X Education Lead Score Case study

Bhavini D. Gadhiwala
Sankalp Rusia

Intent of the study

- 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

Dataset at Bird's Eye View

- 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

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

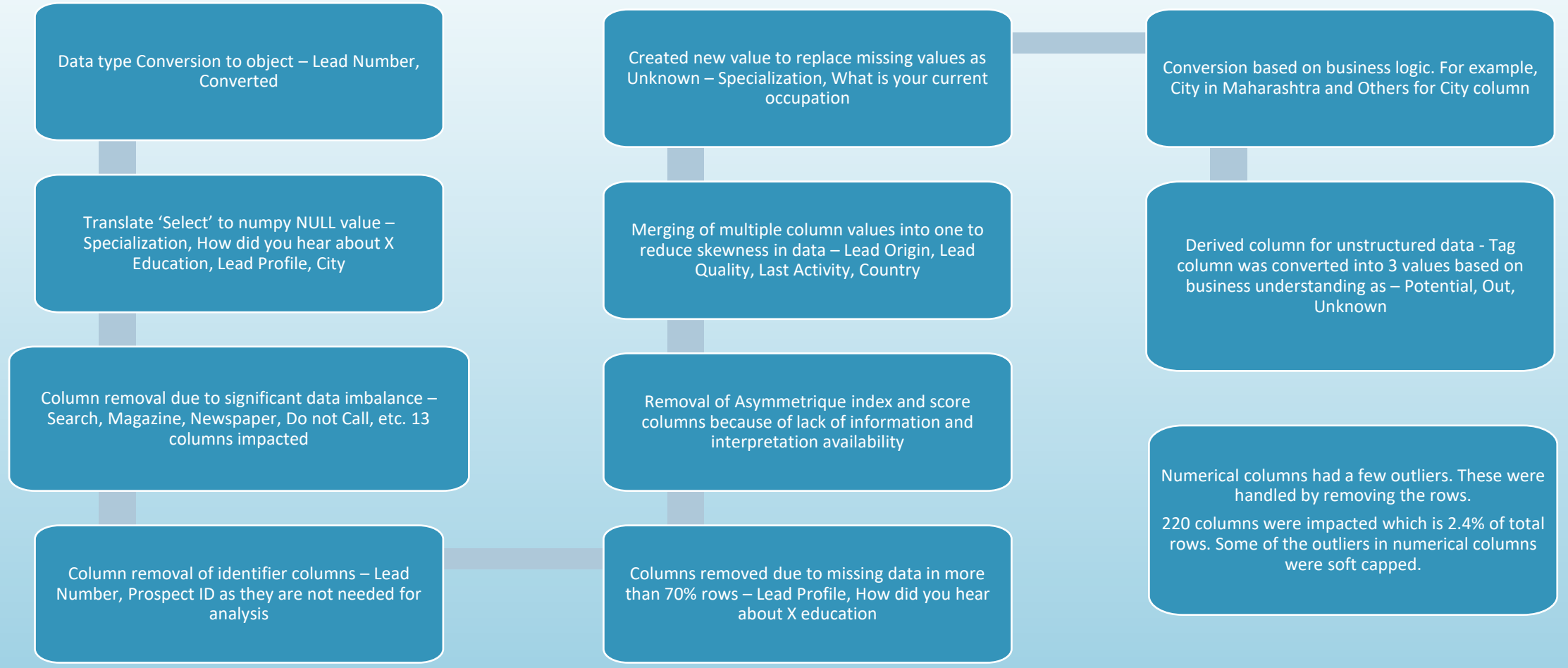
- Many columns had uneven distribution of value counts where there were some values with counts in thousands, whereas many values were sparse in counts like under 10. The examples are below:
  - Last Activity
  - What is your current occupation
  - Lead Source
  - Lead Origin
  - Tags
  - Specialization

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 Cleaning Strategies

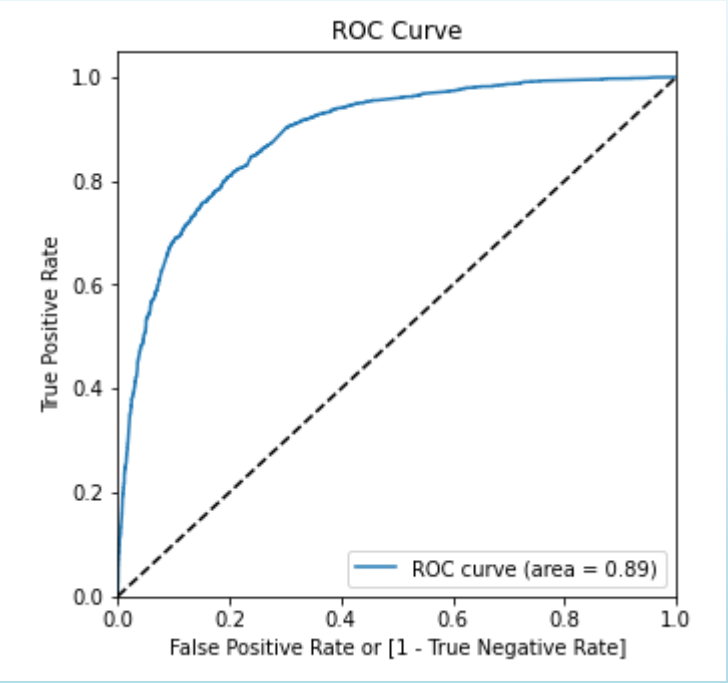


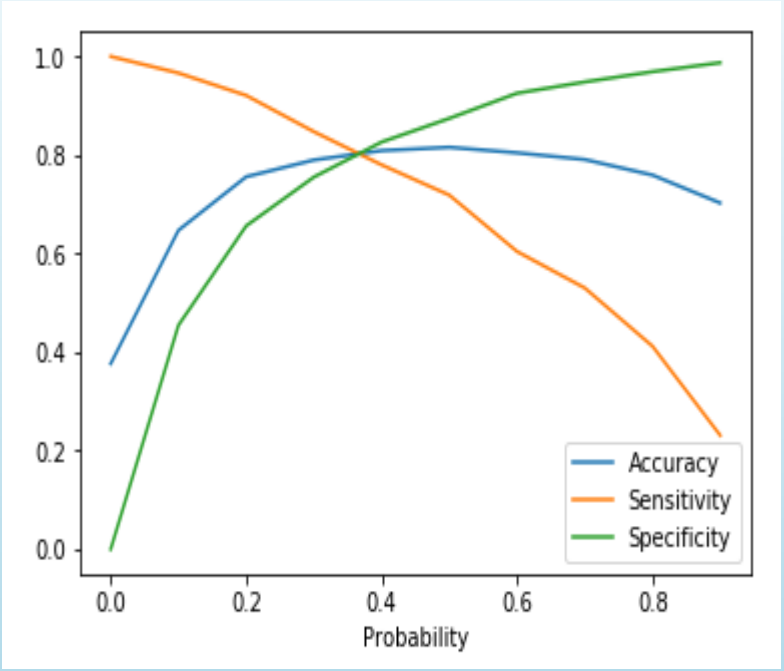
- Data Cleaning
- Creation of dummies
- Train-Test split 70-30
- Standard scaling
- Correlation matrix
- Coarse tuning using RFE -25 variables
- GLM – fine tuning

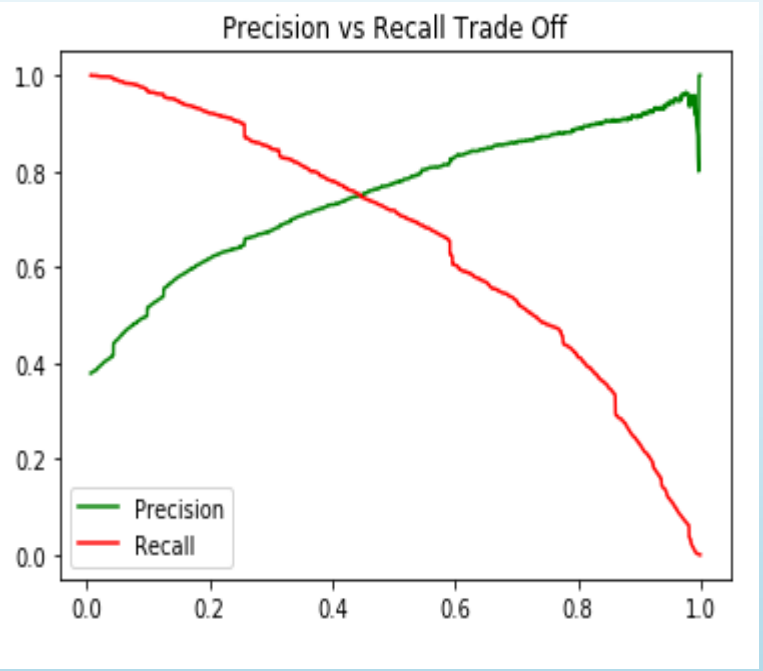
*Sr No	Column Dropped	VIF	P-value
1	Lead Quality_MayBe	10.8	0.0
2	Country_Others	5.6	0.0
3	Tags_Unknown	5.1	0.0
4	Specialization_International Business	3.11	0.4
5	Specialization_Retail Management	1.03	0.4

Data Modelling

* 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

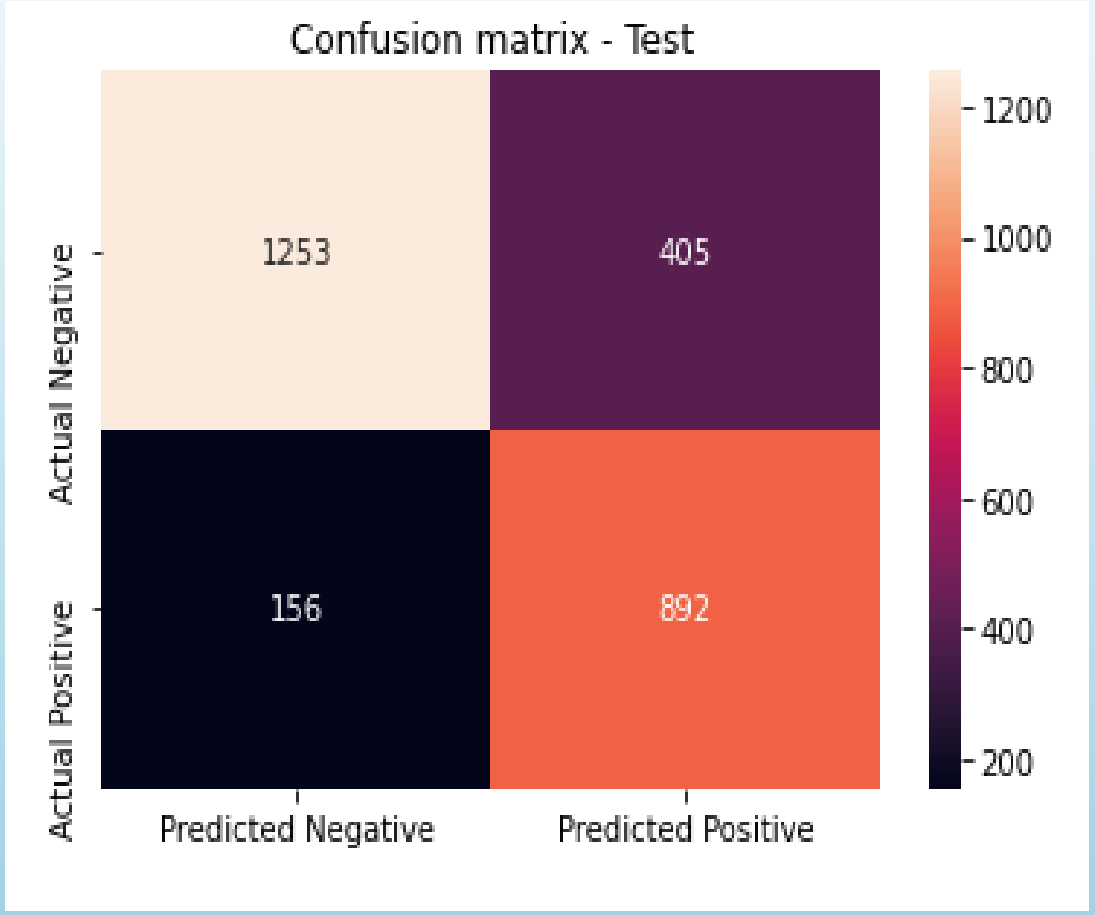
- ✓ Clear observation of Precision vs Recall tradeoff

Model Strengths

Key Target Achievements of the Model

- ✓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.



- 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!