Lead Scoring Case Study XEducation

Presentation By,

Sudheer N Poojari,

Bangera Anush Devendra,

Rajat Pravin Asare.

Jan 2023

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, but its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

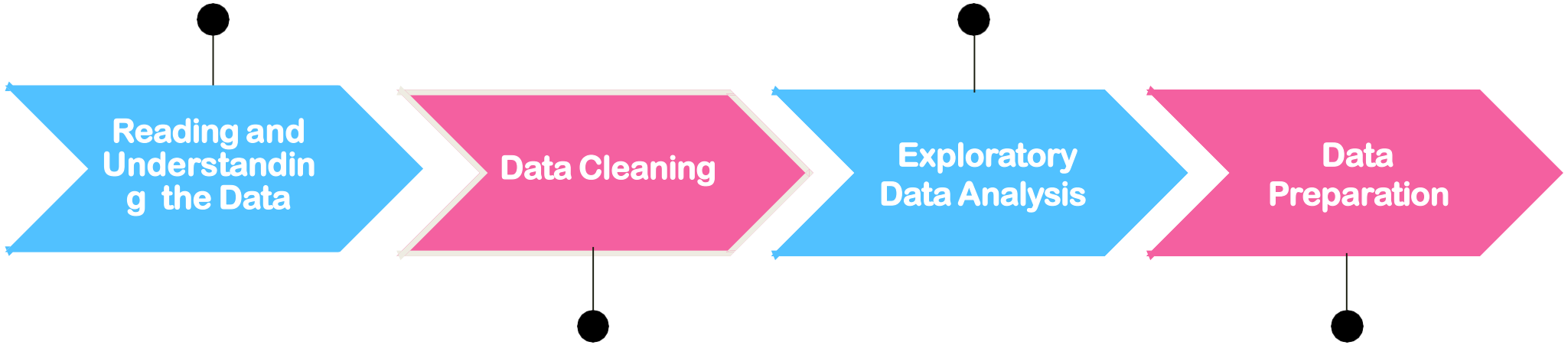
Business Objective

- X education wants to know the most promising leads.
- For that, they want to build a Model which identifies the hot leads.
- Deployment of the model for future use.

Solution Methodology-I

Importing and Observing the past data provided by the Company

Univariate and Bivariate analysis



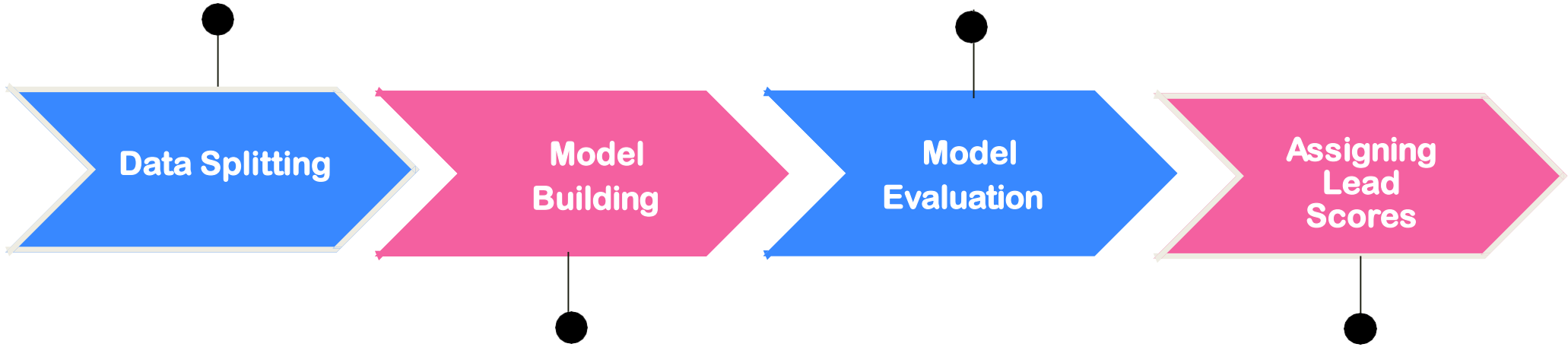
- Missing value imputation
- Removing duplicate data and other redundancies

- Outlier treatment
- Dropping unnecessary columns
- Dummy variable creation
- Feature standardization

Solution Methodology-II

- Performing a train-test split, we have chosen 70:30 ratio
- Re-scaling the features

- Evaluating model based on various evaluation metrics
- Finding the optimal probability threshold



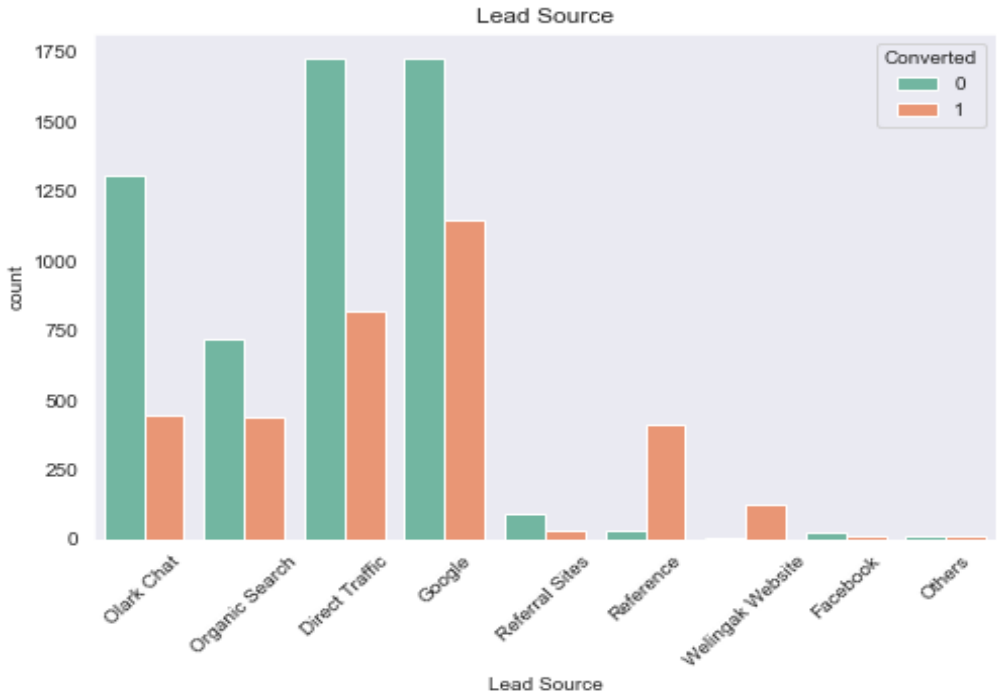
- Feature selection using RFE
- Manual feature elimination based on p-values and VIFs

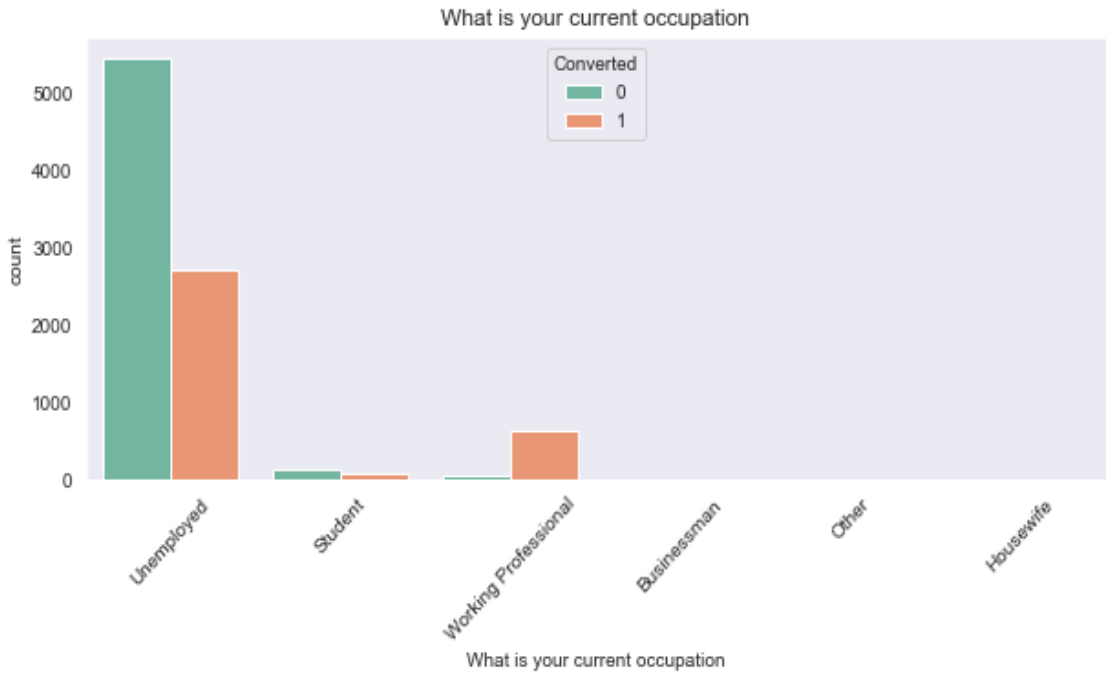
- Finalizing the first model
- Using predicted probabilities to calculate Lead Scores:
  Lead Score = Probability* 100

Data Exploration/Cleaning

- ‘Leads.csv’ contains Total Number of Rows =37, Total Number of Columns =9240.
- Prospect ID & Lead Number are two variables that are just indicative of the ID number of the Contacted People & so Lead Number column can be dropped.
- Replacing 'Select' value with NaN , Since it means no option is selected.
- Following columns have been dropped since percentage of missing value is more than 40%: How did you hear about X Education, Lead Quality, Lead Profile, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score .
- All the missing values of categorical columns have been imputed with respective mode/median value, such as Country, Specialization, What is your current occupation, What matters most to you in choosing a course,Tags,City.
- Rest missing values are under 2% so we can drop these rows which contain null values, Such as TotalVisits, Page Views Per Visit, Last Activity.
- Following columns have been, as their contribution is insignificant, such as 'Tags','Country','Search','Magazine','Newspaper Article','X Education Forums','Newspaper','Digital Advertisement','Through Recommendations','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','A free copy of Mastering The Interview'.

EDA

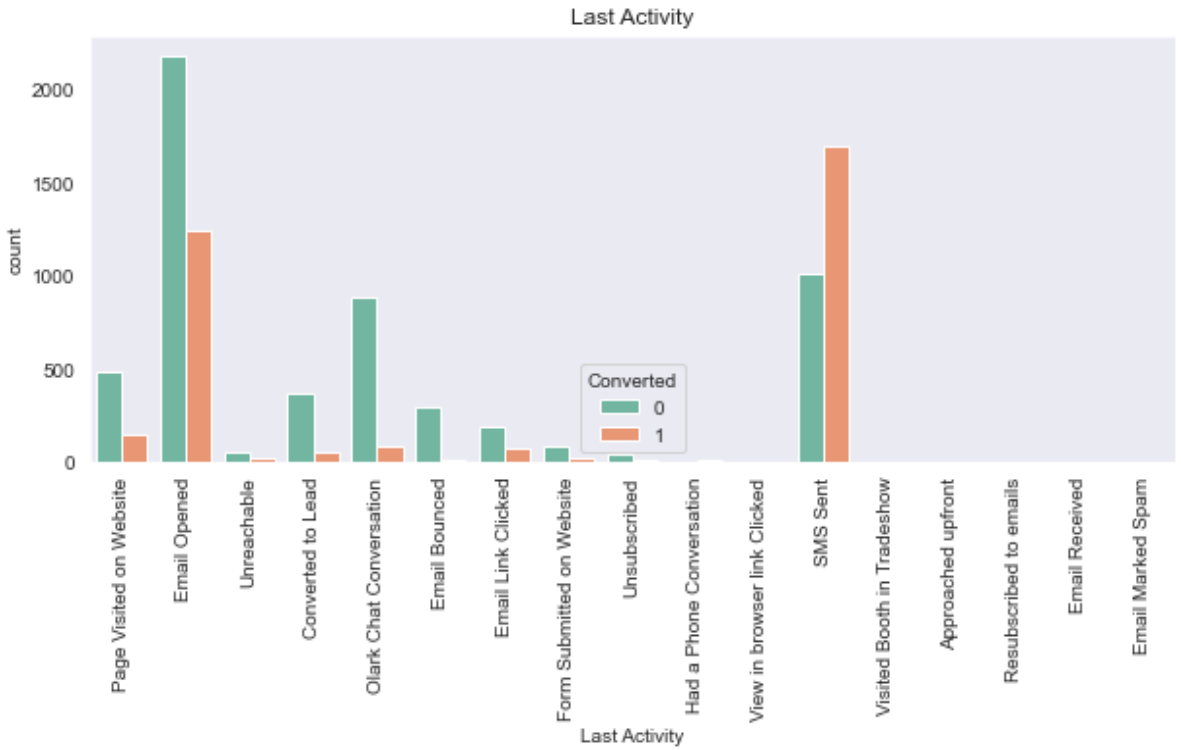


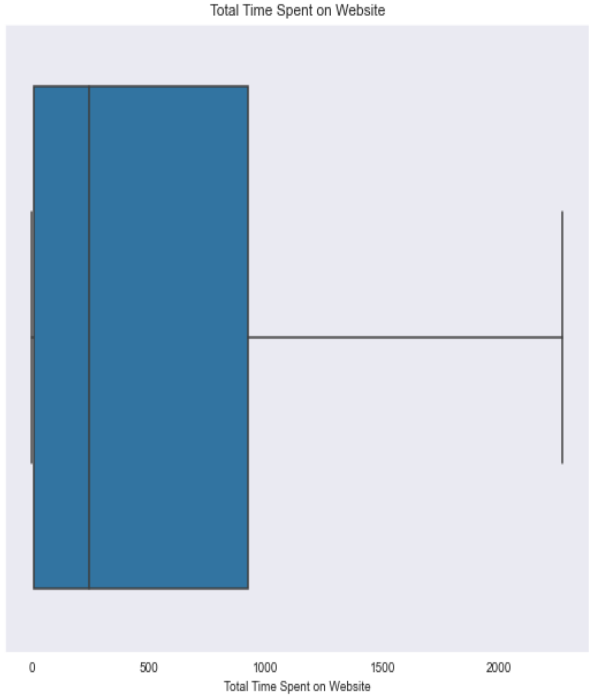


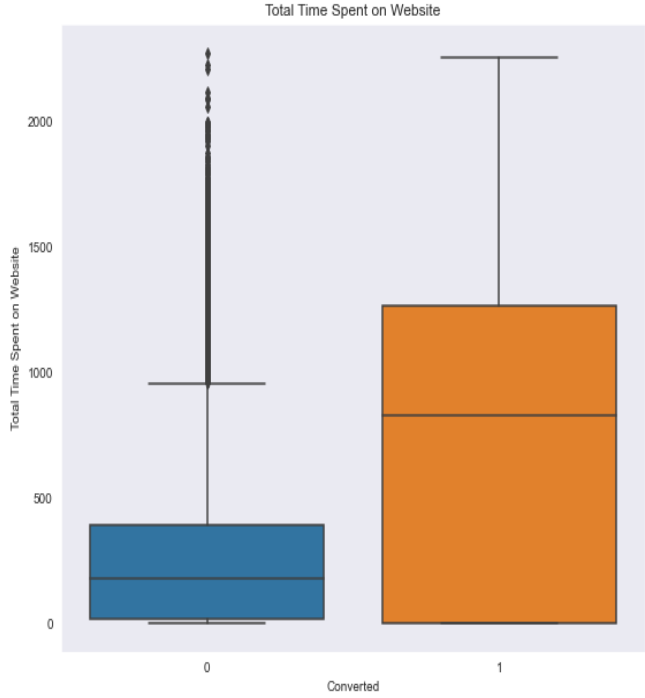
Lead Source: Need to focus on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

What is your current occupation: Working Professionals going for the course have high chances of joining it and Unemployed leads are the most in numbers but has around 30-35% conversion rate.

EDA (cont)







Last Activity:

1.Customers whose last activity was SMS Sent have higher conversion rate which is around 63%.
2.Customers whose last activity was Email Opened constitute majority of the customers. They have around 36% of conversion rate.

Total Time Spent on Website:

Leads spending more time on the website are more likely to be converted, a website should be made more engaging to make leads spend more time.

Model Building

Data Preparation for Modeling

Binary Variables Encoding:

- Variables which have binary (Yes/No) values have been encoding with 1 and 0.
- 1 denotes Yes whereas 0 denotes No.

Train – Test Split:

- The modified 'Leads' dataset has been split into Train and test dataset in the ratio 70:30.
- Train dataset has been used to train the model whereas Test dataset has been used to evaluate the model

Feature Scaling:

- It is important to have all variables on the same scale in order to avoid the dominance of variables with high magnitude in the model.
- "StandardScaler" function has been used to scale the data for modeling which brings all the data points into a standard normal distribution with mean at '0' and standard deviation at '1'.

Modeling-Using Logistic Regression

- Generalised Linear Model (GLM) from StatsModels library has been used to build the Logistic Regression Model.
- Initially, the model was built using 93 features present in X_train dataset.
- Most of the features were found to be insignificant. So, we needed to perform a feature selection technique.

Feature Selection using Recursive Feature Elimination (RFE):

- RFE is an optimization technique for finding the best-performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside, and then repeating the process with the rest of the features..
- We ran RFE to identify the top 20 features for the further model-building process.
- Insignificant features were dropped one by one after checking the P-value and Variance Inflation Factor (VIF).
- Accepted P-value should be kept below 0.05 and VIF should be less than 5.

Final Model – Regression Summary

- The final model contains the 12 most important features which satisfy all the selection criteria.
- All p-values are zero which indicates all these columns are statistically significant.
- If we see coefficients following attributes are positively impacting conversion:
  - Lead Source_Reference
  - What is your current occupation_Working Professional
  - Last Activity_Other_Activity
  - Last Activity_SMS Sent
  - Total Time Spent on Website
  - Lead Source_Olark Chat
  - Lead Source_Welingak Website
- Following negatively impacted:
  - Last Notable Activity_Modified
  - Last Activity_Olark Chat Conversation
  - Lead Origin_Landing Page Submission
  - Specialization_Others
  - Do Not Email

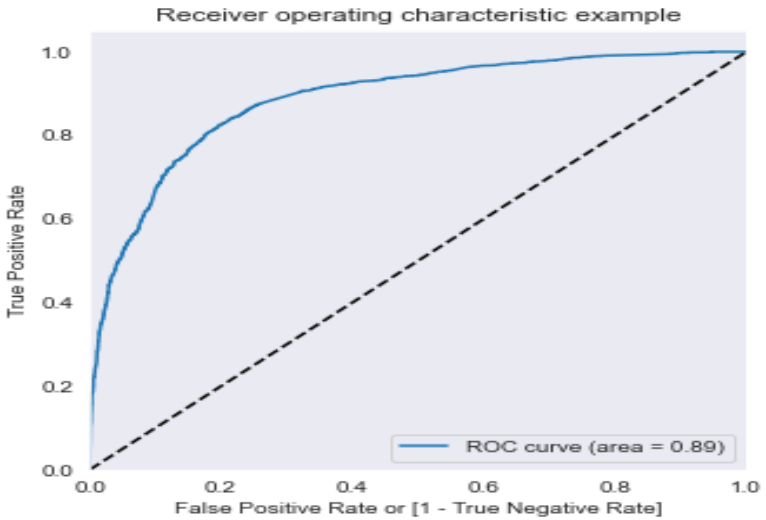
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6338
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2610.5
Date:	Mon, 23 Jan 2023	**Deviance:**	5221.0
Time:	23:46:55	**Pearson chi2:**	6.53e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4001
Covariance Type:	nonrobust		

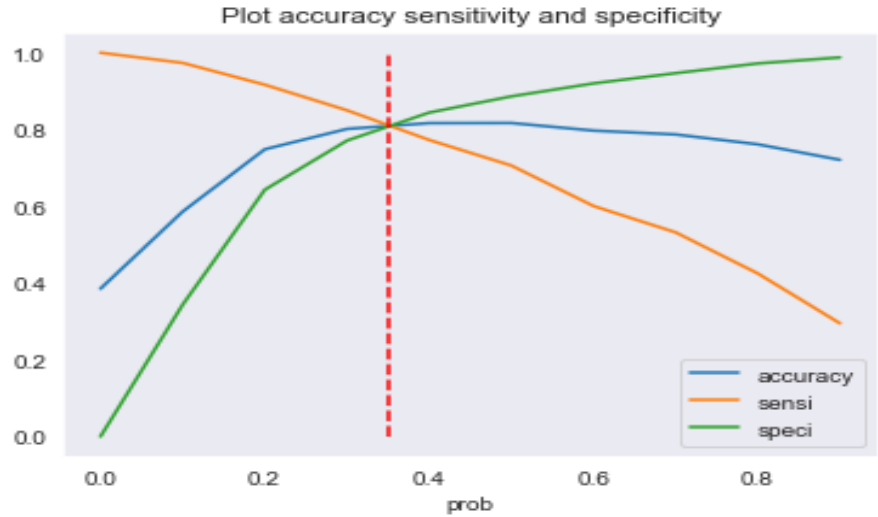
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0376	0.125	-0.300	0.764	-0.283	0.208
Do Not Email	-1.5218	0.177	-8.611	0.000	-1.868	-1.175
Total Time Spent on Website	1.0954	0.040	27.225	0.000	1.017	1.174
Lead Origin_Landing Page Submission	-1.1940	0.128	-9.360	0.000	-1.444	-0.944
Lead Source_Olark Chat	1.0819	0.122	8.847	0.000	0.842	1.322
Lead Source_Reference	3.3166	0.241	13.747	0.000	2.844	3.789
Lead Source_Welingak Website	5.8115	0.728	7.981	0.000	4.384	7.239
Last Activity_Olark Chat Conversation	-0.9613	0.171	-5.610	0.000	-1.297	-0.625
Last Activity_Other_Activity	2.1751	0.463	4.699	0.000	1.268	3.082
Last Activity_SMS Sent	1.2942	0.075	17.308	0.000	1.148	1.441
Specialization_Others	-1.2025	0.125	-9.582	0.000	-1.448	-0.957
What is your current occupation_Working Professional	2.6083	0.194	13.454	0.000	2.228	2.988
Last Notable Activity_Modified	-0.9004	0.081	-11.097	0.000	-1.059	-0.741

Evaluation Metrics

- **Receiver Operating Characteristics (ROC) Curve:**
  - By determining the Area Under the Curve (AUC) of the ROC curve, the goodness of the model is determined.
  - Since the ROC curve is close to the upper left part of the graph, it means this model is a very good model.
  - The value of AUC for our model is 0.89.



- **Plot accuracy sensitivity and specificity:**
  - Tradeoff between sensitivity and accuracy can be observed (cutoff = 0.35).
- **Precision and Recall plot:**
  - An ideal cutoff of 0.42 is observed from the recall and precision plot.
  - We will use both the cutoff and evaluate results for further predictions.



Final Model – Evaluation

Confusion Metrics		
Actual\Predicted	Positive	Negative
Positive	1409	325
Negative	197	792

Evaluation Metrics	
Accuracy	80
Sensitivity	80
Specificity	81

	Prospect ID	Converted	Converted_prob	Final_predicted	Lead_Score
0	3271	0	0.130342	0	13
1	1490	1	0.969057	1	97
2	7936	0	0.112570	0	11
3	4216	1	0.802999	1	80
4	3830	0	0.132924	0	13

- Confusion Metrics were created with thresholds at 0.35.
- Overall Model accuracy is 80%

Recommendations

The company should make calls to the leads by considering the below listed features, as these are more likely to get converted.

- ✓ Leads coming from the lead sources: Welingak Websites and Reference.
- ✓ Leads who are the working professionals.
- ✓ Leads whose last activity was SMS Sent.
- ✓ Leads who spent more time on the websites.
- ✓ Leads coming from the lead sources Olark Chat.

The company should not make calls to the leads by considering the below-listed features, as they are not likely to get converted.

- ✓ Leads whose last activity was Olark Chat Conversation.
- ✓ Leads whose lead origin is Landing Page Submission.
- ✓ Leads whose Specialization was Others.
- ✓ Leads who choose the option of Do not Email as yes.

Conclusion : Company should also focus on Lead Score (which are the probabilities obtained via algorithm) which are greater than 80% to expedite the conversion rate.

THANK YOU!