



# Lead Scoring Case Study

## Presenters:

Snehil Tiwari

Sowparni R

Siddhartha Puttuguntala

Batch - DS C47 | 2023





# Problem Statement

- ▶ To help X education company in identification of the most potential leads also called as 'Hot Leads', these leads are most likely to get converted into customers and can enroll for offered online courses by company.
- ▶ We need to build a logistic regression model that can assign a lead score falling between 0 to 100 to every single captured leads. Here, leads with higher lead score will have a higher conversion rate and leads with the lower lead score will have a lower conversion rate.
- ▶ We need to identify the 'driver variables' and also understand significance that will act as a strong indicators of lead conversion.
- ▶ Identify outliers, if there are any, in the given dataset and further justify the same.
- ▶ Need to consider the technical as well as the business aspects while we evaluate data set and build the model.
- ▶ Summarize conversion predictions using evaluation metrics parameters such as specificity, accuracy, precision and sensitivity.




# Data Understanding

► 'Leads.csv' contains all information regarding the generated leads collected from various sources and marketing activities.

- This data set file contains 9240 rows and 37 columns.
- Out of 37 columns, we have 7 numeric columns and 30 non-numeric columns commonly known as categorical columns.
- Leads current conversion rate is around 38.5%.

► 'Leads Data Dictionary.csv' is a data dictionary csv file that defines the meaning of all the variables present in the 'Leads' dataset.





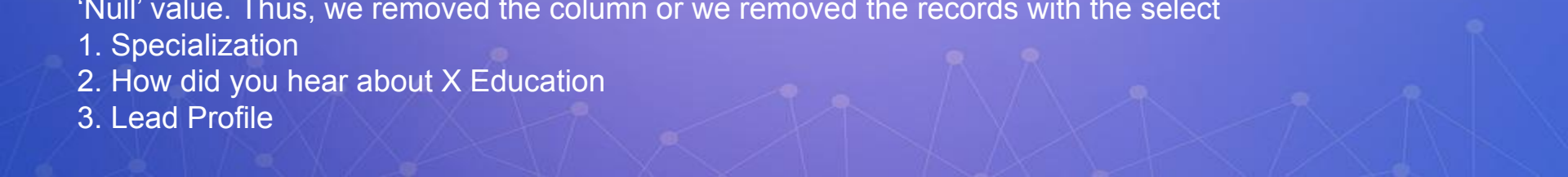
# Data Cleaning and Preparation

Leads.csv:

► Undermentioned columns contains Null values more than 30% initially:

1. Tags
2. Lead Quality
3. Asymmetrique Profile Index
4. Asymmetrique Activity Index
5. Asymmetrique Profile Score
6. Asymmetrique Activity Score

► Undermentioned columns contains default value of 'select' as a 'dominating value'. It is same as 'Null' value. Thus, we removed the column or we removed the records with the select

1. Specialization
  2. How did you hear about X Education
  3. Lead Profile
- 

# Data Cleaning and Preparation

► The columns that have higher percentage of single value compared to others are dropped because they did not add any valuable details to our predictions. These are listed below.

1. Magazine
2. Receive More Updates About Our Courses
3. Update me on Supply Chain Content
4. Get updates on DM Content
5. I agree to pay the amount through cheque
6. What matters most to you in choosing a course
7. How did you hear about X Education
8. Country
9. City

► There are also a couple of 'ID' columns that are used as record identifiers. We dropped them too.

1. Prospect ID
2. Lead Number





# Data Preparation for Model Building


## ► Create Dummy Variables:

The independent variables act as a dummy variables and allows easy interpretation and calculation of the odds ratios, that increases the stability and also significance of coefficients.

## ► Dummy variables got created for the following columns:

1. Lead Origin
2. Lead Source
3. Last Activity
4. Specialization
5. What is your current occupation
6. Last Notable Activity

► After creating Dummy variables, we got total 79 columns. We also dropped the above original columns from Data set.



# Data Preparation for Model Building

## ► Train – Test Split:

- The modified 'Leads' data set were split into 'Train & test' dataset in 70:30 ratio.
- Train dataset were then used to 'Train model' whereas 'Test dataset' are used to evaluate the model.
- Below are shapes of Data sets depicting the train and test split activity.

```
#After observinf X and y values.Let split the data into train and test in the ratio of 70 and 30%  
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size = 0.7, test_size = 0.3, random_state = 100)
```


```
#Printing train and test shape  
print('X_train', X_train.shape)  
print('X_test', X_test.shape)  
print('y_train', y_train.shape)  
print('y_test', y_test.shape)
```

```
X_train (4460, 72)  
X_test (1912, 72)  
y_train (4460,)  
y_test (1912,)
```



# Data Preparation for Model Building


## ► Feature Scaling:

- It is significant to have all the variables on same scale to avoid dominance of variables with high magnitude in model.
  - 'StandardScaler' technique were used to scale data for modeling, that brought all data points into standard normal distribution, mean at '0' and also standard deviation at '1'.
  - Scaling got performed on the below columns:
    1. Total Visits
    2. Total Time Spent on Website
    3. Page Views Per Visit
- 





# Model Building: Using Logistic Regression

- ▶ Generalised Linear Model (GLM) from StatsModels library were used to build Logistic Regression Model.
  - ▶ Initially, model was built using all 72 features present in the X\_train data set.
  - ▶ We then used 'feature selection technique' known as 'RFE' to do identification of the most important variables on which Model got built.
  - ▶ The Feature Selection using 'Recursive Feature Elimination' or (RFE)
  - ▶ The RFE is an optimization technique to find best performing subsets of features.
  - ▶ It is primarily based on the foundation of repeatedly constructing model and choosing the best (based on coefficients), setting the feature aside and then repeating the process with rest of features. This process is then applied, until all features in dataset are exhausted. Features are then ranked according to when they were eliminated.
  - ▶ We executed 'RFE' to identify the top 15 features for further model building process.
  - ▶ Insignificant features were dropped one by one after checking the 'P-value' and 'Variance Inflation Factor' (VIF).
  - ▶ Accepted P-value was considered to be within '0.05' and 'VIF' needs to be less than '5'.
- 

# Continued...

- ▶ Using GLM, we then build Model 1, where we found few columns with p-value that were higher than the acceptable range.
- ▶ We performed 'VIF' to check for multi-collinearity among different selected features.
- ▶ Below are screenshots of 'p-values and VIF'.

	coef	std err	z	P> z	[0.025	0.975]
const	1.0748	0.615	1.747	0.081	-0.131	2.280
Total Time Spent on Website	1.1271	0.047	24.215	0.000	1.036	1.218
Lead Origin_Lead Add Form	1.8870	0.977	1.931	0.054	-0.029	3.803
Lead Source_Olark Chat	1.3898	0.114	12.179	0.000	1.166	1.613
Lead Source_Reference	1.9582	1.003	1.961	0.051	-0.009	3.922
Lead Source_Welingak Website	4.1371	1.402	2.950	0.003	1.389	6.886
Last Activity_Email Bounced	-1.5410	0.353	-4.200	0.000	-2.252	-0.830
Last Activity_Had a Phone Conversation	1.6096	1.007	1.598	0.110	-0.385	3.594
Last Activity_SMS Sent	0.9273	0.083	11.127	0.000	0.764	1.091
What is your current occupation_Student	-1.9547	0.652	-2.998	0.003	-3.233	-0.677
What is your current occupation_Unemployed	-1.9023	0.616	-3.086	0.002	-3.111	-0.694
What is your current occupation_Working Professional	0.8095	0.844	0.948	0.344	-0.853	1.872
Last Notable Activity_Had a Phone Conversation	21.4492	1.19e+04	0.002	0.999	-2.34e+04	2.34e+04
Last Notable Activity_Modified	-0.8517	0.089	-9.521	0.000	-1.027	-0.676
Last Notable Activity_Unreachable	2.5476	0.802	3.177	0.001	0.978	4.119
Specialization_Services Excellence	-2.1327	0.902	-2.366	0.018	-3.900	-0.366

	Features	VIF
1	Lead Origin_Lead Add Form	69.33
3	Lead Source_Reference	53.91
4	Lead Source_Welingak Website	16.54
6	Last Activity_Had a Phone Conversation	2.85
11	Last Notable Activity_Had a Phone Conversation	2.85
9	What is your current occupation_Unemployed	2.41
7	Last Activity_SMS Sent	1.69
12	Last Notable Activity_Modified	1.57
2	Lead Source_Olark Chat	1.37
10	What is your current occupation_Working Profes...	1.31
0	Total Time Spent on Website	1.28
5	Last Activity_Email Bounced	1.06
8	What is your current occupation_Student	1.05
13	Last Notable Activity_Unreachable	1.01
14	Specialization_Services Excellence	1.01

# Continued...

- We dropped the column - Lead Origin\_Lead Add Form as it was having a high VIF value. After dropping this, we then build another model. Below are the snapshot of the re-built model.
- But, we still had a few columns with the higher p-values. Also, the VIF values are within acceptable ranges.

	coef	std err	z	P> z	[0.025	0.975]
const	1.0756	0.615	1.749	0.080	-0.130	2.281
Total Time Spent on Website	1.1239	0.048	24.199	0.000	1.033	1.215
Lead Source_Olark Chat	1.3877	0.114	12.181	0.000	1.164	1.611
Lead Source_Reference	3.8367	0.242	15.872	0.000	3.363	4.310
Lead Source_Welingak Website	6.0171	1.010	5.960	0.000	4.038	7.996
Last Activity_Email Bounced	-1.5429	0.362	-4.256	0.000	-2.253	-0.832
Last Activity_Had a Phone Conversation	1.6074	1.007	1.597	0.110	-0.366	3.581
Last Activity_SMS Sent	0.9278	0.083	11.138	0.000	0.765	1.091
What is your current occupation_Student	-1.9545	0.652	-2.999	0.003	-3.232	-0.677
What is your current occupation_Unemployed	-1.8991	0.616	-3.082	0.002	-3.107	-0.691
What is your current occupation_Working Professional	0.6091	0.644	0.946	0.344	-0.653	1.871
Last Notable Activity_Had a Phone Conversation	21.4462	1.19e+04	0.002	0.999	-2.34e+04	2.34e+04
Last Notable Activity_Modified	-0.8525	0.089	-9.537	0.000	-1.028	-0.677
Last Notable Activity_Unreachable	2.5426	0.802	3.171	0.002	0.971	4.114
Specialization_Services Excellence	-2.1331	0.901	-2.368	0.018	-3.899	-0.368

Features	VIF
Last Activity_Had a Phone Conversation	2.85
Last Notable Activity_Had a Phone Conversation	2.85
What is your current occupation_Unemployed	2.41
Last Activity_SMS Sent	1.69
Last Notable Activity_Modified	1.57
Lead Source_Olark Chat	1.37
What is your current occupation_Working Profes...	1.31
Lead Source_Reference	1.29
Total Time Spent on Website	1.28
Lead Source_Welingak Website	1.07
Last Activity_Email Bounced	1.06
What is your current occupation_Student	1.05
Last Notable Activity_Unreachable	1.01
Specialization_Services Excellence	1.01



# Continued...

- ▶ We re-built model after dropping column - Last Notable Activity\_Had a Phone Conversation
- ▶ Below are snapshot of Model Statistics - Still one column has high p-value.

	coef	std err	z	P> z	[0.025	0.975]
const	1.0790	0.615	1.755	0.079	-0.126	2.284
Total Time Spent on Website	1.1226	0.046	24.179	0.000	1.032	1.214
Lead Source_Olark Chat	1.3866	0.114	12.170	0.000	1.163	1.610
Lead Source_Reference	3.8363	0.242	15.868	0.000	3.362	4.310
Lead Source_Welingak Website	6.0159	1.010	5.958	0.000	4.037	7.995
Last Activity_Email Bounced	-1.5412	0.363	-4.251	0.000	-2.252	-0.831
Last Activity_Had a Phone Conversation	2.9337	0.799	3.670	0.000	1.367	4.500
Last Activity_SMS Sent	0.9261	0.083	11.120	0.000	0.763	1.089
What is your current occupation_Student	-1.9550	0.652	-2.999	0.003	-3.233	-0.677
What is your current occupation_Unemployed	-1.8993	0.616	-3.082	0.002	-3.107	-0.691
What is your current occupation_Working Professional	0.6088	0.644	0.946	0.344	-0.653	1.871
Last Notable Activity_Modified	-0.8613	0.089	-9.645	0.000	-1.036	-0.686
Last Notable Activity_Unreachable	2.5391	0.802	3.167	0.002	0.968	4.110
Specialization_Services Excellence	-2.1328	0.901	-2.368	0.018	-3.898	-0.368

# Continued..

- ▶ We performed the Model rebuilding activity again after dropping column - What is your current occupation\_Working Professional
- ▶ For this Model, all the features have p-value under '0.05' and VIF values are also in range. So, we eliminated Multicollinearity and retained the most significant features.

Features	VIF
What is your current occupation_Unemployed	2.18
Last Activity_SMS Sent	1.53
Last Notable Activity_Modified	1.50
Lead Source_Olark Chat	1.35
Total Time Spent on Website	1.24
Lead Source_Reference	1.17
Lead Source_Welingak Website	1.07
Last Activity_Email Bounced	1.06
What is your current occupation_Student	1.04
Last Activity_Had a Phone Conversation	1.01
Last Notable Activity_Unreachable	1.01
Specialization_Services Excellence	1.00

	coef	std err	z	P> z	[0.025	0.975]
const	1.6392	0.187	8.749	0.000	1.272	2.008
Total Time Spent on Website	1.1225	0.046	24.179	0.000	1.032	1.214
Lead Source_Olark Chat	1.3874	0.114	12.178	0.000	1.164	1.611
Lead Source_Reference	3.8372	0.242	15.872	0.000	3.363	4.311
Lead Source_Welingak Website	6.0159	1.010	5.958	0.000	4.037	7.995
Last Activity_Email Bounced	-1.5359	0.382	-4.242	0.000	-2.246	-0.828
Last Activity_Had a Phone Conversation	2.9353	0.789	3.673	0.000	1.369	4.502
Last Activity_SMS Sent	0.9288	0.083	11.164	0.000	0.766	1.092
What is your current occupation_Student	-2.5161	0.284	-8.850	0.000	-3.073	-1.959
What is your current occupation_Unemployed	-2.4607	0.187	-13.145	0.000	-2.828	-2.094
Last Notable Activity_Modified	-0.8613	0.089	-9.644	0.000	-1.036	-0.686
Last Notable Activity_Unreachable	2.5412	0.802	3.170	0.002	0.970	4.112
Specialization_Services Excellence	-2.1154	0.897	-2.360	0.018	-3.873	-0.358



# Final Model and Interpretation

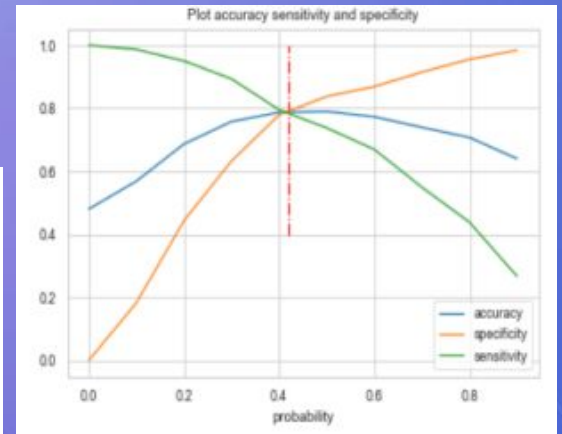
- ▶ Final model contains 12 most important features, which satisfies all selection criteria.
- ▶ Prediction was performed using the final Model and optimal probability Threshold value came out to be 0.42.
- ▶ Plot accuracy sensitivity and specificity:

Trade-off between sensitivity and accuracy can be observed (cutoff = 0.42)

▶ Confusion Matrix:

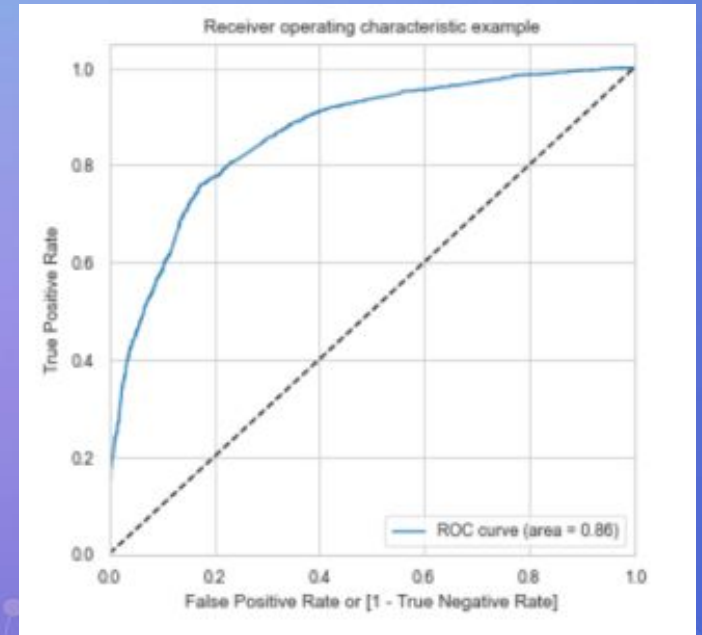
- Accuracy – 78.54%
- Sensitivity – 77.86%
- Specificity – 79.17%
- False Positive Rate – 20.82%
- Positive predictive value – 77.53%

Predicted	0	1	All
Actual			
0	1836	483	2319
1	474	1667	2141
All	2310	2150	4460



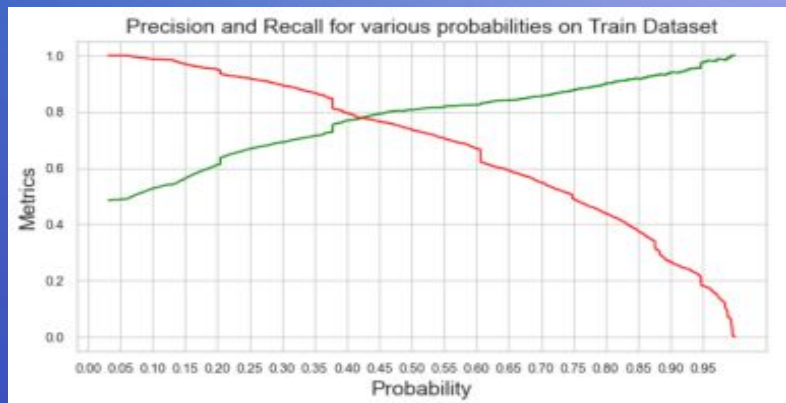
# Evaluation Metrics

- ▶ 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 0.86.

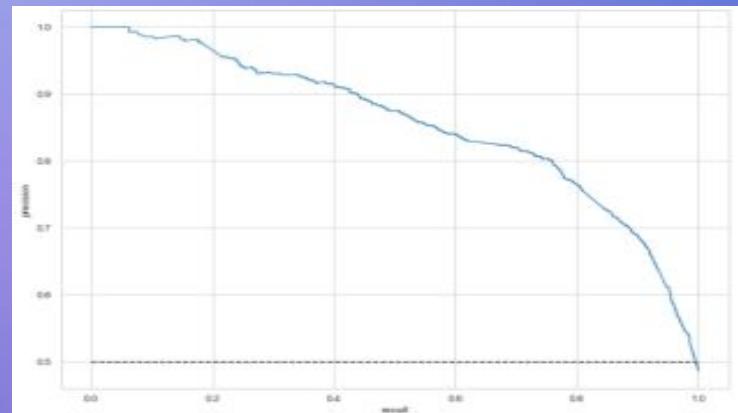


# Continued...

- Using the Precision and the Recall method , we checked parameters such as Accuracy and other evaluation metrics.
  - Using Precision and Recall, we are getting optimal threshold as 0.42



- We are getting AUC Score = 85.46%



- Precision and Recall Curve

# Predictions on Test Dataset

► Using the optimal threshold of 0.42, different metrics were calculated on Test dataset and below are details:

► Confusion Matrix:

Predicted	0	1	All
Actual			
0	787	201	988
1	199	725	924
All	986	926	1912

► Accuracy – 79%

► Sensitivity – 78.46%

► Specificity – 79.65%

► Precision – 78.29%

► Recall – 78.46%

► AUC Score – 85.37%

► Classification Report

	precision	recall	f1-score	support
0	0.80	0.80	0.80	988
1	0.78	0.78	0.78	924
accuracy			0.79	1912
macro avg	0.79	0.79	0.79	1912
weighted avg	0.79	0.79	0.79	1912



# Conclusion and Recommendations

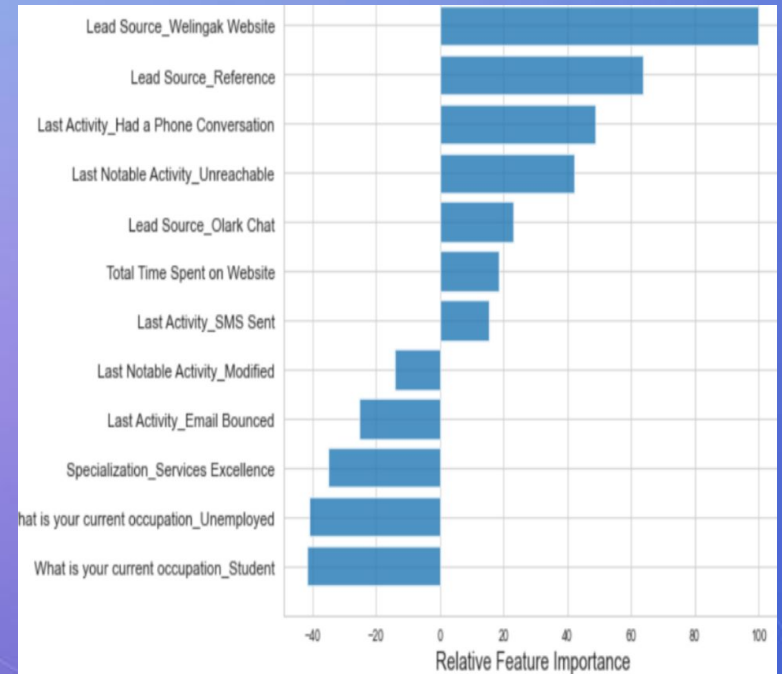
► **Lead Score is computed based on the Probability.** Higher the probability, higher the lead score. Resulting in higher the chance of converting the potential Lead and will lead to increase in overall revenue.

► **Followings are top three features** that contributed in decision making and understanding, whether the given lead will get converted or not .

1. Lead Source\_Welingak Website
2. Lead Source\_Reference
3. Last Activity\_Had a Phone Conversation

► **Top three categories that contribute to decision**

1. Lead Source - Students who have Lead Source as 'Reference' are more likely to get converted.
2. Lead Source – Students browsing the courses Welingak Website have very high chances of getting converted .
3. Last Notable Activity – Students with Last notable activity as Unreachable tend to convert more while opting for the course.







# Thank You