Lead Score Model Building Report

Problem

Analysis to Enhance Sales for X Education's Online Courses

X Education seeks to boost their sales of online courses by leveraging a data-driven approach. The company has collected data from a previous source, encompassing 9,000 records with various parameters. This analysis aims to develop a predictive model that will assist the sales team in achieving their targets more efficiently and accurately, thus allowing the company to reach its goals in less time.

- **Sales and Marketing Approach:** X Education markets its online courses across multiple online platforms, including search engines and social networking sites, to simplify the registration process for prospective students.

- **Lead Management:** Individuals interested in enrolling must provide their details for an initial discussion. These individuals are categorized as leads. If a lead shows significant interest during the discussion, they are classified as "hot leads." The sales team is expected to focus on these hot leads to increase the likelihood of conversion.

- **Current Challenges:** The sales team currently struggles to identify hot leads, resulting in follow-ups with random leads. This approach has maintained a conversion rate of 30% throughout the year, as all leads are pursued without considering their behaviour or likelihood of conversion.

- **Objective:** The company aims to enhance lead identification to focus on those with the highest probability of conversion. By doing so, X Education hopes to increase its conversion rate from 30% to 80%, thereby meeting annual targets more swiftly and reallocating the sales team to other tasks.

Steps:

- Reading and understanding data
- Data Quality Check
- Data Preparation and Cleanup
- Outlier Treatment
- Dummy Variables Creation
- Test-Train split of data
- Scaling
- Check correlations
- Model building (GLM, RFE, VIF)
- Model Evaluation
- Make Predictions on the Test Set
- Precision-Recall view
- Add lead score

1. **Reading and understanding data**

The data has 9240 rows and 37 columns.

2. **Data Quality Check**

In the initial step, after collecting the data, we proceeded with data verification, which involved reviewing the data and analyzing its properties. During this process, we identified several issues:

- **Null Values:** We observed the presence of null values in the dataset. These were categorized into two types: NaN and 'Select', with the latter indicating optional fields that were left unfilled by users.
- **Biased Columns:** We also discovered columns with biased data, where a single value was predominant, resulting in low variance.

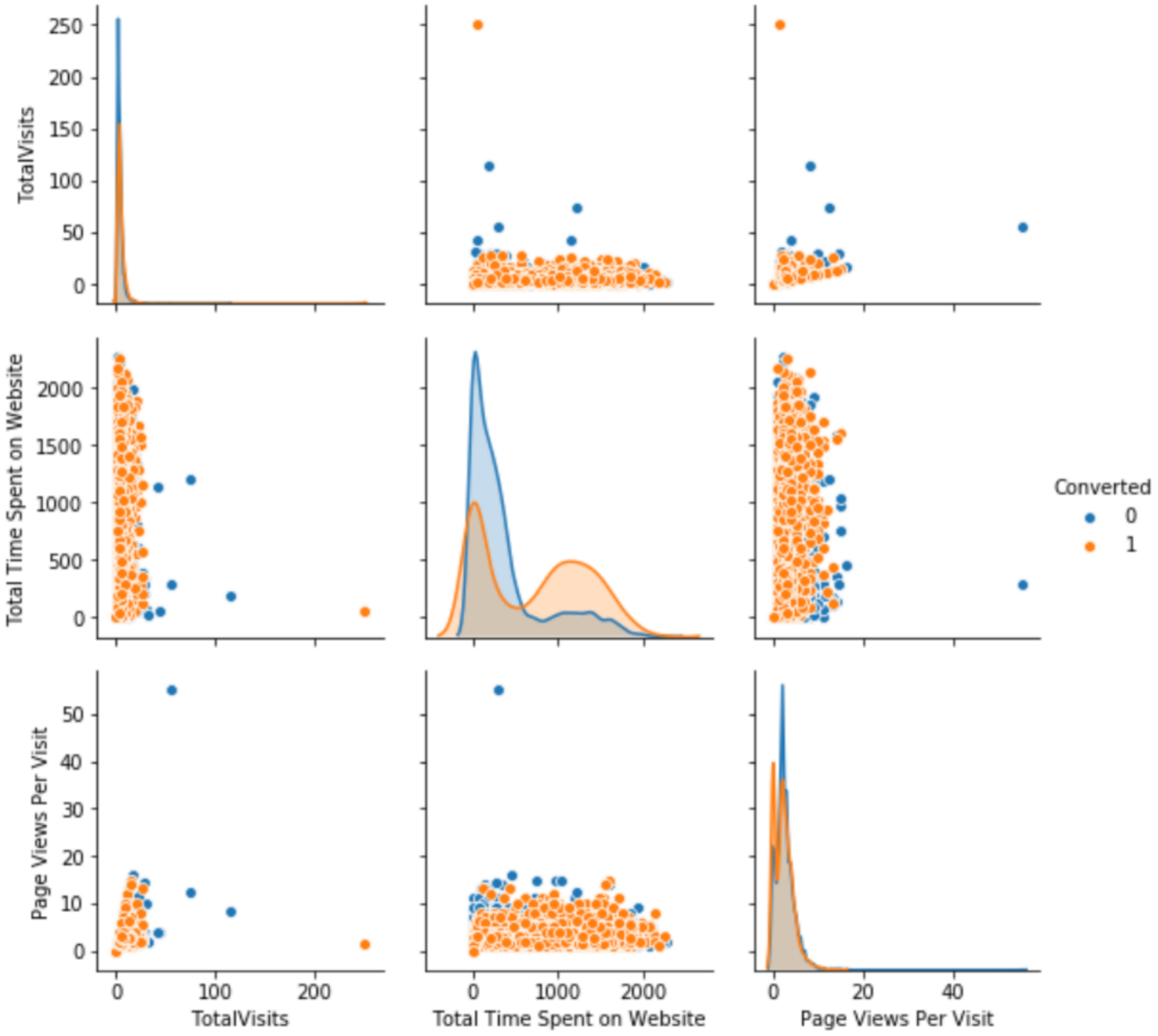
Addressing these issues was crucial for ensuring the quality and reliability of the data for subsequent analysis.

3. **Data cleaning & preparation:**

- **Handling Null Values:** After verifying the data, we proceeded with treating null values using the following approach:

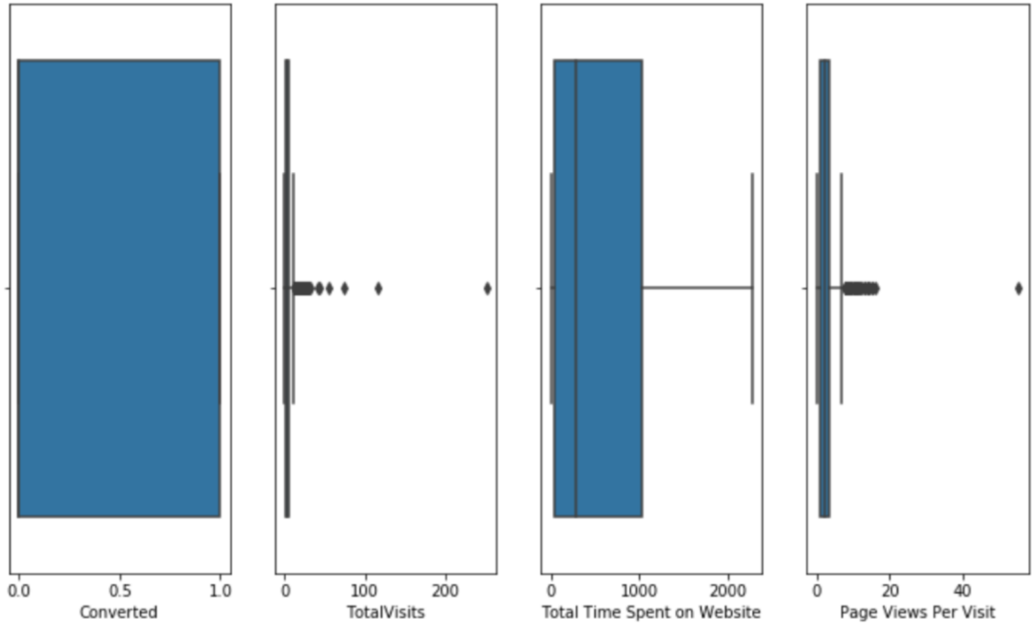
  - **Column Removal:** We first deleted columns with more than 30% null values and removed columns with a high frequency of 'Select' values.
  - **Irrelevant Variables:** We identified that the *City* and *Country* variables were not useful for our analysis since leads from any location should be treated equally. Therefore, we decided to drop these columns.
  - **'Select' Values:** We found that some columns contained 'Select' as a placeholder for missing values. We evaluated these columns and removed those with a significant number of 'Select' entries, such as *Lead Profile* and *How did you hear about X Education*. The *Specialization* column, which also had 'Select' values, was retained due to the relevance of its other data.
  - **Biased Columns:** We noticed that some columns had predominantly single values (e.g., *Do Not Call*, *Search*, *Magazine*). Since these columns mostly contained 'No' and had little variance, they were deemed unhelpful and were removed.
  - **Specific Column Handling:** The column *What is your current occupation* had many null values. Instead of dropping the entire column, which would reduce our feature set significantly, we chose to remove only the rows with null values to retain potentially valuable data.

- **Data Retention:** After performing the cleanup, approximately 69% of the original rows were retained for further analysis.



TotalVisits
Total Time Spent on Website
Page Views Per Visit
250
200
150
100
50
0
2000
1500
1000
500
0
50
40
30
20
10
0
0
100
200
0
1000
2000
0
20
40
Converted
0
1

4. Outlier Analysis and Treatment



Based on the outlier analysis from the plots for *TotalVisits* and *Page Views Per Visit*, we removed approximately 98 records.

5. Dummy Creation:

 Categorical Variable Conversion: To build the logistic regression model for identifying the conversion rate of hot leads, we converted all categorical variables into numerical format using dummy variable creation with the pandas.get_dummies method.

 Data Preparation: After completing the above tasks, we cleaned and prepared the dataset for modeling, ensuring it was ready for use in the logistic regression analysis.

6. Test-Train split

The dataset is split into 70% train and 30% test data.

7. Scaling

We used MinMaxScaler to scale the variables to bring them into same value range.

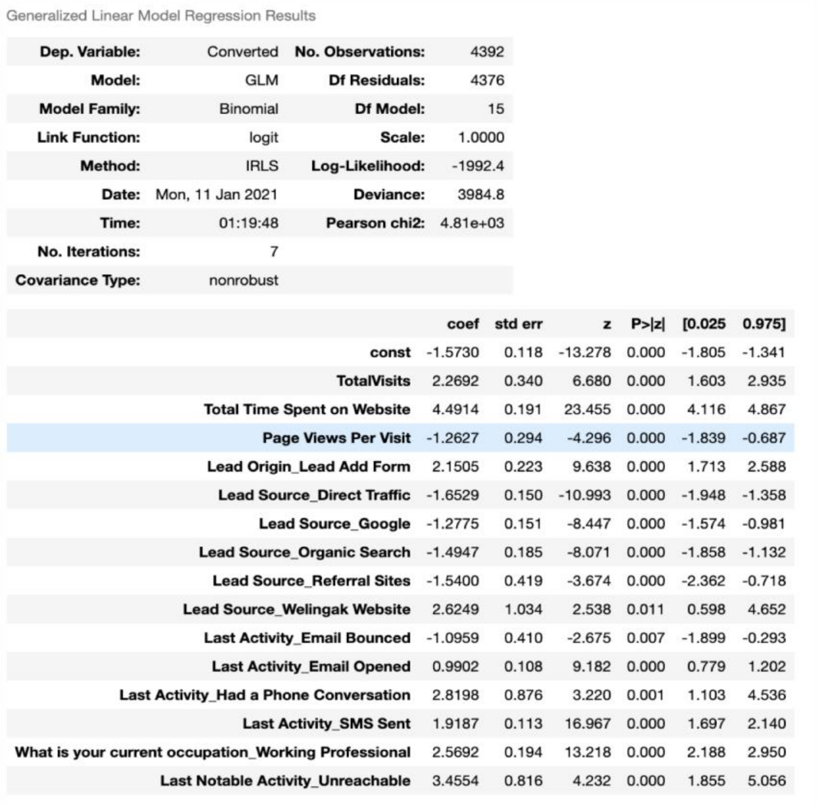
8. Correlations

There are columns with high correlation. We can drop them. However not dropping it here and will eliminate the features in modelling part.

9. Model Building

Model 1:

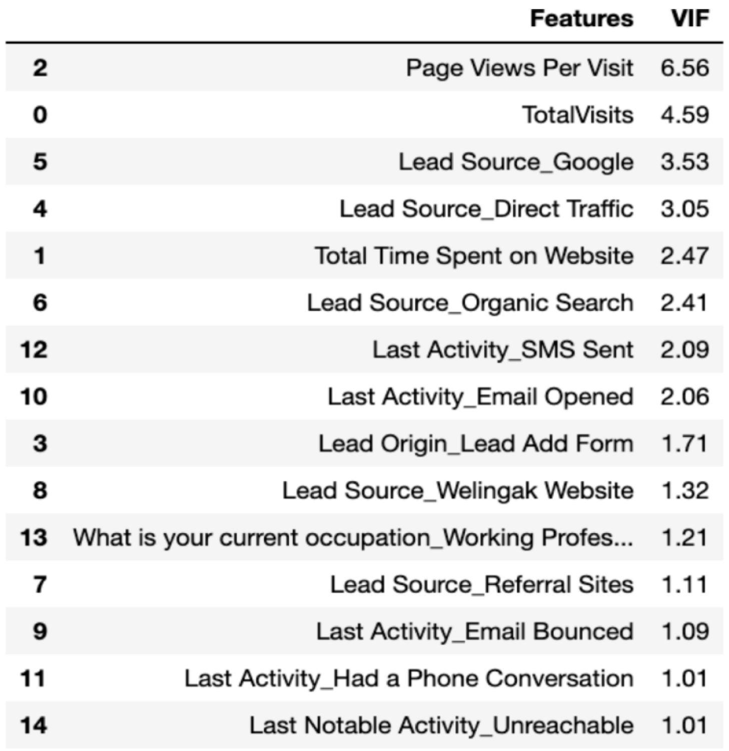
- **Initial Model Building:** We initially constructed a sample model to review and visualize the summary of all variables, including their p-values and coefficients. To enhance the model's robustness, we employed the Recursive Feature Elimination (RFE) technique to select the top 15 features from the dataset based on coefficient values and feature importance.
- **Feature Selection and Model Refinement:** Using the top 15 features identified through RFE, we proceeded to build the model, carefully examining the coefficients and p-values of these selected features to ensure their relevance and contribution to the model's performance.



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4392
Model:	GLM	**Df Residuals:**	4376
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1992.4
Date:	Mon, 11 Jan 2021	**Deviance:**	3984.8
Time:	01:19:48	**Pearson chi2:**	4.81e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

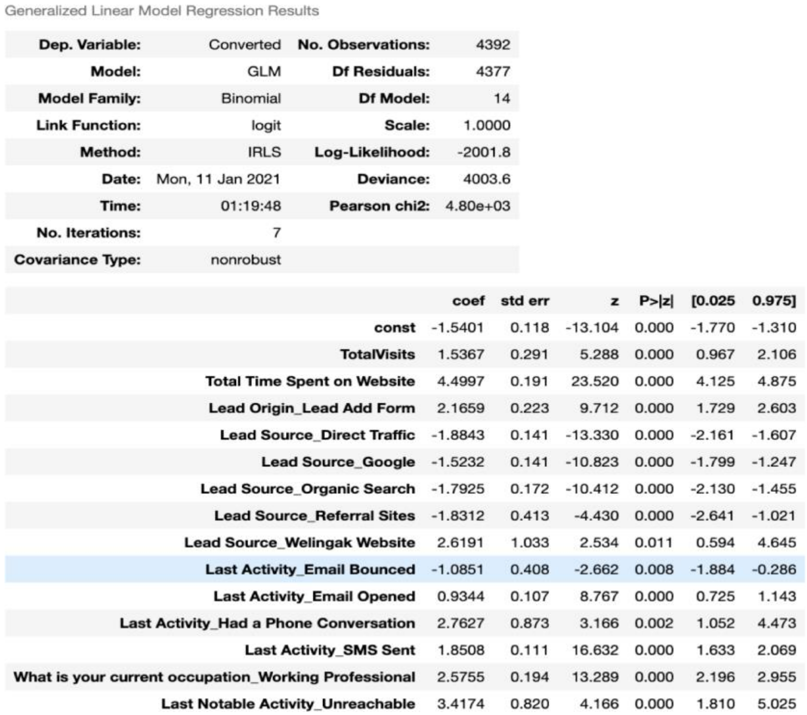
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5730	0.118	-13.278	0.000	-1.805	-1.341
TotalVisits	2.2692	0.340	6.680	0.000	1.603	2.935
Total Time Spent on Website	4.4914	0.191	23.455	0.000	4.116	4.867
Page Views Per Visit	-1.2627	0.294	-4.296	0.000	-1.839	-0.687
Lead Origin_Lead Add Form	2.1505	0.223	9.638	0.000	1.713	2.588
Lead Source_Direct Traffic	-1.6529	0.150	-10.993	0.000	-1.948	-1.358
Lead Source_Google	-1.2775	0.151	-8.447	0.000	-1.574	-0.981
Lead Source_Organic Search	-1.4947	0.185	-8.071	0.000	-1.858	-1.132
Lead Source_Referral Sites	-1.5400	0.419	-3.674	0.000	-2.362	-0.718
Lead Source_Welingak Website	2.6249	1.034	2.538	0.011	0.598	4.652
Last Activity_Email Bounced	-1.0959	0.410	-2.675	0.007	-1.899	-0.293
Last Activity_Email Opened	0.9902	0.108	9.182	0.000	0.779	1.202
Last Activity_Had a Phone Conversation	2.8198	0.876	3.220	0.001	1.103	4.536
Last Activity_SMS Sent	1.9187	0.113	16.967	0.000	1.697	2.140
What is your current occupation_Working Professional	2.5692	0.194	13.218	0.000	2.188	2.950
Last Notable Activity_Unreachable	3.4554	0.816	4.232	0.000	1.855	5.056



	Features	VIF
2	Page Views Per Visit	6.56
0	TotalVisits	4.59
5	Lead Source_Google	3.53
4	Lead Source_Direct Traffic	3.05
1	Total Time Spent on Website	2.47
6	Lead Source_Organic Search	2.41
12	Last Activity_SMS Sent	2.09
10	Last Activity_Email Opened	2.06
3	Lead Origin_Lead Add Form	1.71
8	Lead Source_Welingak Website	1.32
13	What is your current occupation_Working Profes...	1.21
7	Lead Source_Referral Sites	1.11
9	Last Activity_Email Bounced	1.09
11	Last Activity_Had a Phone Conversation	1.01
14	Last Notable Activity_Unreachable	1.01

- **Feature Significance and Model Iteration:** Features with high p-values are considered less significant. Therefore, we drop such features and rebuild the model using the remaining ones. This iterative process continues until we achieve a model with lower p-values for all features.
- **Model Adjustment:** In this case, all variables had p-values below 0.05. However, the Variance Inflation Factor (VIF) indicated a high score for the *Page Views Per Visit* column, which was highly correlated with *TotalVisits*. To address multicollinearity, we removed the *Page Views Per Visit* column and re-ran the model.

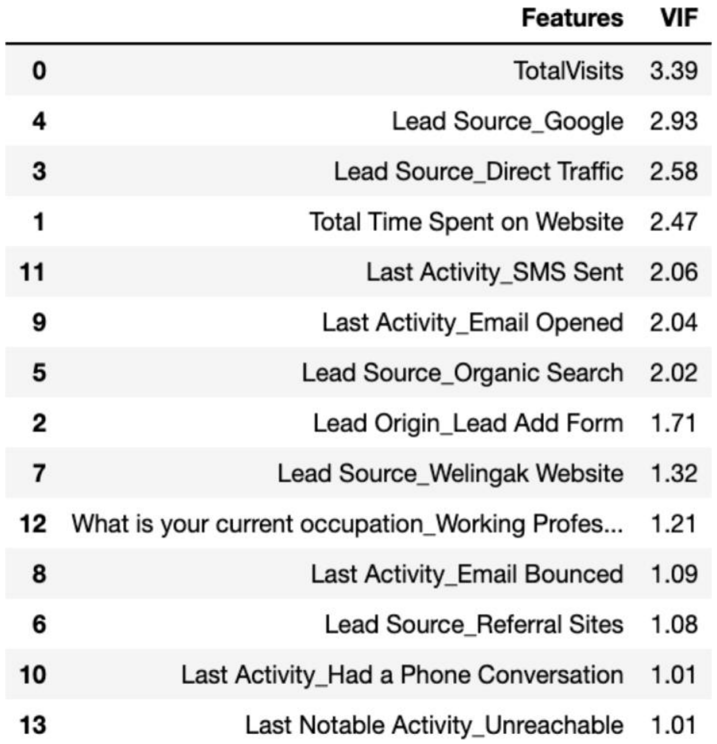
Model 2



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4392
Model:	GLM	**Df Residuals:**	4377
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2001.8
Date:	Mon, 11 Jan 2021	**Deviance:**	4003.6
Time:	01:19:48	**Pearson chi2:**	4.80e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-1.5401	0.118	-13.104	0.000	-1.770	-1.310
TotalVisits	1.5367	0.291	5.288	0.000	0.967	2.106
Total Time Spent on Website	4.4997	0.191	23.520	0.000	4.125	4.875
Lead Origin_Lead Add Form	2.1659	0.223	9.712	0.000	1.729	2.603
Lead Source_Direct Traffic	-1.8843	0.141	-13.330	0.000	-2.161	-1.607
Lead Source_Google	-1.5232	0.141	-10.823	0.000	-1.799	-1.247
Lead Source_Organic Search	-1.7925	0.172	-10.412	0.000	-2.130	-1.455
Lead Source_Referral Sites	-1.8312	0.413	-4.430	0.000	-2.641	-1.021
Lead Source_Welingak Website	2.6191	1.033	2.534	0.011	0.594	4.645
Last Activity_Email Bounced	-1.0851	0.408	-2.662	0.008	-1.884	-0.286
Last Activity_Email Opened	0.9344	0.107	8.767	0.000	0.725	1.143
Last Activity_Had a Phone Conversation	2.7627	0.873	3.166	0.002	1.052	4.473
Last Activity_SMS Sent	1.8508	0.111	16.632	0.000	1.633	2.069
What is your current occupation_Working Professional	2.5755	0.194	13.289	0.000	2.196	2.955
Last Notable Activity_Unreachable	3.4174	0.820	4.166	0.000	1.810	5.025



	Features	**VIF**
0	TotalVisits	3.39
4	Lead Source_Google	2.93
3	Lead Source_Direct Traffic	2.58
1	Total Time Spent on Website	2.47
11	Last Activity_SMS Sent	2.06
9	Last Activity_Email Opened	2.04
5	Lead Source_Organic Search	2.02
2	Lead Origin_Lead Add Form	1.71
7	Lead Source_Welingak Website	1.32
12	What is your current occupation_Working Profes...	1.21
8	Last Activity_Email Bounced	1.09
6	Lead Source_Referral Sites	1.08
10	Last Activity_Had a Phone Conversation	1.01
13	Last Notable Activity_Unreachable	1.01

Based on the summary report of the model, we finalized it because all features had p-values below 0.05 and VIF scores were less than 5.

The logistic regression model provided valuable insights into addressing X Education's business problem by utilizing a Generalized Linear Model (GLM). Our analysis revealed that all variables in the model had p-values below 0.05, indicating statistical significance, and VIF scores were all below 5, suggesting no multicollinearity issues. The coefficients derived from the model offer a detailed understanding of how each feature contributes to the probability of lead conversion.

Here’s a detailed breakdown of the key variables and their impact on lead conversion:

- **Total Time Spent on Website (coef = 4.4997):** This variable has the highest coefficient, indicating a strong positive relationship with the probability of lead conversion. Leads who spend more time on the website are significantly more likely to convert, as the time spent is a strong indicator of their interest and engagement.

- **Last Notable Activity (coef = 3.4174):** When the last notable activity of a lead is recorded as 'not present' or 'unreachable', it surprisingly has a positive coefficient, suggesting that these leads have a higher likelihood of conversion. This could imply that despite initial difficulties in reaching them, these leads may eventually show strong interest or commitment.

- **Last Activity (coef = 2.7627):** The coefficient for this variable indicates that leads whose most recent activity was a phone call have a higher chance of conversion. Phone calls may signify more serious interest or a higher level of engagement compared to other activities.

- **Lead Source (coef = 2.6191):** The source of the lead plays a significant role in conversion probability. Leads originating from the Welingak website have a higher likelihood of conversion, suggesting that this source is particularly effective in attracting potential customers who are more likely to enroll in the courses.

- **Current Occupation (coef = 2.5755):** Working professionals are more likely to convert compared to other occupations. This coefficient suggests that professionals may be more interested in the courses offered or better positioned to make decisions about enrolling.

Overall, the model’s results provide actionable insights for the sales team at X Education, allowing them to focus their efforts on high-potential leads and tailor their strategies based on the identified influential factors. This approach is expected to significantly enhance lead conversion rates and help the company achieve its sales targets more efficiently.

10. Model Evaluation

To evaluate the performance of the model, we calculated accuracy, sensitivity, and specificity on both the training and test datasets:

- **Accuracy:** 79.6%
- **Sensitivity:** 75.3%
- **Specificity:** 83.6%

Confusion Matrix Analysis:

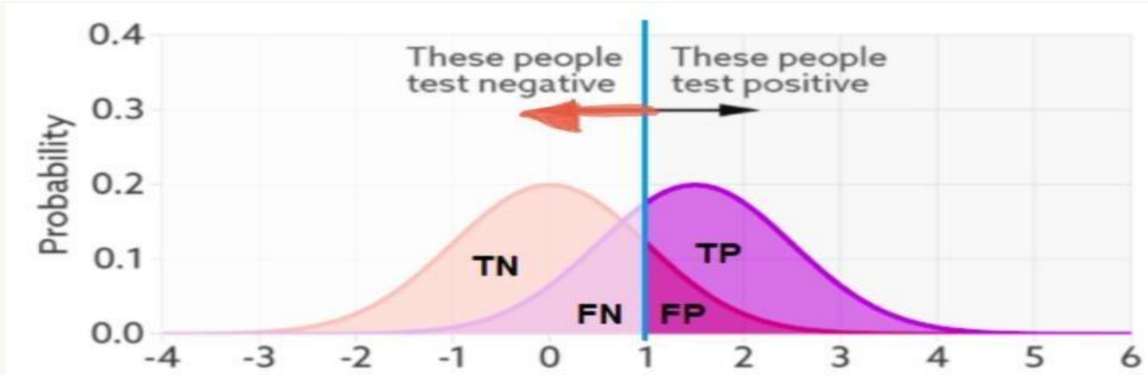
- **Confusion Matrix Design:** We constructed the confusion matrix by predicting the dependent variable (y) to compare and assess the accuracy, sensitivity, and specificity of the actual versus predicted values.

- **Accuracy:** This metric represents the ratio of correctly predicted values to the total number of values. It provides an overall measure of the model's performance.

- **Sensitivity (True Positive Rate or Recall):** Sensitivity is defined as the ratio of correctly predicted positive results to the total actual positive results. It measures the model's ability to identify true positives effectively.

- **Specificity:** Specificity is the ratio of correctly predicted negative results to the total actual negative results. It assesses the model's ability to identify true negatives accurately.

- **Precision:** In addition to accuracy, sensitivity, and specificity, if X Education wants to understand the ratio of positive predictions out of the total predicted results, precision is also considered. Precision measures the proportion of true positive results among all positive predictions made by the model.

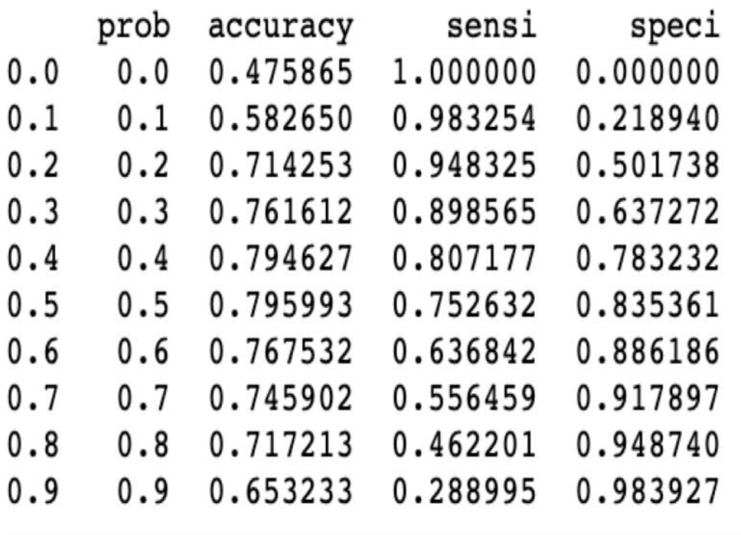


- **Confusion Matrix Distribution:** The confusion matrix provides insights into the values of True Positive, True Negative, False Positive, and False Negative:
  - **True Positive (TP):** The number of actual positive instances that were correctly predicted as positive.
  - **True Negative (TN):** The number of actual negative instances that were correctly predicted as negative.
  - **False Positive (FP):** The number of actual negative instances that were incorrectly predicted as positive.
  - **False Negative (FN):** The number of actual positive instances that were incorrectly predicted as negative.
- **Prediction and Cutoff Value:** We initially used a cutoff value of 0.5 to classify predictions. If the predicted probability was 0.5 or higher, it was classified as positive; otherwise, it was classified as negative.
- **Optimization Technique:** Recognizing that a static cutoff of 0.5 might not yield the best results, we applied additional optimization techniques to determine the optimal threshold value. This adjustment helps in achieving better performance and accuracy in the model's predictions.

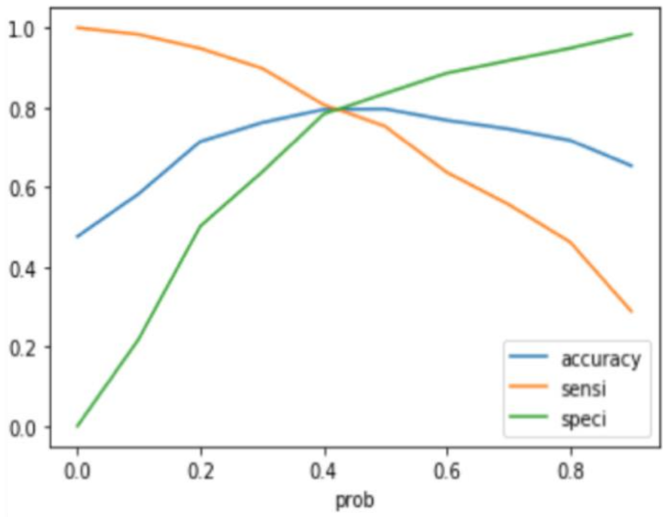
ROC (Receiver Operating Characteristic):-

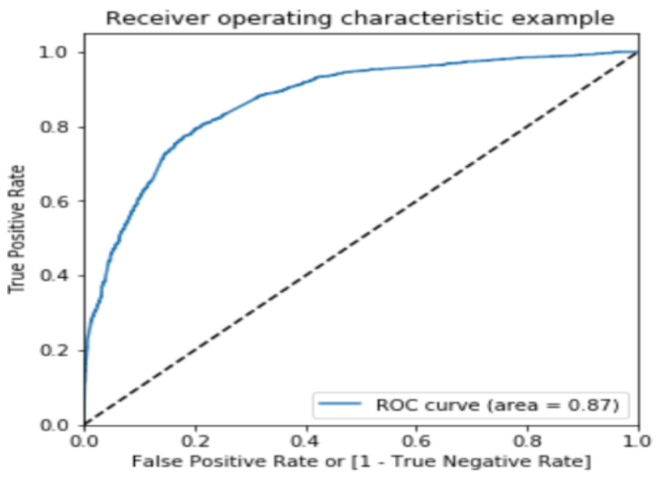
The distribution of sensitivity and specificity is used to define the Receiver Operating Characteristic (ROC) curve, which assesses the model's performance. A larger area under the ROC curve indicates a better model, reflecting its effectiveness in distinguishing between classes.

ROC analysis is a valuable tool for selecting optimal models and eliminating suboptimal ones, regardless of the cost context or class distribution.



```
     prob  accuracy     sensi     speci
0.0   0.0  0.475865  1.000000  0.000000
0.1   0.1  0.582650  0.983254  0.218940
0.2   0.2  0.714253  0.948325  0.501738
0.3   0.3  0.761612  0.898565  0.637272
0.4   0.4  0.794627  0.807177  0.783232
0.5   0.5  0.795993  0.752632  0.835361
0.6   0.6  0.767532  0.636842  0.886186
0.7   0.7  0.745902  0.556459  0.917897
0.8   0.8  0.717213  0.462201  0.948740
0.9   0.9  0.653233  0.288995  0.983927
```



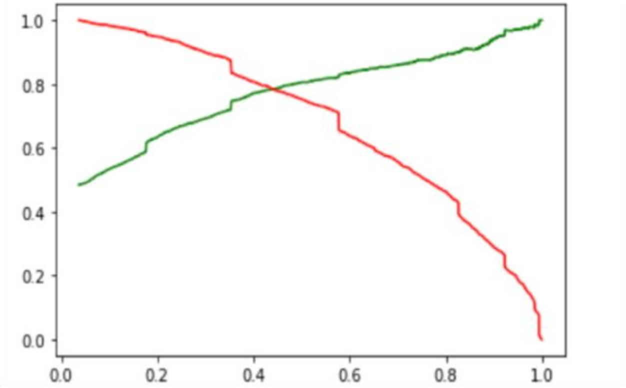


- **Optimal Threshold Determination:** To find the optimal threshold value, we calculated and analyzed accuracy, sensitivity, and specificity across different cut-off ranges from 0.0 to 0.9. The intersection point of these parameters, which was identified as 0.42, was determined to be the optimal cut-off point.

- **Model Verification:** Using the optimal threshold of 0.42, we evaluated the training data by predicting the dependent variable. The resulting metrics were:
  - **Accuracy:** 79.5%
  - **Sensitivity:** 79.6%
  - **Specificity:** 79.4%

- **Model Testing:** To assess the model's correctness, we predicted lead conversion probabilities and obtained a robust ROC curve, along with optimal rates for accuracy, sensitivity, and specificity.

11. Precision-Recall Tradeoff



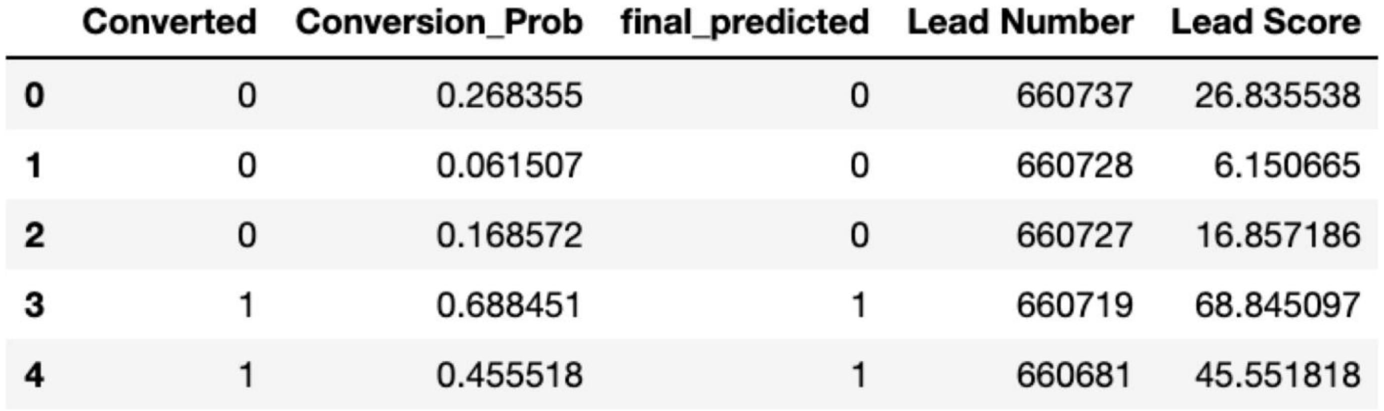
Here, the optimal threshold is identified as 0.44, with the following performance metrics:

- **Accuracy:** 79.5%
- **Precision:** 78.4%
- **Recall:** 78.5%

Based on the business requirement, we assigned a lead score ranging from 1 to 100 to each record to distinguish between high-quality and low-quality leads.

12. Add Lead Score

Using conversion probability predicted by the model to calculate lead score.



	Converted	Conversion_Prob	final_predicted	Lead Number	Lead Score
0	0	0.268355	0	660737	26.835538
1	0	0.061507	0	660728	6.150665
2	0	0.168572	0	660727	16.857186
3	1	0.688451	1	660719	68.845097
4	1	0.455518	1	660681	45.551818

Based on the model, we assigned a lead score to each lead. This score helps the company target potential leads more effectively. A higher score indicates that a lead is 'hot' and more likely to convert, while a lower score suggests that the lead is 'cold' and less likely to convert. The accuracy of the model is approximately 80%, aligning with the objectives outlined in the problem statement.

Overall Benefit Analysis:

Addressing X Education's business challenge, we developed a model to enhance the lead conversion rate from various platforms. Given the current conversion rate is only 30%, our model assigns a score between 0 and 100 to effectively identify 'hot' leads, achieving approximately 80% accuracy.

Key Benefits:

- **Efficient Lead Identification:** The model reduces the effort required by salespersons to identify hot leads, allowing them to focus on converting these leads more effectively.
- **Increased Conversion Rate:** Salespersons can meet their hot lead conversion targets more quickly, leading to higher productivity.
- **Resource Optimization:** Once targets are met, sales resources can be redirected to other valuable company tasks, optimizing overall resource use.
- **Financial Growth:** Improved lead conversion rates will contribute to increased revenue and business growth for X Education.
- **Enhanced Employee Retention:** Skilled personnel can be better managed with bonuses and incentives, reducing the risk of turnover.

Overall, X Education stands to gain substantial financial benefits and improve its operational efficiency.

CONCLUSION:

- **Recommendations:** Based on the comprehensive analysis, we identified how to enhance sales and increase the lead conversion rate. The sales team should prioritize leads identified as 'hot' by the model. This approach will boost efficiency and reduce the effort needed to engage with leads.

- **Analysis and Model Development:** The analysis involved insights from exploratory data analysis and the development of a logistic regression model. The model assigns a predicted score to each lead, enabling the sales team to target high-potential leads more effectively.

- **Model Performance and Business Impact:** The model, trained to distinguish between good and bad leads, has a sensitivity of nearly 80%, aligning with the business requirements. This will aid in improving conversion rates and achieving business goals more efficiently.