



Summary Report

Lead Scoring
Case Study

**Chandrashekhar Purandare
Maganuru Basavaraju**

2nd March 2020

Problem Statement

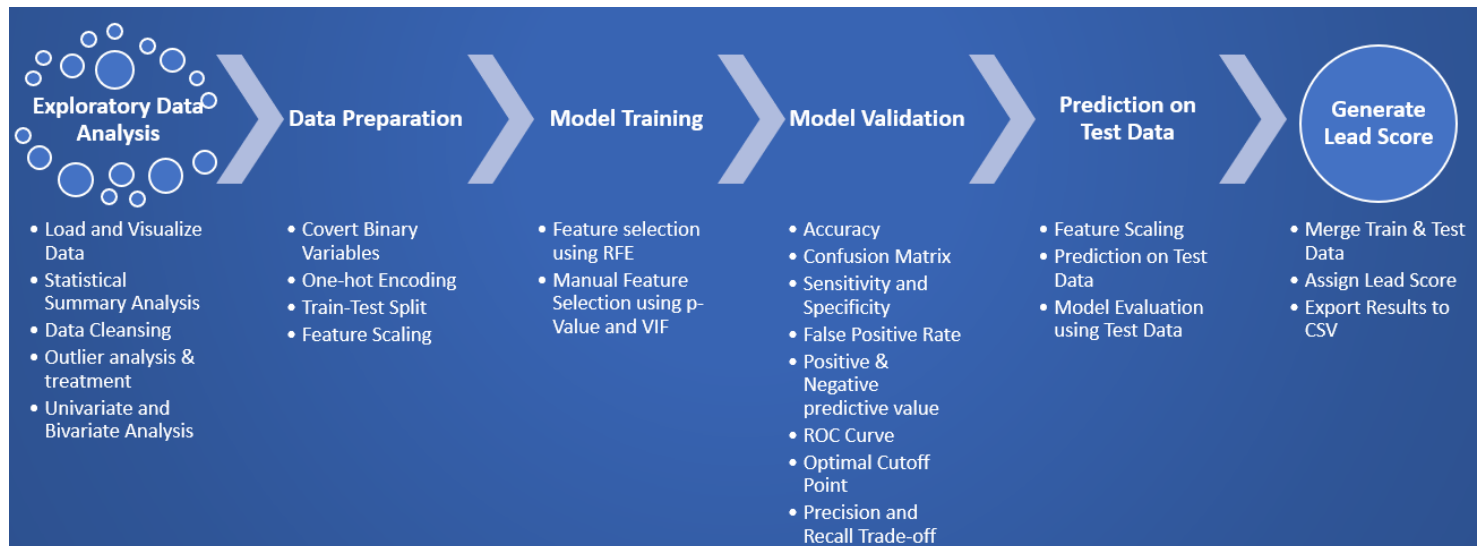
- + X Education - an education company which sells online courses to industry professionals markets its courses on several websites and search engines like Google.
- + X Education gets a lot of leads, its lead conversion rate is very poor at ~30%
- + CEO has mandate to increase the conversion rate to 80% from current 30%.

Business Objective

- + Current lead conversion rate of X education is low at around 30% which needs to be improved.
- + Categorize the Leads into hot and cold by assigning a lead score to each of the open leads to improve the Lead Conversion Ratio.
- + The CEO of X Education wants to increase lead conversion rate to around 80%
- + X Education also needs to consider the seasonality in business peak and lean periods.

Analysis Process

Process followed to perform analysis of Leads to classify hot and cold leads by assigning Lead Score.



Find below detailed **steps** and **key learnings**.

Exploratory Data Analysis

- Load and visualized the Leads data
- Handle duplicate Data
 - No duplicate data found
- Handle missing data
 - Considered the default “Select” level as null when user did not make any explicit selection from the list
 - Dropped columns with >70% missing values as there is no way to impute such a large missing value
 - Dropped index and score columns with no clarity on how to impute missing data
- Removed columns with single unique values and very low variance, as those doesn't impact prediction capability of model.

- Remove Outlier
 - Used quartile range and box plot to decide on capping values for outlier
- Univariate Analysis
 - Come up with inferences about which aspects need to be focused to improve the lead conversion rate such making website more engaging, target working professional with high rate of conversion and so on.

Data Preparation

- Convert binary columns
- One-hot encoding used to create dummy variables with multiple levels
- Split the train and test data using 70% & 30% guideline
- Feature scaling applied on Training data

Model Training

- Perform automated feature selection using RFE
- Fine grain the feature selected using p-Value and VIF manually
 - Validate final model features has less than 0.05 significance level indicating significant features
 - Ensure VIF value less than 3 for all features indicating no multicollinearity.

Model Validation

- Verify multiple matrix such as accuracy, sensitivity, specificity, false positive rate, positive & negative predictive value to ensure goodness of the final model.
- Analyzed ROC Curve
- Performed Precision & Recall trade-off analysis and check the F1 score to verify goodness of the model.
- Optimal cutoff where sensitivity and specificity are balanced is considered to decide probability threshold value.

Generate Lead Score

- Merged the train and test data
- Calculate Lead score based on predicated probability and assigned for each lead.
- Exported the Lead Score data in CSV, which can be used by X Education for further follow-up and improve conversion rate.

	Lead Number	Conversion_Prob	Converted	final_predicted	Lead_Score
0	660737	0.022784	0	0	2
1	660728	0.008597	0	0	1
2	660727	0.957345	1	1	96
3	660719	0.003435	0	0	0
4	660681	0.906935	1	1	91
5	660680	0.049356	0	0	5
6	660673	0.906935	1	1	91
7	660664	0.049356	0	0	5
8	660624	0.115536	0	0	12
9	660616	0.115536	0	0	12

Final Recommendations

Important variables for Lead Conversion



Top 3 Features to be focused

6	Tags_Lost to EINS	100.00
5	Tags_Closed by Horizzon	86.68
8	Tags_Will revert after reading the email	43.22