

# SUMMARY REPORT

## LEAD SCORE CASE STUDY

### PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.

### Business Understanding

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. Now, although X Education gets a lot of leads, 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 GOAL

X Education requires to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company requires to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, has given a ballpark of the target lead conversion rate to be around 80%.

### GOALS OF CASE STUDY

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so we will need to handle these as well

### DATA UNDERSTANDING

- Provided with a leads dataset from the past with around 9000 data points.

- This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.

## TARGET VARIABLE

- The **target variable**, in this case, is the column '**Converted**' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted

## Exploratory Data Analysis Approach

The dataset is being analysed using following steps:

- Data Sourcing
- Data Cleansing
- Binary Mapping
- Dummy variable Creation
- Segmented Univariate Analysis
- Bivariate Analysis
- Arriving at Insights

Detecting Outliers:-

- Numerical columns are analysed for identifying outliers using **Box plots**
- Outlier Datapoints are detected using **Quantile value method**
- Datapoints capped to an upper quantile of 0.95 and lower quantile of 0.05

## Business Insights

- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
- Google and Direct traffic generates higher number of leads.
- Conversion Rate of reference leads and leads through welingak website is high
- Recommended to generate more leads through Reference and Wellingak Website
- Focus on improving lead conversion through olark chat, organic search, direct traffic, and google
- The median time spend on website for converted leads is more than that of not converted. Website should be made more engaging to make leads spend more time.
- Working Professionals going for the course have high chances of joining it
- Unemployed leads are the most in numbers and has a good conversion rate compared to others. Focus on unemployed category

## MODEL BUILDING

- TEST-TRAIN SPLIT
- FEATURES SCALING using Standard Scalar
- Feature selection using RFE
- Run the model with RFE selected variables
- Models checked for p-values and VIF
- Final model arrived upon with reasonable VIF values and p-values
- Creating a dataframe with the actual conversion flag and the predicted probabilities
- Creating new column 'predicted' with 1 if Convert\_Prob > 0.5 else 0

Plotting the ROC Curve

Finding Optimal Cutoff Point as 0.3

- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity
- plot accuracy sensitivity and specificity for various probabilities.
- Optimal cut off obtained at probability of 0.3

## MODEL EVALUATION

- Recalculating column 'predicted' with 1 if Convert\_Prob > 0.3 else 0
- Assigning Lead Score for each leads in a scale of 1-100
- Recalculating model evaluation metrics
- Confusion Matrix and Accuracy
- Precision and Recall Trade off
- Making predictions on the test set
- Evaluating metrics on Test predictions

## MODEL EVALUATION METRICS for Cutoff of 0.3

- CONFUSION MATRIX

[[3693 212]

[ 299 2147]]

- Accuracy :- 0.91954
- Sensitivity :- 0.877
- Specificity :- 0.9457
- False postive rate:-0.05

- Positive predictive value:-0.910
- Negative predictive value:-0.925

## **PRECISION AND RECALL**

- Precision :- 0.9101

$TP / TP + FP$

- Recall :-0.8777

$TP / TP + FN$

- F1-score :- 0.8936

$2 * (Precision * Recall) / (Precision + Recall)$

## **TEST DATA PREDICTION and METRICS**

- CONFUSION MATRIX

[[1635 99]

[ 138 851]]

- Accuracy :- 0.9129
- Sensitivity :- 0.860
- Specificity :- 0.942
- Precision:-0.895
- Recall:-0.860
- F1 score :- 0.8777

## ASSIGNING LEAD SCORE TO ORIGINAL CLEANED DATSET

- ORIGINAL CLEANED DATSET and Final Predicted Lead score combined to identify the HOT LEADS SUBSET
- Leads with Leads Score Greater than 30 are considered as HOT LEADS
- Further Analysis could be done on the subset to identify most promising conversions

## FINAL MODEL STATISTICS

### 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
Method:	IRLS	Log-Likelihood:	-1366.8
Date:	Mon, 26 Aug 2019	Deviance:	2733.6
Time:	19:43:36	Pearson chi2:	1.37E+04
No. Iterations:	8	Covariance Type:	nonrobust

	coef	std err	z	P> z	[0.025	0.975]
const	-0.9571	0.089	-10.708	0	-1.132	-0.782
Do Not Email	-1.3559	0.249	-5.441	0	-1.844	-0.867
Lead Source_Welingak Website	4.1527	0.742	5.598	0	2.699	5.607
Last Activity_SMS Sent	2.2261	0.111	20.14	0	2.009	2.443
What is your current occupation_Other	-1.055	0.113	-9.301	0	-1.277	-0.833
Tags_Busy	0.0951	0.235	0.404	0.686	-0.366	0.556
Tags_Closed by Horizon	6.8452	0.723	9.473	0	5.429	8.261
Tags_Lost to EINS	7.4833	0.844	8.87	0	5.83	9.137
Tags_Ringing	-3.9083	0.253	-15.469	0	-4.404	-3.413
Tags_Will revert after reading the email	4.651	0.198	23.53	0	4.264	5.038
Tags_switched off	-3.9303	0.525	-7.48	0	-4.96	-2.901
Lead Quality_Worst	-3.7399	0.614	-6.093	0	-4.943	-2.537
Last Notable Activity_Modified	-1.8047	0.122	-14.756	0	-2.044	-1.565

## **FINAL ANALYSIS and INSIGHTS**

Analysis could be done on the Hot leads subset to arrive at insights for increasing conversion rate

■ Have to prioritize on Hot leads(Lead Score >30) with following attributes:-

1. Tags\_Lost to EINS
2. Tags\_Closed by Horizzon
3. Tags\_Will revert after reading the email
4. Lead Source\_Welingak Website
5. Last Activity\_SMS Sent

---

---