# X Education Lead Score Case study

Bhavini D. Gadhiwala Sankalp Rusia



- Improve the lead conversion rate.
- Study parameters provided and find top parameters that lead to conversion.
- Assign a lead score of probability value to each lead.
- This score will be used to nurture the 'Hot Leads' that has high probability of conversion.
- The target is to achieve 80% of lead conversion

#### Intent of the study

- 9240 rows and 37 columns
- 17 columns with missing values
- 5 numerical columns and 32 categorical columns
- 4 columns with 'Select' columns
  - These are from forms where the dropdown values were not selected by users

#### Dataset at Bird's Eye View

#### Columns with Highly Skewed data

- Country
- Through Recommendations
- Update me on supply Chain Content
- Get updates on DM content
- Newspaper
- Receive More Updates About Our Courses
- X Education Forums
- I agree to pay the amount through cheque
- Do Not call
- What matters most to you in choosing a course

#### Columns with Uneven value counts

- Many columns had uneven distribution of value counts where there were some values with counts in thousands, whereas many values were sparse in counts like under 10. The examples are below:
- Last Activity
- What is your current occupation
- Lead Source
- Lead Origin
- Tags
- Specialization

#### Columns with unstructured data

- The Tags column had a lot of values but it was quite unstructured
- The column City had values all over the place
- Another example is Last Notable Activity
- Lead Quality and Lead Source
- The columns What is your current occupation had values all other the place

### Exploring Data – Univariate Analysis

Data type Conversion to object – Lead Number, Converted

Translate 'Select' to numpy NULL value –
Specialization, How did you hear about X
Education, Lead Profile, City

Column removal due to significant data imbalance — Search, Magazine, Newspaper, Do not Call, etc. 13 columns impacted

Column removal of identifier columns – Lead Number, Prospect ID as they are not needed for analysis Created new value to replace missing values as Unknown – Specialization, What is your current occupation

Merging of multiple column values into one to reduce skewness in data – Lead Origin, Lead Quality, Last Activity, Country

Removal of Asymmetrique index and score columns because of lack of information and interpretation availability

Columns removed due to missing data in more than 70% rows – Lead Profile, How did you hear about X education

Conversion based on business logic. For example, City in Maharashtra and Others for City column

Derived column for unstructured data - Tag column was converted into 3 values based on business understanding as – Potential, Out, Unknown

Numerical columns had a few outliers. These were handled by removing the rows.

220 columns were impacted which is 2.4% of total rows. Some of the outliers in numerical columns were soft capped.

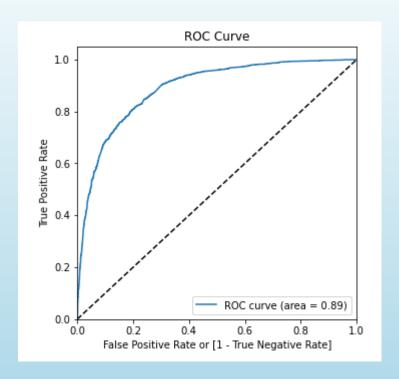
## **Data Cleaning Strategies**

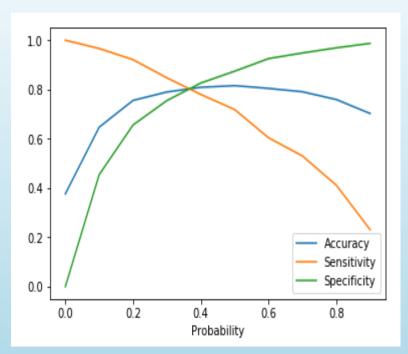
- Data Cleaning
- Creation of dummies
- Train-Test split 70-30
- Standard scaling
- Correlation matrix
- Coarse tuning using RFE -25 variables
- GLM fine tuning

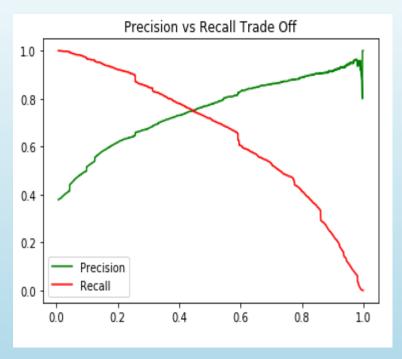
*Sr No	Column Dropped	VIF	P-value
1	Lead Quality_MayBe	10.8	0.0
2	Country_Others	5.6	0.0
3	Tags_Unknown	5.1	0.0
4	Specialization_International Business	3.11	0.4
5	Specialization_Retail Management	1.03	0.4

Data Modelling

<sup>\*</sup> The iteration number of GLM using statsmodels. Parameters removed based the high VIF or P-value or both







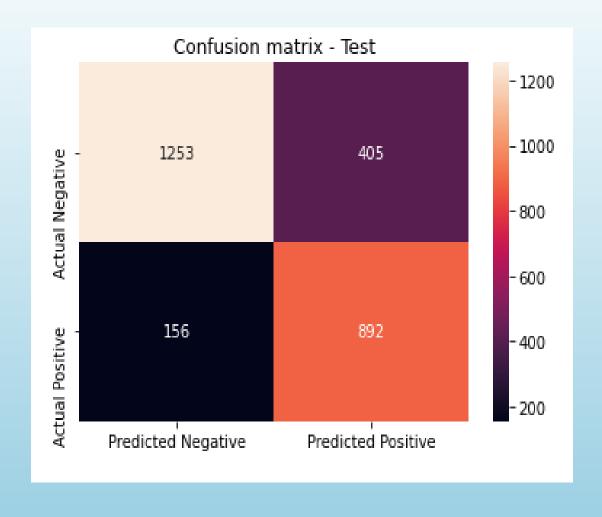
- ✓ Higher area under ROC curve
- ✓ Indicates greater accuracy

- ✓ Optimal cutoff probability of 0.3
- ✓ Clear observation of Precision vs Recall tradeoff

### **Model Strengths**

- ✓ Recall 85%
- ✓ Sensitivity 69%
- ✓ Specificity 76%
- ✓ Accuracy 79%

Business goal is to pursue 'hot leads' and reduce cold calls. High value of recall will fulfill the goal.



### Key Target Achievements of the Model

- Working professionals should be targeted most
- A user spending more time and visiting more often on website is a potential hot lead
- All the sources of leads should be considered seriously as the smaller values when clubbed together is creating impact. This is re-iterated in other words by the Lead Quality\_Low coefficient.
- Olark Chat is one very good lead source with potential candidates
- The leads that have not disclosed their current occupation were found to have negative impact on the lead score

#### Recommendations





















Thank You!