingak Website	1.31
1	Total Time Spent on Website	1.23
2	LeadOrigin_Landing Page Submission	1.17
7	Current_Occupation_Working Professional	1.17
0	Do Not Email	1.10

BUILDING THE MODEL (cont…)

- All features **p-value and VIF values** lies under perfection
- Therefore these features are the **final for model creation**
- Model is now ready for **prediction**
- No **Multicollinearity** present in the model



VARIABLES IMPACTING THE CONVERSION RATE

- Do Not Email
- Total Time Spent on Website
- LeadOrigin_Landing Page Submission
- LeadOrigin_Lead Add Form
- LeadSource_Olark Chat
- LeadSource_Welingak Website
- Specialization_Not Specified
- Current_Occupation_Working Professional

PREDICTIN THE CONVERSION PROBABILITY & PREDICTED COLUMN

Creating a data frame with the **actual converted flag and the predicted probabilities**

	Converted	Converted_Prob	LeadID
0	0	0.143404	5493
1	0	0.112766	8064
2	0	0.062266	4716
3	0	0.191506	9117
4	1	0.432252	2402

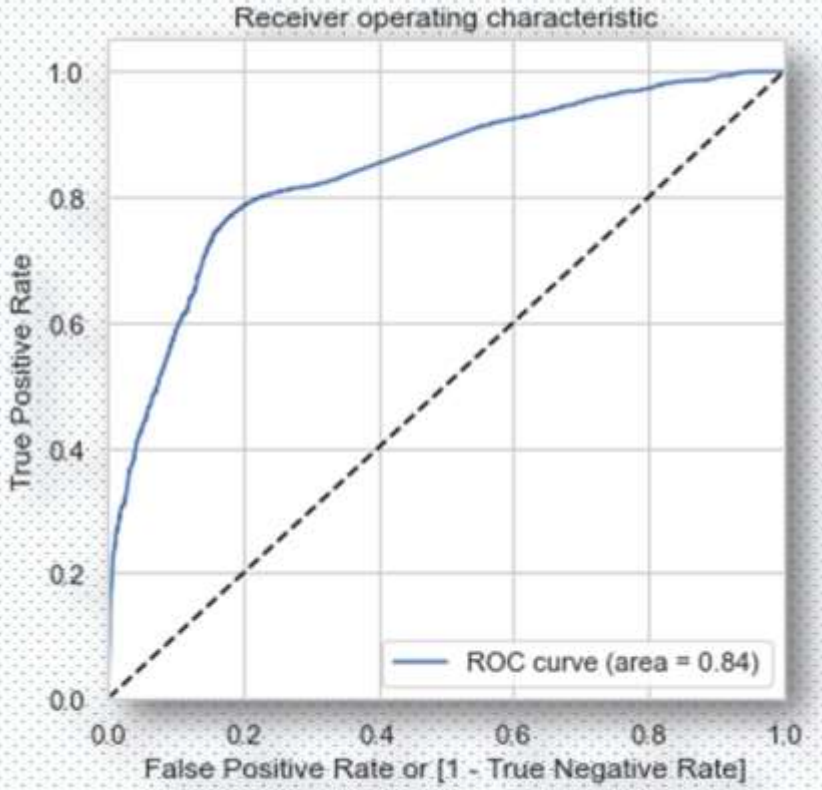
Creating new column **'predicted'** with 1 if **Converted_Prob > 0.5** else 0

	Converted	Converted_Prob	LeadID	predicted
0	0	0.143404	5493	0
1	0	0.112766	8064	0
2	0	0.062266	4716	0
3	0	0.191506	9117	0
4	1	0.432252	2402	0

ROC CURVE

An **ROC curve** demonstrates several things:

- It shows the tradeoff between **sensitivity and specificity** (any **increase in sensitivity** will be accompanied by a **decrease in specificity**
- The **closer the curve follows the left-hand border** and then the top border of the ROC space, the **more accurate the test**
- The **closer the curve comes to the 45-degree diagonal** of the ROC space, the **less accurate the test.**
- **Area under the ROC Curve = 0.84 i.e. 84% means model is a good fit**

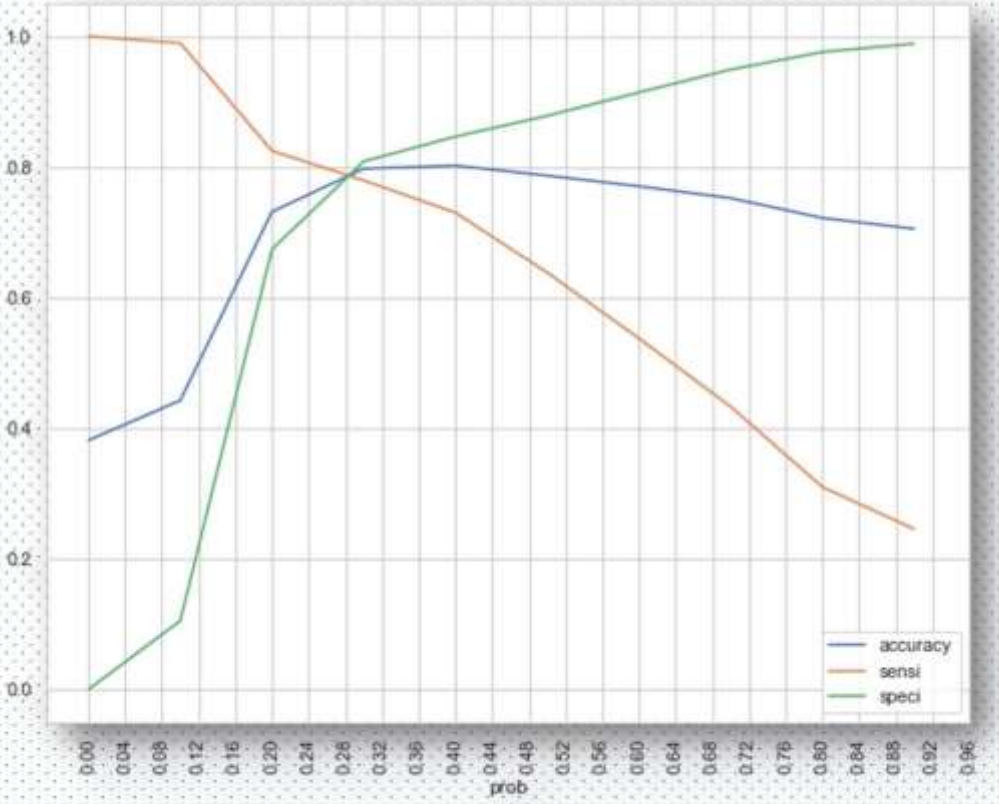


OPTIMAL PROBABILITY THRESHOLD

```
# Plot accuracy sensitivity and
specificity for various probabilities

cutoff_df.plot.line(x='prob',y=['accuracy'
,'sensi','speci'],figsize=(10,8))

plt.xticks(np.arange(0, 1,
0.04),rotation=90)

plt.show()
```

From the curve 0.28 is found to be the **optimum point** for cutoff probability

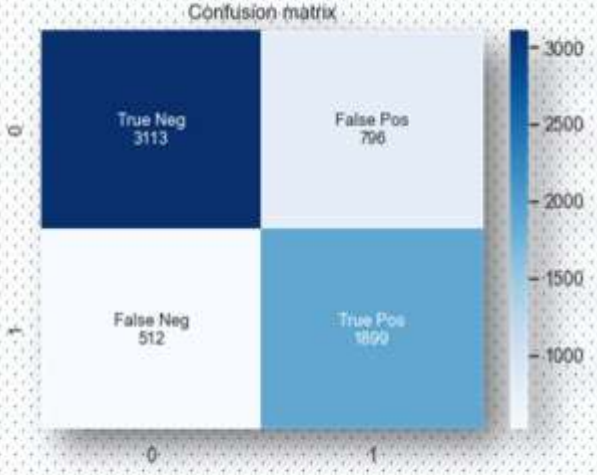
At this value of threshold all 3 metrices (**accuracy, sensitivity and specificity**) found to be above 80% which is acceptable value



Model Evaluation
– Sensitivity and Specificity on Train Data Set -

Top Records of the Train Dataset

	Converted	Converted_Prob	LeadID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted
0	0	0.143404	5493	0	1	1	0	0	0	0	0	0	0	0	0
1	0	0.112766	8064	0	1	1	0	0	0	0	0	0	0	0	0
2	0	0.062266	4716	0	1	0	0	0	0	0	0	0	0	0	0
3	0	0.191506	9117	0	1	1	0	0	0	0	0	0	0	0	0
4	1	0.432252	2402	0	1	1	1	1	1	0	0	0	0	0	1

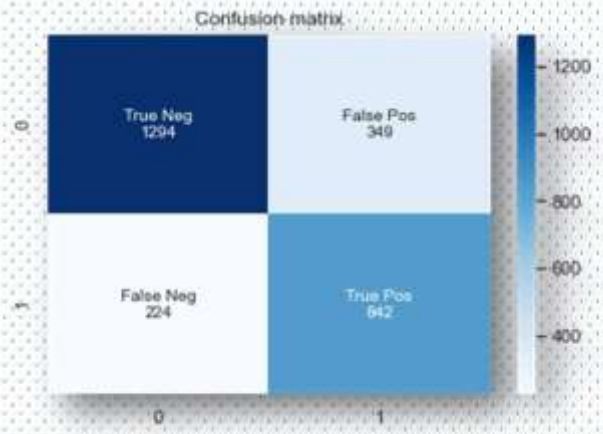


- Accuracy – 79.3%
- Sensitivity – 78.7 %
- Specificity – 79.6%
- False Positive Rate – 20%
- Positive Predictive Value – 70%
- Negative Predictive Value – 86%

Model Evaluation
– Sensitivity and Specificity on Test Data Set -

Top Records of the Test Dataset

	LeadID	Converted	Converted_Prob	final_predicted
0	4664	0	0.191506	0
1	5394	0	0.532147	1
2	1595	1	0.946008	1
3	2111	0	0.279018	0
4	9182	0	0.191506	0



Lead Score = 100 * Conversion Probability

	LeadID	Converted	Converted_Prob	final_predicted	Lead_Score
0	4664	0	0.191506	0	19
1	5394	0	0.532147	1	53
2	1595	1	0.946008	1	95
3	2111	0	0.279018	0	28
4	9182	0	0.191506	0	19

- Accuracy – **78.8%**
- Sensitivity – **78%**
- Specificity – **78.7%**
- False Positive Rate – **21%**
- Positive Predictive Value – **71%**
- Negative Predictive Value – **85%**

CONCLUSION

- **Logistic Regression Model**: Rather then predicting target variable value the model predict the probability of the value of target variable.

- Cut off value came as `0.28`
  - any lead with greater than 0.28: as **Hot Lead** (customer will convert)
  - any lead with 0.28: predicted as **Cold Lead** (customer will not convert)

- Total `8` columns/features to build Logistic Regression Model
- **Important Features**:
  - Do Not Email
  - Total Time Spent on Website
  - Lead Origin_Landing Page Submission
  - LeadOrigin_Lead Add Form
  - LeadSource_Olark Chat
  - LeadSource_Welingak Website
  - Specialization_Not Specified
  - Current_Occupation_Working Professional

- The final model has **Precision of 0.706**, this means **71%** of predicted hot leads are True Hot Leads.

- **The lead conversation rate** on the final predicted model is around **79% on train set and 78% on test set.**