LEAD SCORING CASE STUDY

By Kranthi Potharla

Introduction

- In this study, we aim to develop a logistic regression model for X Education that assigns each lead a score from 0 to 100, reflecting their potential for conversion. This scoring mechanism will enable the company to prioritize leads more effectively, aiming for an ambitious conversion rate target of 80%. The study will also tackle various challenges faced by X Education, offering strategic recommendations to leverage the lead scoring model for achieving their business objectives. Furthermore, the model will be designed with flexibility in mind, ensuring it can adapt to future shifts in the company's strategic needs.

BUSINESS UNDERSTANDING

- - X Education offers online courses targeted at industry professionals.
- - Promotes courses through online channels, including Google.
- - Prospective customers explore courses on the X Education website.
- - Visitors submitting contact information via a website form are categorized as leads.
- - The sales team attempts to convert these leads into paying customers via phone or email.
- - Despite efforts, a significant number of leads do not convert to customers.
- - The typical lead conversion rate for X Education is around 30%.

Data Understanding

- The dataset is comprised of 'Leads.csv' and 'Leads Data Dictionary.xlsx'.

- 'Leads.csv' holds approximately 9000 entries with the 'Converted' column indicating lead conversion status: 1 for converted and 0 for not converted.

- 'Leads Data Dictionary.xlsx' serves as a guide, detailing the variables found in the 'Leads.csv' file.

STEPS OF ANALYSIS

- DATA IMPORTING & CLEANING

- EXPLORATORY DATA ANALYSIS

- DATA PREPARATION

- MODEL BUILDING & EVALUATION

- MAKING PREDICTIONS ON TEST DATASET

DATA CLEANING

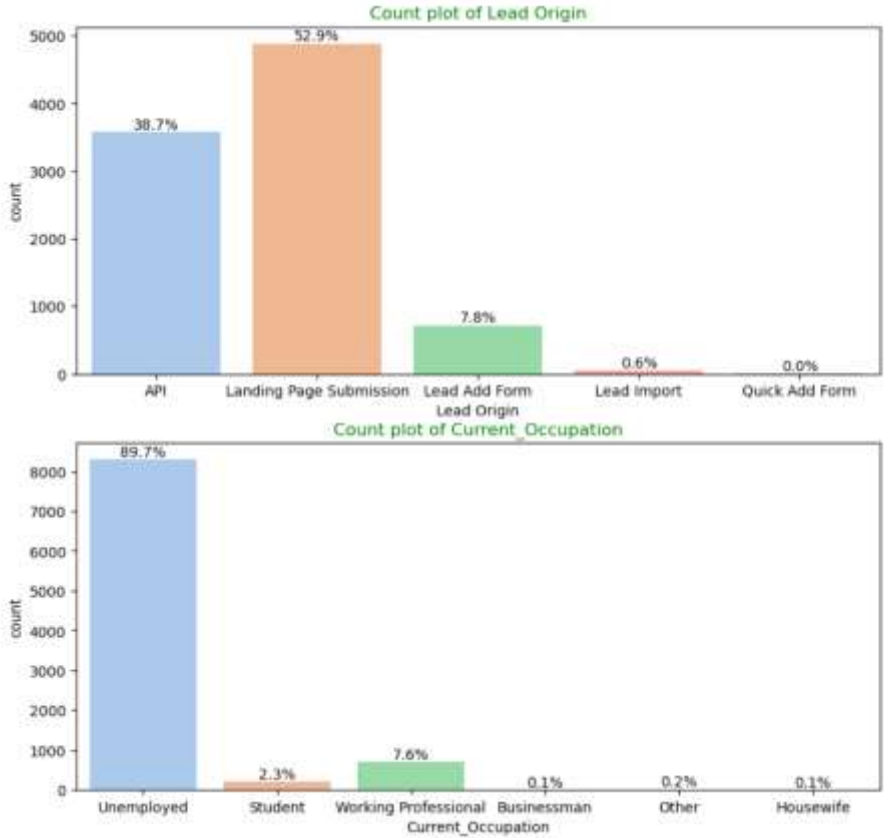
- - "Select" level signifies unselected options in categorical variables, indicating null values.
- - Columns with more than 40% missing values were eliminated.
- - Missing values in categorical variables were filled considering their frequency and specific conditions.
- - Uninformative columns for the study, like 'City', 'Tags', 'Country', and 'What matters most to you in choosing a course', were removed.
- - Certain categorical variables underwent imputation.
- - Irrelevant columns for modeling, such as 'Prospect ID', 'Lead Number', and 'Last Notable Activity', were discarded.
- - Mode imputation was applied to numerical data after assessing their distribution.
- - Categories with skewness were identified and removed to prevent logistic regression model bias.
- - Extreme values in 'TotalVisits', 'Total Time Spent on Website', and 'Page Views Per Visit' were capped.
- - Rare categories were consolidated into an "Others" group.
- - Data standardization was conducted to ensure uniform casing, correcting inconsistencies like "Google" versus "google".

EXPLORATORY DATA ANALYSIS

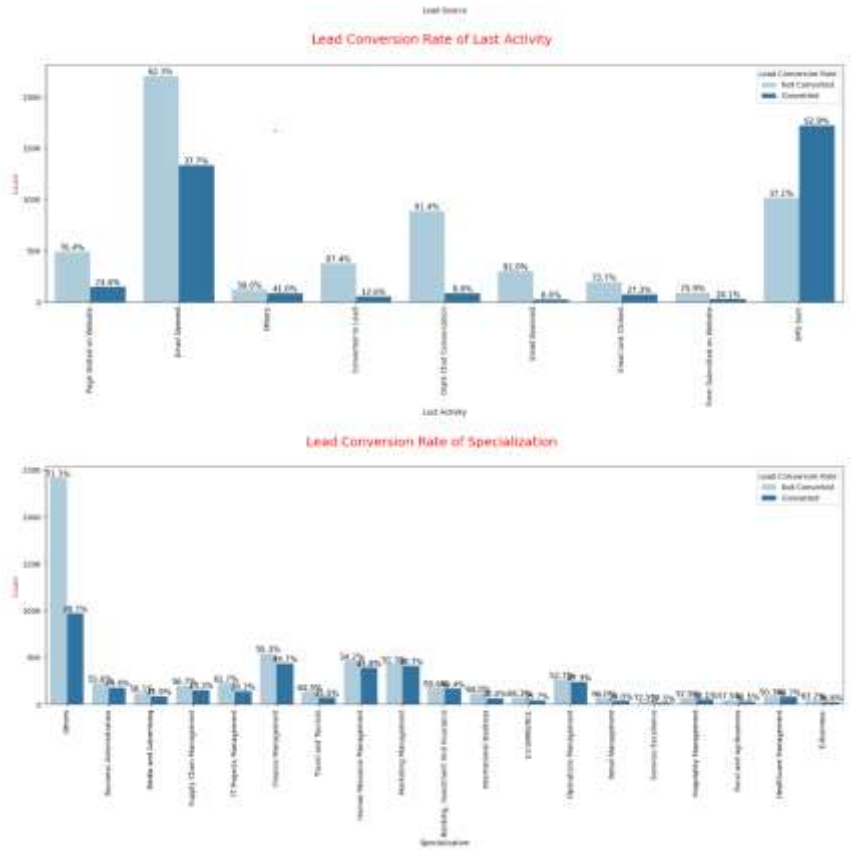
- The exploratory data analysis (EDA) phase for X Education involved rigorous data cleaning and preparation steps to ensure the dataset's readiness for modeling. This process included identifying and handling null values, eliminating columns with high percentages of missing data, and imputing missing values in both categorical and numerical variables based on specific criteria. Uninformative columns that didn't align with the study's objectives were dropped. Additionally, efforts were made to standardize data, group low-frequency values, and treat outliers, setting a solid foundation for building accurate and reliable logistic regression models for lead scoring.

UNIVARIATE ANALYSIS

- For 'Lead Origin', over half (52.9%) of the leads came from 'Landing Page Submission', with 'API' being the second most common source at 38.7%.
- Regarding 'Current Occupation', the data shows that a vast majority (89.7%) of the leads are unemployed.

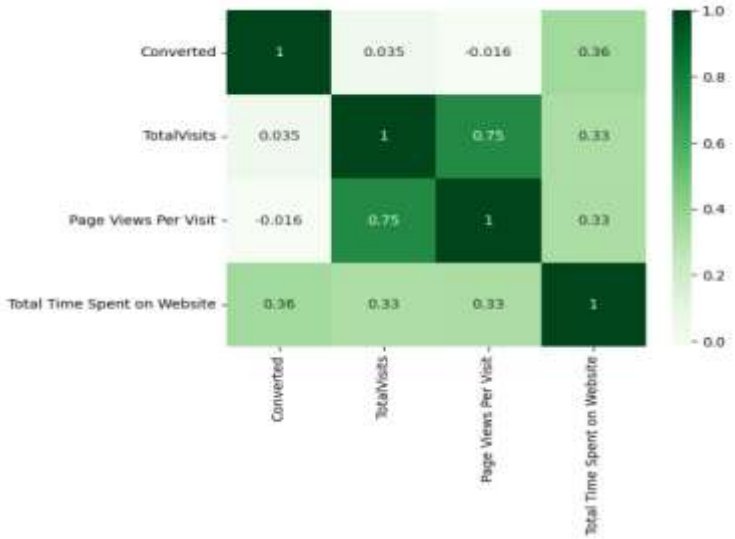


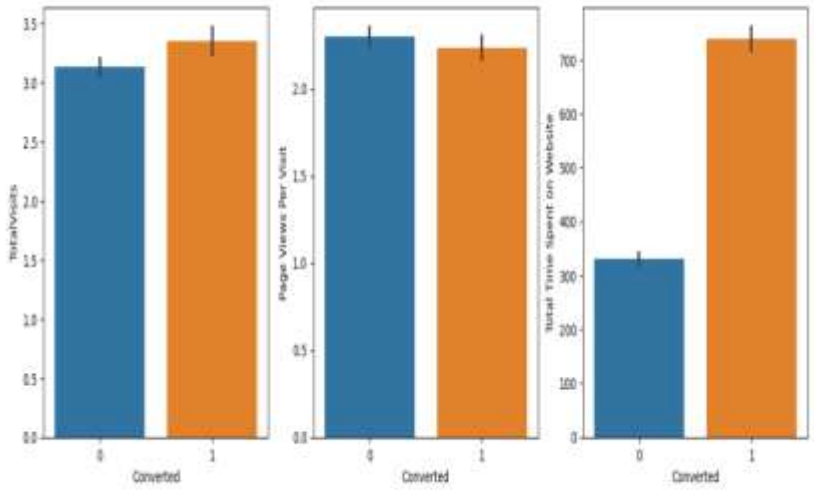
BIVARIATE ANALYSIS



Specialization: Marketing Management, HR Management, Finance Management and Operations Management all show good LCRs, indicating a strong interest among customers in these specializations.

CORRELATION ANALYSIS





Inference:

- There is a strong positive correlation between 'Total Visits' and 'Page Views per Visit', indicating that customers who visit the website more frequently tend to view more pages per visit.
- Customers who spend more time on the website have a higher LCR, indicating that increasing the time spent on the website can lead to higher conversion rates.

DATA PREPARATION

- Binary-level categorical columns were previously converted to 1/0 to suit logistic regression analysis.
- One-hot encoding was employed to create dummy variables for categories like Lead Origin, Lead Source, Last Activity, Specialization, and Current Occupation.
- The dataset was divided into training (70%) and testing (30%) sets to both train the model and test its efficacy on new data.
- Feature scaling via standardization was applied, ensuring all variables were equally weighted, preventing any single feature from overshadowing others.
- To minimize multicollinearity, correlated predictors such as Lead Origin_Lead Import and Lead Origin_Lead Add Form were eliminated.

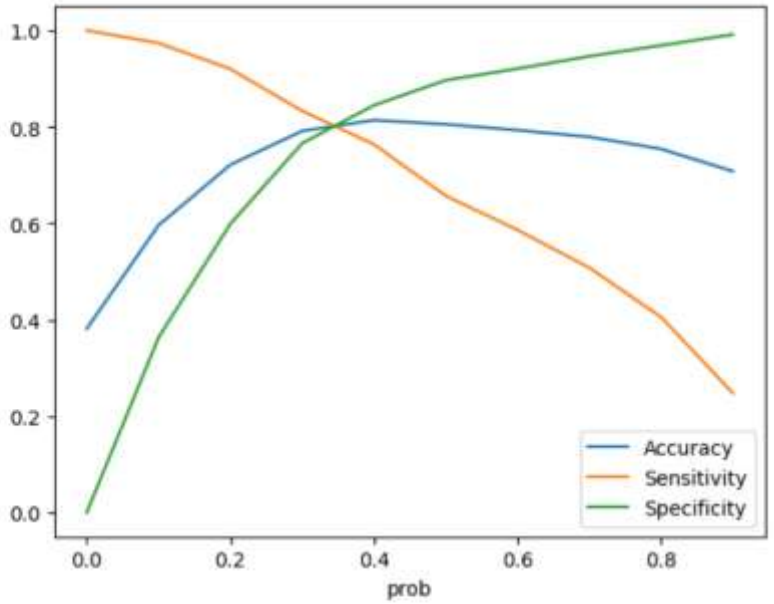
MODEL BUILDING

- The dataset's high dimensionality, with numerous features, posed challenges in model performance and computational efficiency.
- Recursive Feature Elimination (RFE) was utilized to streamline the feature set, narrowing down from 48 to 15 essential columns.
- Initial logistic regression models were basic, progressively refined through manual feature reduction to exclude variables with a p-value over 0.05, enhancing model accuracy and reliability.
- The fourth iteration of the logistic regression model emerged as the final choice, marked by statistically significant predictors ($p < 0.05$) and absence of multicollinearity (VIF < 5), optimizing both evaluation and prediction phases.

MODEL EVALUATION

CONFUSION MATRIX - 1		
Actual/Predicted	not_converted	converted
not_converted	3588	414
converted	846	1620

Accuracy	0.8052
Sensitivity	0.6569
Specificity	0.8966
False Positive Rate	0.1034
Precision	0.7965
Recall	0.6569
Negative Predictive Value	0.8092



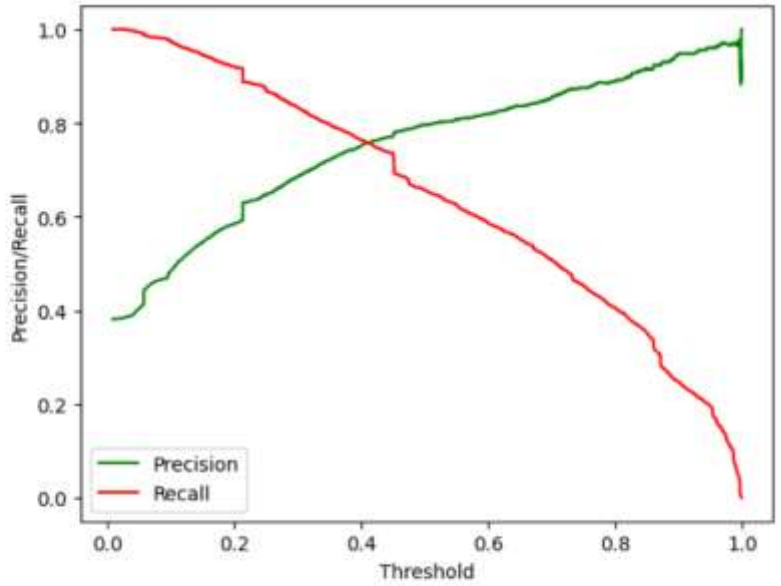
Inference:

- Based on the curve analysis, a cutoff probability of 0.35(approx.) is suggested as the optimal point for classification.

MODEL EVALUATION

CONFUSION MATRIX - 2		
Actual/Predicted	not_converted	converted
not_converted	**3064**	**938**
converted	**412**	**2054**

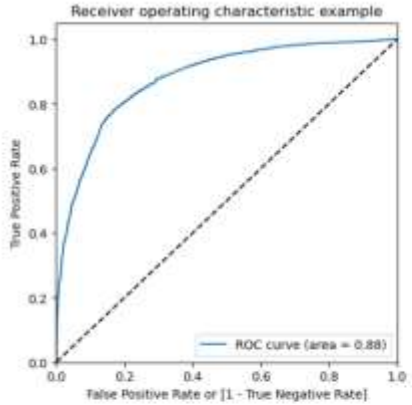
Accuracy	0.8057
Sensitivity	0.7972
Specificity	0.8108
False Positive Rate	0.1892
Precision	0.722
Recall	0.7972
Negative Predictive Value	0.8665



Inference:

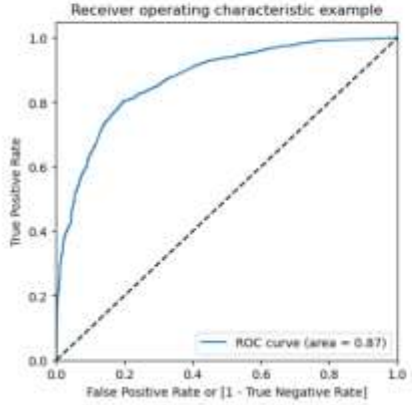
- Based on the precision-recall curve, a threshold of 0.4 provides a good balance between precision and recall.

MAKING PREDICTIONS ON TEST DATASET



ROC Curve – Train Data Set

- The Area under ROC curve was found to be 0.88 out of 1, indicating that the model is a good predictor.
- The curve is plotted as close to the top left corner of the plot as possible, which indicates that the model has a high true positive rate and a low false positive rate at all threshold values.



ROC Curve – Test Data Set

- The Area under ROC curve was found to be 0.87 out of 1, indicating that the model is a good predictor.
- The curve is plotted as close to the top left corner of the plot as possible, which indicates that the model has a high true positive rate and a low false positive rate at all threshold values.

MAKING PREDICTIONS ON TEST DATASET

	Prospect ID	Converted	Converted_Prob	final_predicted	Lead_Score
0	4269	1	0.697934	1	70
1	2376	1	0.860665	1	86
2	7766	1	0.889241	1	89
3	9199	0	0.057065	0	6
4	4359	1	0.87151	1	87
5	9186	1	0.503859	1	50
6	1631	1	0.419681	1	42
7	8963	1	0.154531	0	15
8	8007	0	0.072344	0	7
9	5324	1	0.298849	0	30

Inference:

The customers with a high lead score have a higher chance of conversion and low lead score have a lower chance of conversion.

CONCLUSION

CONFUSION MATRIX - 3		
Actual/Predicted	not_converted	converted
not_converted	1359	318
converted	227	868

Accuracy	0.8034
Sensitivity	0.7927
Specificity	0.8104
False Positive Rate	0.1896
Precision	0.7319
Recall	0.7927
Negative Predictive Value	0.8569

Inference:

Train Data Set:

- Accuracy: 80.57%
- Sensitivity: 79.72%
- Specificity: 81.08%

Test Data Set:

- Accuracy: 80.34%
- Sensitivity: 79.27%
- Specificity: 81.04%

The evaluation metrics of the model are consistently close to each other, indicating that the model is performing consistently across different evaluation metrics in both the test and train datasets. This consistency suggests that the model is reliable and is not overfitting to the training data. It also implies that the model is generalizing well to new data, which is important for real-world applications. The similar performance across evaluation metrics also means that there are no significant biases in the model's predictions. This is a positive sign for the model's performance and provides confidence in its ability to make accurate predictions in the future.

We know that the relationship between ln(odds) of 'y' and feature variable "X" is much more intuitive and easier to understand. The equation is:

ln(odds) = - 1.0236 x const + 1.0498 x Total Time Spent on Website - 1.259 x Lead Origin_Landing Page Submission + 0.9072 x Lead Source_Olark Chat + 2.9253 x Lead Source_Reference + 5.3887 x Lead Source_Welingak Website + 0.9421 x Last Activity_Email Opened - 0.5556 x Last Activity_Olark Chat Conversation + 1.2531 x Last Activity_Others + 2.0519 x Last Activity_SMS Sent - 1.0944 x Specialization_Hospitality Management - 1.2033 x Specialization_Others + 2.6697 x Current_Occupation_Working Professional

- 'Lead Origin_Lead Add Form', 'Current_Occupation_Working Professional' and 'Total Time Spent are effective factors that contribute to a good conversion rate.
- Working professionals and Unemployed customers tend to have higher conversion rates.
- Referral leads generated by old customers have a significantly higher conversion rate
- Google and Direct Traffic are channels that are showing promising conversion rates.
- Leads whose 'Last Activity' is 'SMS Sent' or 'Email Opened' tend to have a higher conversion rate.
- **The 'Others' specialization category is the most common among customers followed by Finance Management, HR Management and Marketing Management.**

RECOMMENDATIONS

- Prioritize features like 'Lead Origin_Lead Add Form', 'Current_Occupation_Working Professional', and website engagement time in lead generation for higher conversion rates.
- Target working professionals more aggressively due to their higher conversion likelihood and financial readiness.
- Enhance referral programs with incentives to leverage their high conversion potential.
- Utilize digital marketing channels, including Google ads and email campaigns, more frequently to boost conversions.
- Focus on leads engaging through 'SMS Sent' or 'Email Opened' activities for higher success rates.
- Enhance website user experience and content to keep potential customers engaged longer, improving conversion chances.
- Tailor course offerings and marketing strategies to popular specializations, like Marketing Management and HR Management, to attract relevant audiences.


thank you!