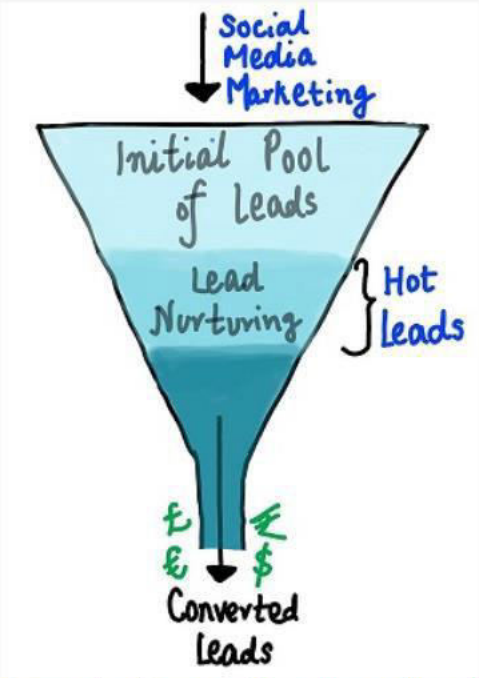
Lead Scoring Case Study SUMMARY
By Navneet Anand

Problem Statement:

X Education is an online education company that sells courses to industry professionals. The company has a high lead generation rate, but only 30% of leads convert into paying customers. The CEO wants to develop a model and strategy to increase the lead conversion rate.



Lead Conversion Process - Demonstrated as a funnel

Raw Data:

X Education collects data from multiple sources: Marketing, websites, Forms, Calls, Referrals etc.

Steps for solving the problem:

1. Loading and understanding Data with description.
2. EDA

1. **Data Preprocessing**
    i. Numerical Columns
        - We checked the missing values.
        - Outliers
    ii. Categorical Columns
        - We checked missing values,
        - What labels are present in each column and counts?
2. **Data Visualization**
    i. Univariate Analysis – Explore each variable in the dataset.
    ii. Bivariate Analysis – Explore each variable with Target variable.
3. **Feature Engineering**
    i.Dummy variables created
    ii.Imputed Missing value with others. To avoid information loss, we did not drop the missing rows.
    iii. Dropped unnecessary columns.
4. **Correlation check**
    Using the heat map, we check the correlation between independent columns to independent columns & independent columns to dependent columns.
    Also, we dropped the columns if correction is too high, to avoid the multicollinearity.

3. Train test Split

Split the data with 80-20 ratio. 80% training set and 20% test set.

4. Feature Scaling

Multiple columns are in different scales, so to normalize the range of features we used min-max scaling.

5. Feature Section

After the feature engineering, 68 independent variables were created, and all 68 features are not important so we must select important features for model building.

For selecting import features we used “rfe” and “vif” techniques.

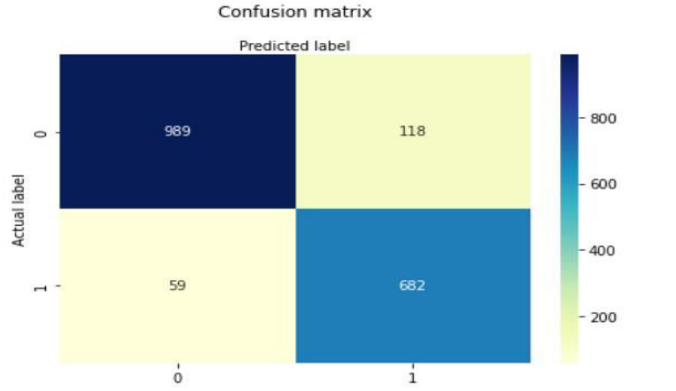
6. Model Training

Using Logistic regression, we train our model with the most important features. We used 21 features for building the model.

7. Model Evaluation

We evaluated our model using a confusion matrix and focused on recall. We obtained a cutoff value of 0.3, which resulted in a recall of 92%, a precision of 85%, an F1 score of 89%, and an accuracy of 90% on the test data. The high recall value indicates that our model is good at identifying positive leads, even if it is less precise.

Evaluation Matrix Result:



	precision	recall	f1-score	support
without Converted	0.94	0.89	0.92	1107
Converted	0.85	0.92	0.89	741
accuracy			0.90	1848
macro avg	0.90	0.91	0.90	1848
weighted avg	0.91	0.90	0.90	1848

Conclusion:

Our model has achieved a recall of 92% on the test dataset, which means that it correctly predicts 92% of the positive leads. This meets the CEO's expectation of a lead conversion rate of around 80%.

We should focus on the following customer segments for lead conversion:

- Working professionals
- Customers who spend more time on the X Education website
- Customers who visit the website regularly and have the highest page views per visit
- Customers who have responded to previous emails

We should not focus on the following customer segments:

- Customers who are interested in other courses
- Students

By focusing on the right customer segments, we can further improve our lead conversion rate and achieve the CEO's goal of increasing revenue from online courses.