LEAD SCORE CASE STUDY

PANKAJ
and
UTKARSH

PROBLEM STATEMENT

- X Education is an esteemed education company specializing in offering high-quality online courses tailored to meet the needs of industry professionals. To reach out to potential learners, the company employs marketing strategies across various platforms, including popular search engines like Google.
- Once visitors land on the company's website, they have the opportunity to explore the diverse range of courses available or may choose to complete a form to express their interest in a specific course. Additionally, engaging video content is also available to captivate the audience. By providing their contact details, such as email addresses or phone numbers, these individuals are categorized as leads. Furthermore, X Education's lead generation is not limited to website interactions alone, as they also receive leads through referrals from satisfied past customers.
- Once these valuable leads are acquired, the efficient sales team takes over, utilizing various communication methods like calls and emails to establish meaningful connections. The lead conversion rate at X Education stands at an impressive 30%, highlighting the company's effectiveness in turning potential learners into satisfied students.

PRIMARY OBJECTIVE:

- The company's primary objective is to identify the most promising leads, referred to as "Hot Leads," and improve the lead conversion process. To achieve this, the company requires a lead scoring model that assigns a lead score to each potential customer based on their likelihood of conversion. Leads with higher lead scores are expected to have a greater chance of converting into paying customers, while those with lower scores are anticipated to have a lower probability of conversion.
- The CEO has provided a rough estimation of the target lead conversion rate, which stands at 80%. This implies that the company aims to achieve an 80% success rate in converting leads into actual customers. By implementing an efficient lead scoring system, X Education aims to prioritize and focus its resources on the most promising leads, ultimately increasing the overall conversion rate and maximizing the return on investment for its marketing and sales efforts.

THE OVERALL APPROACH FOR THE PROJECT INVOLVES THE FOLLOWING STEPS:

- Preprocessing the data by handling missing values and performing data cleaning.
- Conducting exploratory data analysis, encompassing univariate, bivariate, and multivariate analyses.
- Scaling features and creating dummy variables as part of the data preparation.
- Constructing a logistic regression model to predict outcomes.
- Evaluating the model using metrics such as specificity, sensitivity, precision, and recall.
- Drawing conclusions and providing recommendations based on the model's findings

Approach Followed:

Data Cleaning and Prepration

- Acquiring data from the source.
- Preparing and cleansing the data to ensure its suitability for analysis.
- Identifying and eliminating duplicate entries.
- Addressing outliers in the dataset.
- Conducting exploratory data analysis to gain insights and understanding from the data.

Data spliting and Feature scaling

- The process of dividing the data into separate test and train datasets.
- Performing feature scaling on numerical variables to standardize their values.

Model Building

- Employing RIF, VIF, and p-value techniques for feature selection.
- Utilizing Logistic Regression to identify the optimal model.
- Computing diverse evaluation metrics to assess model performance.

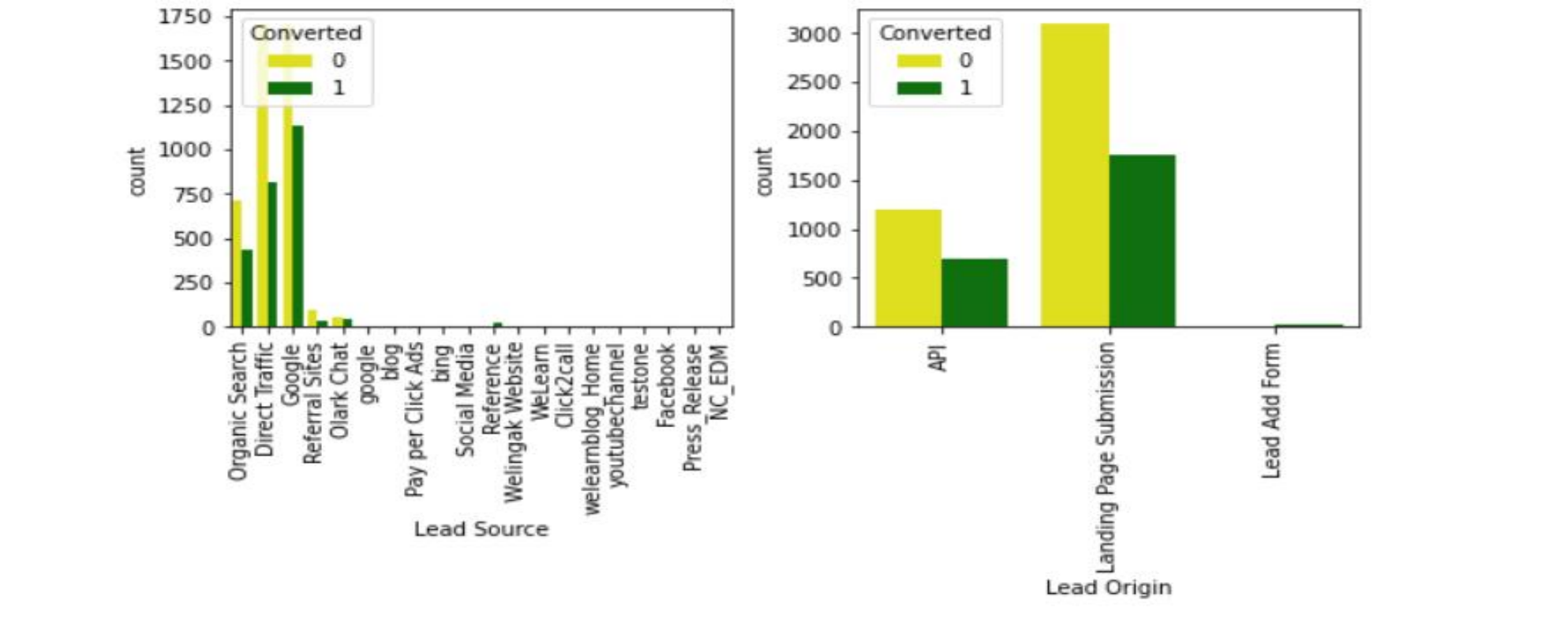
Result

- Calculate the lead score and verify if the target final prediction achieves a conversion rate higher than 80%.
- Assess the performance of the final prediction using the test set.

CONVERTING DATA

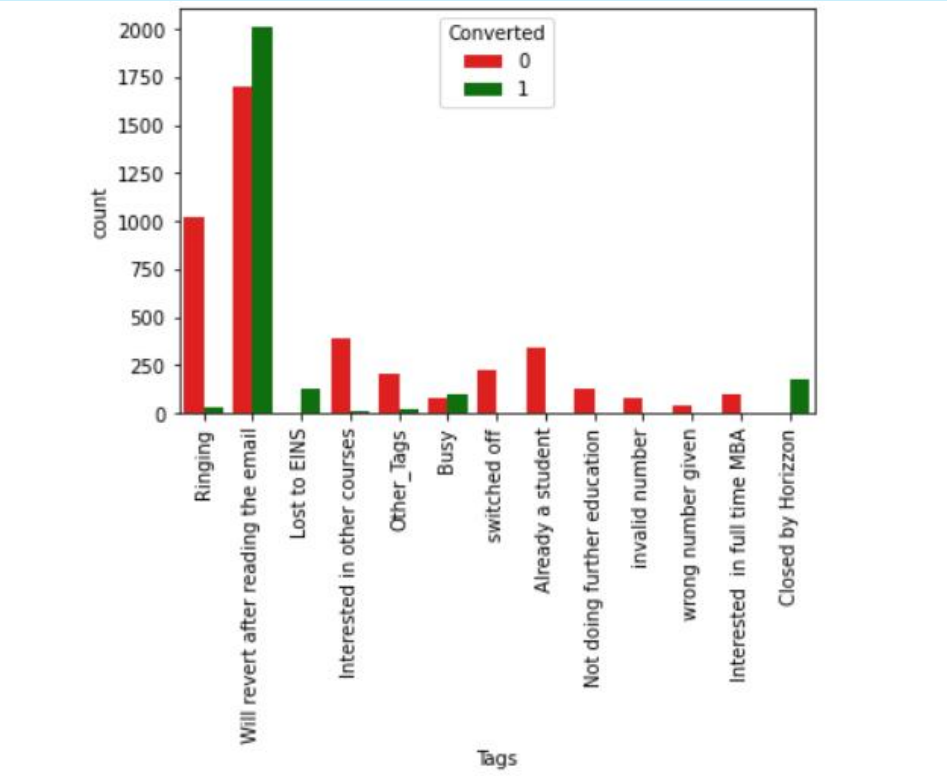
- CONVERTING THE VARIABLE WITH VALUES YES/NO to 1/0s
- CONVERTING THE ‘SELECT’ VALUES WITH NaNs
- DROPIING THE COLUMNS HAVING >70% OF NULL VALUES
- DROPPING UNNECESSARY COLUMNS
- DROPPING THE ROWS AS THE NULL VALUES WERE <2

EDA:

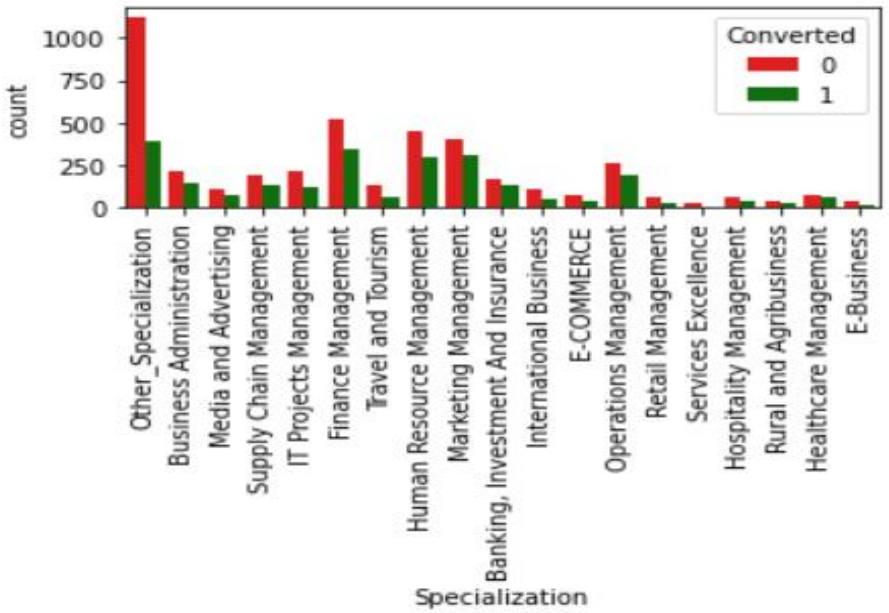


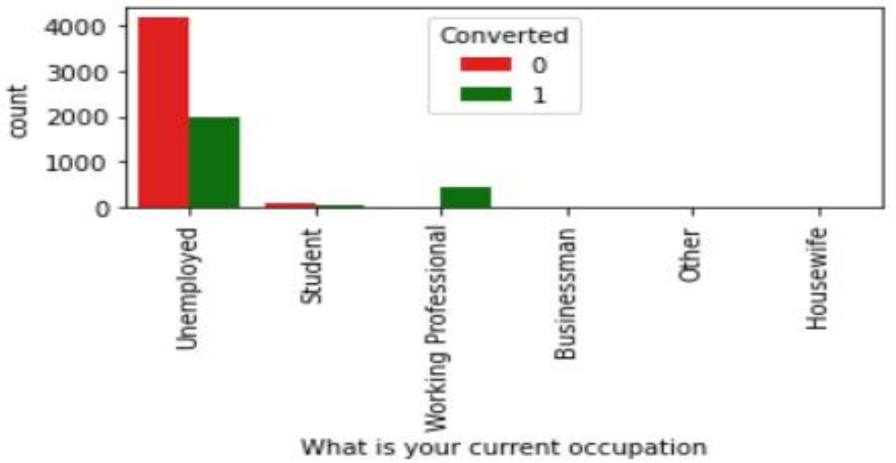
1. The highest number of leads originates from Google and Direct Traffic sources.
2. The leads coming from Reference and Welingak Website exhibit the highest conversion rate.
3. Although API and Landing Page Submission generate a substantial number of leads, their conversion rate is relatively low, standing at approximately 30%.
4. Despite the Lead Add Form contributing a small number of leads, it boasts an impressively high conversion rate.

The actions "Will revert after reading the email" and "Closed by Horizzon" demonstrate a significantly high conversion rate.



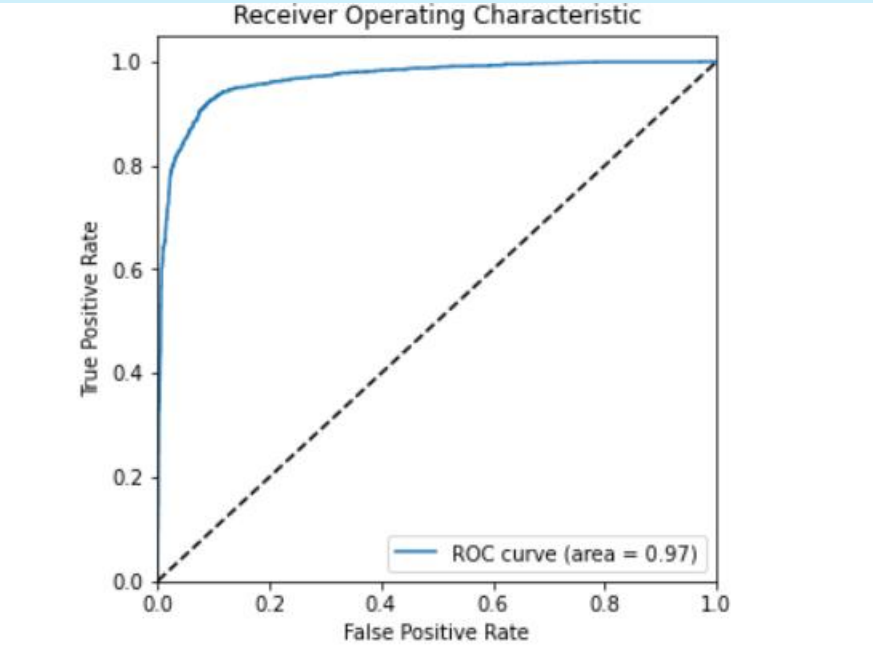
Based on the above plot, it is challenging to draw any specific conclusions regarding the impact of Specialization on the conversion rate.
However, it is evident from the plot that working professionals exhibit a notably high conversion rate.
Furthermore, the number of leads categorized as "Unemployed" surpasses that of any other category.

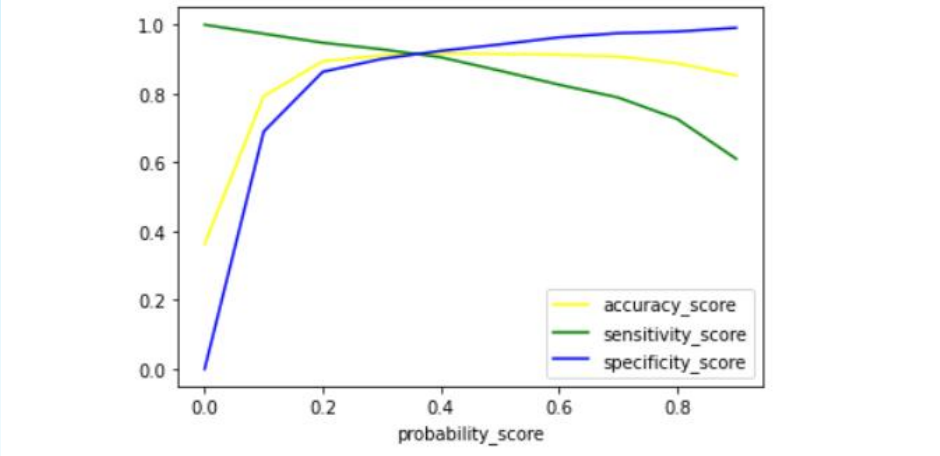




MODEL BUILDING

The dataset was divided into training and test sets with a split ratio of 70:30.
Recursive Feature Elimination (RFE) was utilized to select the most influential features, resulting in the identification of the top 15 variables.
A model was constructed by eliminating variables with a p-value greater than 0.05 and a VIF (Variance Inflation Factor) exceeding 5.
Subsequently, predictions were made on the test dataset.
The overall accuracy achieved by the model was recorded at 91.0%.





MODEL EVALUATION

ACCURACY	**91**
PRECISION	93
SENSITIVITY	90
SPECIFICITY	83

PREDICTED ACTUAL	**NOT CONVERTED**	**CONVERTED**
NOT CONVERTED	2702	319
CONVERTED	116	1607

Based on the graphical representation and considering the other scores, it is evident that the most favorable threshold is 0.2.

	probability_score	accuracy_score	sensitivity_score	specificity_score	precision_score
0.0	0.0	0.363196	1.000000	0.000000	0.363196
0.1	0.1	0.792791	0.973883	0.689507	0.641437
0.2	0.2	0.893971	0.947766	0.863290	0.798143
0.3	0.3	0.910835	0.928613	0.900695	0.842105
0.4	0.4	0.917369	0.905398	0.924197	0.871996
0.5	0.5	0.915051	0.866512	0.942734	0.896158
0.6	0.6	0.913364	0.825885	0.963257	0.927640
0.7	0.7	0.907673	0.788741	0.975505	0.948360
0.8	0.8	0.888069	0.726640	0.980139	0.954268
0.9	0.9	0.852867	0.610563	0.991063	0.974977

MODEL PREDICTION

ACCURACY	38
PRECISION	38
SENSITIVITY	1
SPECIFICITY	0.7

PREDICTED ACTUAL	NOT CONVERTED	CONVERTED
NOT CONVERTED	9	1264
CONVERTED	0	761

```
const                                    -1.366904
Do Not Email                             -1.374797
Total Time Spent on Website               1.083616
Last Activity_SMS Sent                    1.485138
Tags_Busy                                 3.899089
Tags_Closed by Horizzon                   8.753978
Tags_Lost to EINS                         8.745353
Tags_Ringing                             -1.276227
Tags_Will revert after reading the email  3.717455
Tags_switched off                        -2.665014
Lead Quality_Not Sure                    -3.273366
Lead Quality_Worst                       -3.539447
Last Notable Activity_Modified           -1.294615
dtype: float64
```

CONCLUSION

The logistic regression model serves to predict the probability of customer conversion. We have evaluated both sensitivity-specificity and precision-recall metrics, but we selected the optimal cutoff based on sensitivity-specificity for the final predictions.

The Lead Score calculated indicates that the conversion rate of the final predicted model is approximately 91% on the test data, slightly lower than the 95% achieved on the train data. Nevertheless, in business terms, this model has the adaptability to meet the company's future requirements.

The top variables contributing to lead conversion in the model are identified as follows:

- Tags_Lost to EINS
- Tags_Closed by Horizzon
- Lead Quality_Worst

Overall, the model demonstrates promising performance and seems to be effective in predicting lead conversions successfully.