LEAD SCORES ANALYSIS

X Education

Prepared by:

Mohan Chand Kommalapati

PROBLEM STATEMENT

- X Education attracts a large pool of leads every month through its marketing channels.
- However, only about 30 in 100 leads actually purchase a course.
- Sales executives spend valuable time pursuing leads that rarely convert.

Goals of the Case Study: There are quite a few goals for this case study. They are as follows:

- 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., 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 that your model should be able to adjust to if the company's requirements change in the future, so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it out based on the logistic regression model you got in the first step. Also, make sure you include this in your final PowerPoint presentation, where you'll make recommendations.

DATA SET

- Dataset size: ~9,000 historical lead records.
- Key attributes include:
  - Website behavior: visits, time spent, forms filled.
  - Campaign info: lead source, referrals, tags.
  - Interactions: calls, emails, last activity.
- Target variable: Converted (1 = purchase, 0 = no purchase).
- Preprocessing steps:
  - Removed missing/irrelevant entries.
  - Converted categorical fields into numerical form.
  - Standardized numeric features for modeling.

ANALYTICAL APPROACH

- Step-1: Reading the Data
- Step-2: Understanding Data
- Step-3: Data Preparation
- Step-4 : Looking at correlations
- Step-5: Split data into input and target variables (Train – Test split)
- Step-6 : Feature Scaling
- Step -7: Model Building
- Step-8: Metrics beyond Accuracy
- Step-9 : Plotting the ROC curve
- Step-10: Finding the optimal cutoff point
- Step - 11: Making predictions on the test data (Model Evaluation)

ANALYTICAL APPROACH

Step-1: Reading the Data:

- Imported the dataset (Leads.csv) into Python.
- Verified successful loading and checked dataset size (~9,000 rows, 30+ features).
- Ensured the data types of each column were properly recognized (categorical, numeric, object).
- First look at the data helped confirm its structure and readiness for analysis.

```
# Load and Read the dataset
df = pd.read_csv("Leads.csv")
df.head()
```

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	Get updates on DM Content	Lead Profile	City	Asymmetrique Activity Index	Asymm Prof
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	Select	Select	02.Medium	02
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	...	No	Select	Select	02.Medium	02
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	...	No	Potential Lead	Mumbai	02.Medium	
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	...	No	Select	Mumbai	02.Medium	
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	...	No	Select	Mumbai	02.Medium	

5 rows × 37 columns

ANALYTICAL APPROACH

Step-2: Understanding Data:

- An initial exploration was performed to understand the nature of the data.
- The target variable Converted showed about 30% positive conversions, highlighting class imbalance.
- The features included a mix of numerical (e.g., Total Visits), categorical (e.g., Lead Source), and derived tags (e.g., Last Activity).
- Early inspection revealed issues like missing values and irrelevant levels such as "Select," which required cleaning.
- This step helped us identify challenges and opportunities in the dataset.

ANALYTICAL APPROACH

Step-3: Data Preparation:

- Cleaned the dataset by treating missing values appropriately. Columns with excessive missing values were dropped, while others were imputed.
- Removed irrelevant or redundant variables that provided little value for prediction.
- Checked for outliers and treated the outliers.
- Converted categorical variables into **dummy variables** to make them usable by the model.
- After preparation, the d
- ataset became structured, consistent, and modeling-ready.

```
df.head()
```

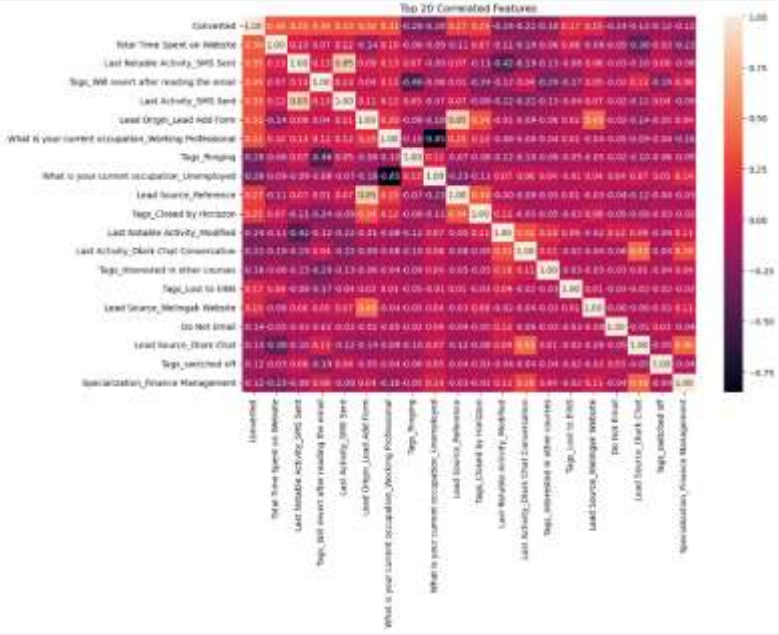
	Lead Number	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Search	Magazine	Newspaper Article	...	Last Notable Activity_Form Submitted on Website	Last Notable Activity_Had a Phone Conversation	Last Notable Activity_Modified	Last I Activit Conv
0	660737.0	0	0	0.0	0.0	0.0	0.0	0	0	0	...	0	0	1	
1	660728.0	0	0	0.0	5.0	674.0	2.5	0	0	0	...	0	0	0	
2	660727.0	0	0	1.0	2.0	1532.0	2.0	0	0	0	...	0	0	0	
3	660719.0	0	0	0.0	1.0	305.0	1.0	0	0	0	...	0	0	1	
4	660681.0	0	0	1.0	2.0	1428.0	1.0	0	0	0	...	0	0	1	

5 rows × 165 columns

ANALYTICAL APPROACH

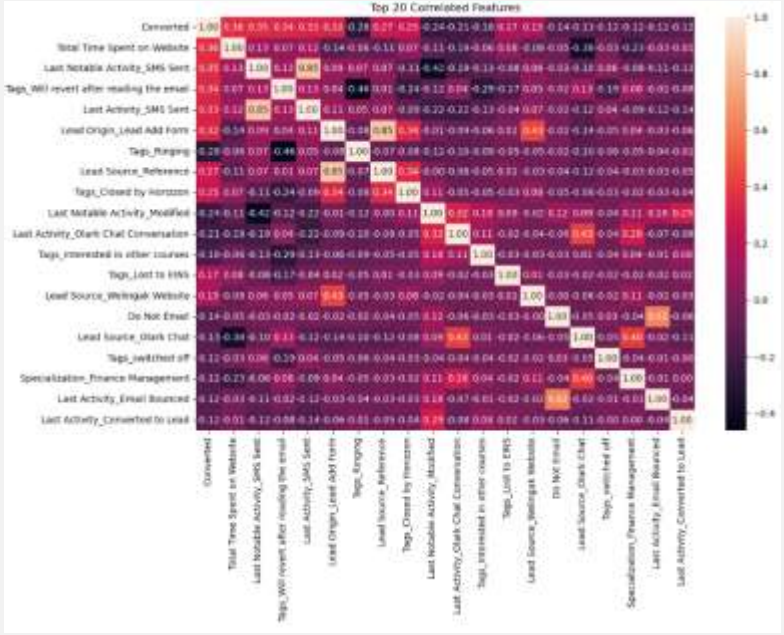
Step-4 : Looking at correlations:

- A heatmap was used to visualize how features relate to each other and to the target.
- This helped in spotting multicollinearity—where two or more predictors are highly correlated.
- High multicollinearity can destabilize logistic regression coefficients. Problematic features were flagged for removal in the model refinement stage.



Before removing highly correlated variables

After removing highly correlated variables



ANALYTICAL APPROACH

Step-5: Split data into input and target variables (Train – Test split):

- To evaluate the model's real-world performance, the dataset was split into training and test sets.
- Typically, 80% of the data was used for training the model, and 20% was reserved for testing.
- Stratified sampling ensured the same proportion of converted vs non-converted leads in both sets.
- This step ensured that the model would be trained on sufficient data while still being validated on unseen records.

ANALYTICAL APPROACH

Step-6 : Feature Scaling:

- Numerical variables like Total Visits, Page Views Per Visit and Total Time Spent on Website were on very different scales.
- Scaling transformed these variables into a standardized range (mean = 0, variance = 1).
- This is crucial in logistic regression so that variables are comparable and coefficients remain stable.
- Without scaling, features with large numeric ranges could dominate the model unfairly.

ANALYTICAL APPROACH

Step -7: Model Building:

- Logistic regression was selected because it is interpretable and outputs conversion probabilities.
- A stepwise process was used to retain only statistically significant predictors.
- p-values helped identify irrelevant variables, while Variance Inflation Factor (VIF) was used to remove correlated ones.
- After multiple iterations, a final set of strong predictors was obtained.

```
Index(['Do Not Email', 'Lead Origin_Lead Add Form',
       'Lead Source_Welingak Website', 'Tags_Busy', 'Tags_Closed by Horizzon',
       'Tags_Lost to EINS', 'Tags_Ringing',
       'Tags_Will revert after reading the email', 'Tags_switched off',
       'Last Notable Activity_Had a Phone Conversation',
       'Last Notable Activity_SMS Sent', 'Last Notable Activity_Unsubscribed'],
      dtype='object')
```

- The model was then fit to the training data, producing probability scores for each lead.

ANALYTICAL APPROACH

Step-8: Metrics beyond Accuracy:

- Accuracy alone can be misleading, especially with imbalanced data.
- Additional metrics were calculated:
  - **Precision** – proportion of predicted conversions that were correct.
  - **Recall (Sensitivity)** – proportion of actual conversions identified.
  - **Specificity** – ability to correctly identify non-conversions.
  - **F1-score** – harmonic mean of precision and recall.
- These metrics provided a holistic understanding of the model's performance.

```
# Let's find out the sesitivity of the model
TP/float(TP+FN)
```

```
0.624113475177305
```

```
# Let's findout the specificity
TN/float(TN+FP)
```

```
0.968503937007874
```

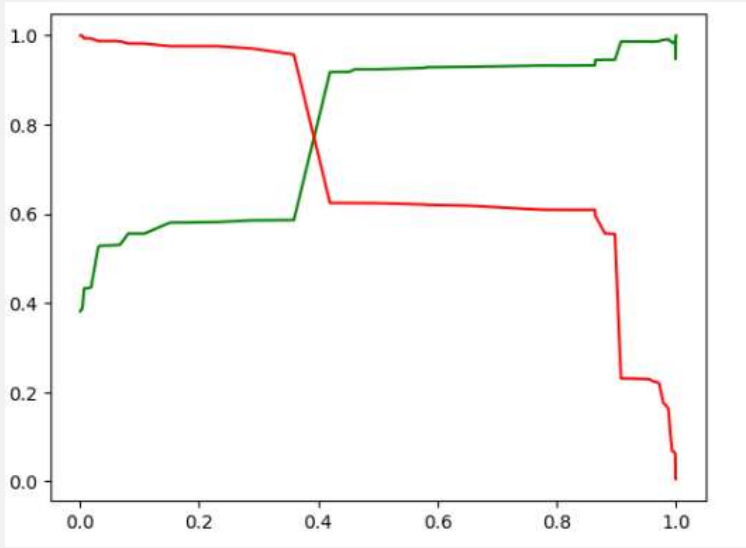
```
Precision = TP/float(FP+TP)
Precision
```

```
0.9181011997913406
```

```
Recall = TP/float(TP+FN)
Recall
```

```
0.624113475177305
```

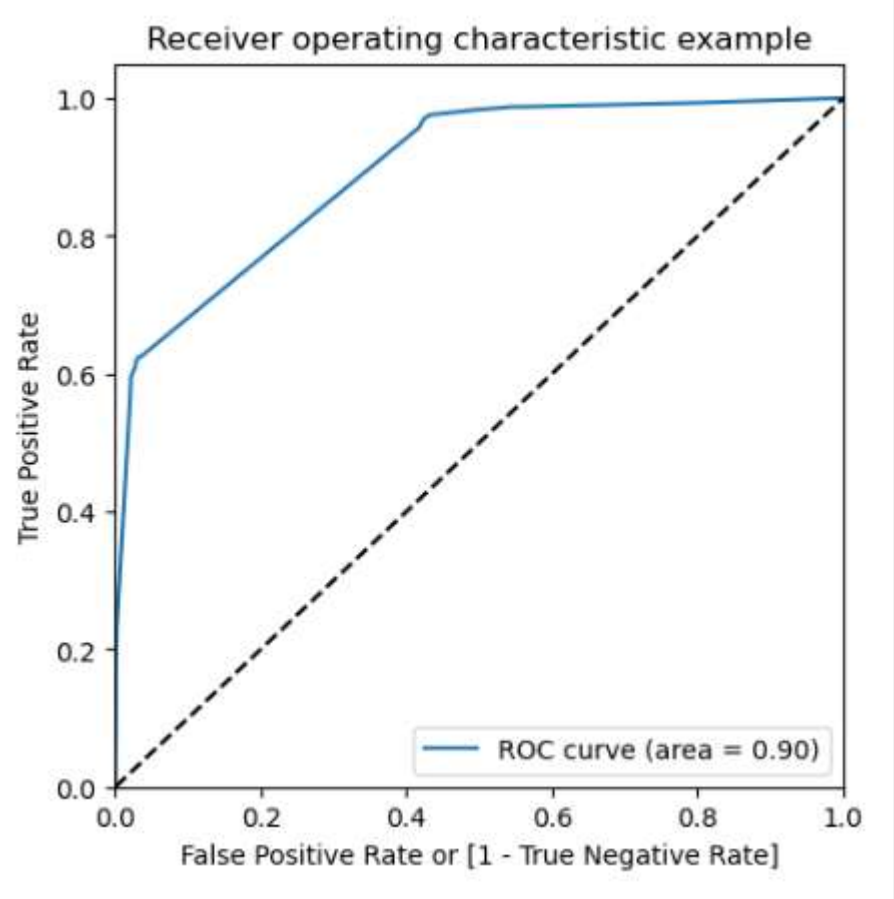
Precision and Recall Tradeoff:



ANALYTICAL APPROACH

Step-9 : Plotting the ROC curve:

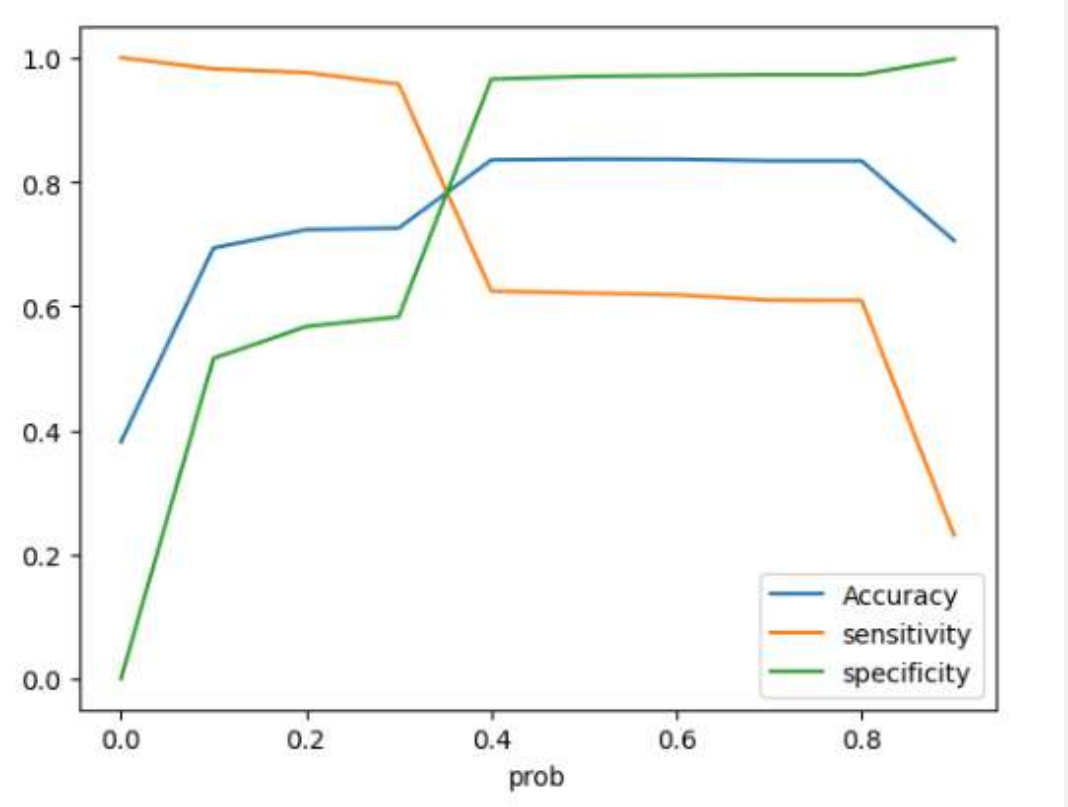
- The Receiver Operating Characteristic (ROC) curve was plotted to assess model performance across thresholds.
- The curve shows the trade-off between True Positive Rate (Recall) and False Positive Rate.
- The Area Under Curve (AUC) was computed, giving a single measure of model quality.
- A higher AUC (closer to 1) indicated strong discriminative ability of the model.
- This provided visual evidence that the model could differentiate between converters and non-converters.



ANALYTICAL APPROACH

Step-10: Finding the optimal cutoff point:

- The default probability cutoff of 0.5 often fails to balance business needs.
- A range of cutoff values (0.0–1.0) was tested.
- For each cutoff, sensitivity, specificity, and accuracy were calculated.
- Here, The cutoff around 0.35–0.40 provided the best balance, ensuring high recall without losing too much specificity.



ANALYTICAL APPROACH

Step - 11: Making predictions on the test data (Model Evaluation):

- The final model and chosen cutoff were applied to the test dataset.
- Performance metrics on the test data closely matched those on the training data, showing good generalization.
- Precision, recall, and AUC confirmed that the model successfully identified high-potential leads.
- The model could now assign lead scores (0–100) to new prospects, guiding sales prioritization.
- This final step validated that the approach was effective and business-ready.

	Prospect_ID	Converted	Converted_prob	final_predicted	Lead score	Interpretation
0	4608	1.0	0.987269	1	99	Very Hot Leads
1	7935	0.0	0.107967	0	11	Cold Leads
2	4043	0.0	0.107967	0	11	Cold Leads
3	7821	0.0	0.004608	0	0	Cold Leads
4	856	0.0	0.358775	0	36	Cold Leads
...	...	...	...	...	...	...
1843	7387	1.0	0.358775	0	36	Cold Leads
1844	3063	1.0	0.971569	1	97	Very Hot Leads
1845	603	0.0	0.004608	0	0	Cold Leads
1846	4210	1.0	0.358775	0	36	Cold Leads
1847	7352	0.0	0.004608	0	0	Cold Leads

1848 rows × 6 columns

MODEL INSIGHTS

- Logistic regression chosen for its simplicity and interpretability.
- The model provides probability scores that can be scaled to a Lead Score (0–100).
- Best cutoff probability ≈ 0.35–0.40, ensuring both good recall and acceptable specificity.
- Most Influential Factors: Top predictors of conversion likelihood:
  - Tags: Closed by Horizon
  - Tags: Lost to EINS
  - Last Activity: Phone Conversation
- Interpretation:
  - Tags linked with closure or loss strongly predict non-conversion.
  - Direct interaction (phone calls) increases conversion chances.

ACTION PLAN

- Introduce Lead Scoring System for day-to-day sales operations.
- Dynamically adjust threshold depending on business stage.
- Monitor conversion rates and compare against historical baseline.
- Continue refining the model with new data.
- Test advanced methods (tree-based models) for possible accuracy gains.

TAKEAWAYS

- Lead scoring system provides a structured way to prioritize leads.
- The approach helps balance between aggressive growth and efficient resource use.
- Top drivers reveal clear insights into customer behavior.
- End result: higher productivity, better conversions, smarter sales strategy.

THANK YOU