UpGrad & IIITB | DS-C61 Course: Machine Learning Case Study: Lead Scoring

Submitted By: Prathamesh Kulkarni

Contents

Problem Statement

1. X Education Company markets its courses online.
2. When people land on X Company's website to browse course, they may or may not fill up a form for a course.
3. When company acquires email address or phone number of potential buyers of courses, they are classified to be a lead.
4. When company representatives make calls to such leads, there is a chance that they may or may not buy a course.
5. The existing percentage of leads getting converted is approximately 30%.
6. The company wants to build a scoring system to assign scores to all such leads, so that they can be focused more on, and the conversion rate increases to the target levels of 80%.

Goals of Analysis

1. Using historical data of the company to learn patterns in the data of leads which were converted in the past.
2. Building a Logistic Regression Model by applying Machine Learning techniques on historical data.
3. Gaining insights into the data and finding the most significant attributes of potential buyers, which point to higher chances of lead conversion.
4. Answering a few questions posed by the business, which could help the company in the future to tweak the strategy and convert more leads or just enough leads (as situation demands.)

Analysis Approach (Before Building ML Model)

Step 1	Step 2	Step 3	Step 4	Step 5
Data inspection to find any discrepancies/problems/missing values and determining methods to handle the same.	Preparing the data for machine learning by converting data values into numerical forms and handling missing values as necessary.	Visualizing and analyzing data using graphs and plots to check if any further fine-tuning of data is required (for example - to handle data outliers.)	Splitting the data into training and test datasets, so that one dataset can be used purely for building the machine learning model and the other one to evaluate performance of the model on unseen data.	Scaling numerical variables on the same scale for efficiency and effectiveness of model building.

Analysis Approach (Building and Evaluating ML Model)

Step 1	Step 2	Step 3	Step 4	Step 5

Building the first Logistic Regression ML model using StatsModels library on whole treated dataset, considering all feature variables.

Using RFE for auto-selecting top 25 feature variables and building second model based on those 25 variables.

Iteratively narrowing down number of features selected, based on their significance and VIF values.

Evaluating metrics such as accuracy, sensitivity, specificity, precision and recall of the final model which contains all significant variables with very low VIF values.

Applying final model to test dataset and evaluating metrics again.

Analysis Outcomes - Visualizations

1. We plotted a heatmap of correlations between all feature variables of the dataset to numerical form, however since the number of variables was huge, the heatmap was inconclusive.
2. We also checked 3 feature variables for outliers in the data and treated them as required:
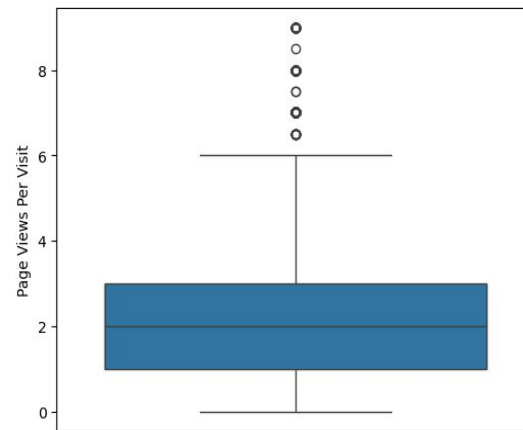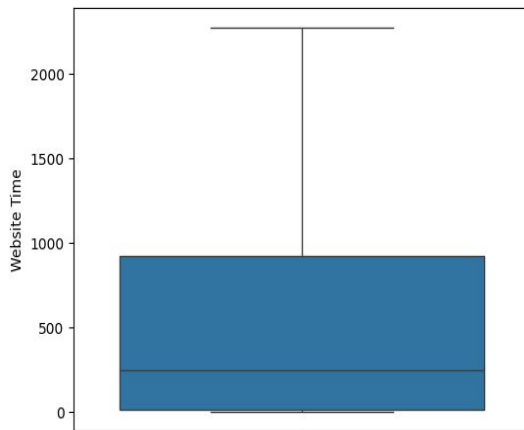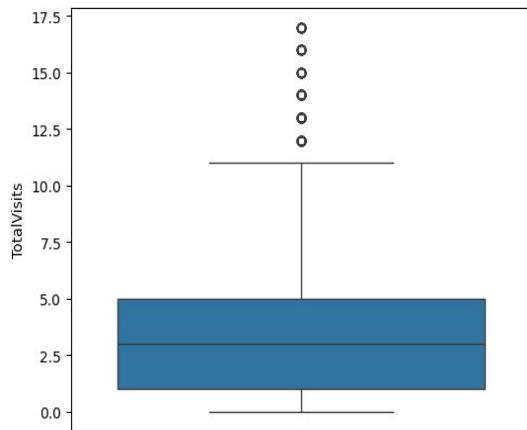
Analysis Outcomes - Visualizations (Continued)

After building the Logistic Regression ML model, we plotted False Positives predicted by the model against True Positives (ROC Curve), which ideally should be as far from 45° diagonal of ROC Space. Following is the plot that we got:
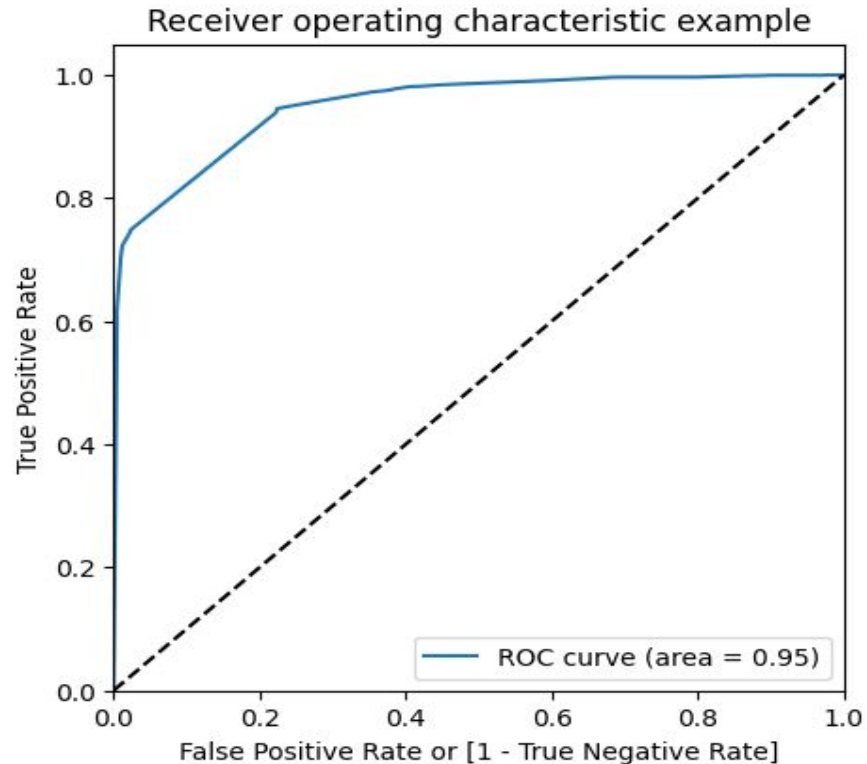
Analysis Outcomes - Visualizations (Continued)

For striking good balance between sensitivity and specificity, we plotted 10 levels of probability against sensitivity and specificity, and found that at probability level of approx. 0.34, sensitivity and specificity meet accuracy. In other words, beyond probability of 0.34, a lead conversion can be predicted to be 1 and below 0.34 as 0.
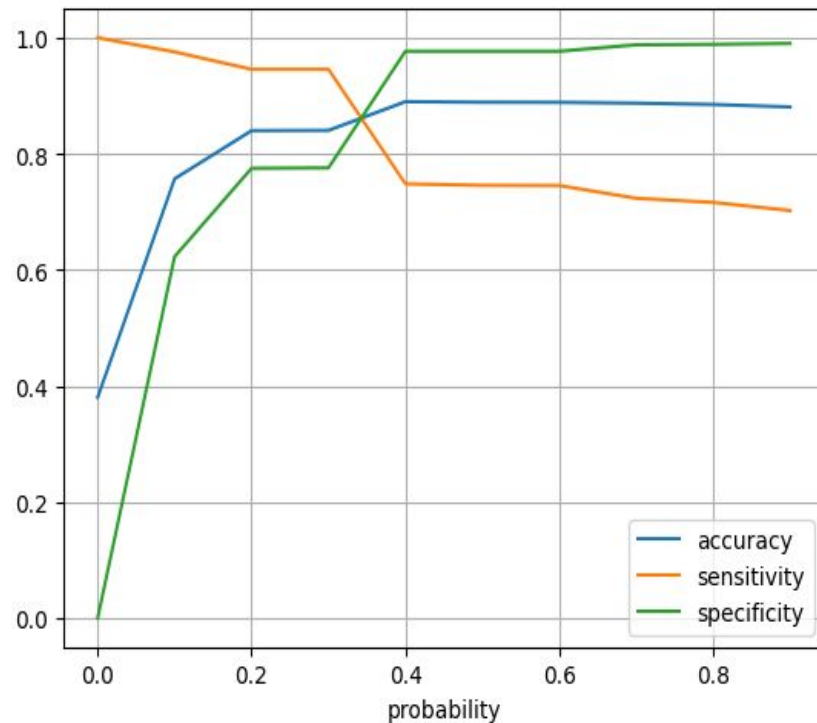
Analysis Outcomes - Logistic Regression Model

1. We began building the first LR ML model with all feature variables after treating and preparing the data and iteratively reduced the number.
2. The final model, built after 12 iterations of model creation and analysis, has **15 significant features**:

Feature Variable	Coefficient	Feature Variable	Coefficient	Feature Variable	Coefficient
Lead Source_Welingak Website	4.449	Tags_Busy	1.3066	Tags_Will revert after reading the email	4.6554
Last Activity_Email Bounced	-2.1832	Tags_Closed by Horizzon	6.5189	Tags_switched off	-3.4303
Last Activity_Olark Chat Conversation	-1.5393	Tags_Interested in other courses	-1.861	Lead Quality_Worst	-2.4525
Last Activity_Page Visited on Website	-1.1318	Tags_Lost to EINS	5.8156	Asymmetrique Activity Index_03.Low	-1.7713
Tags_Already a student	-2.5678	Tags_Ringing	-2.8074	Last Notable Activity_Modified	-1.3672

Analysis Outcomes - Logistic Regression Model (Continued)

Following are the performance metrics of the final model built on training and test datasets:

Training Dataset		Test Dataset	
Accuracy	0.84	Accuracy	0.84
Sensitivity	0.95	Sensitivity	0.94
Specificity	0.78	Specificity	0.79
Precision	0.72	Precision	0.72
Recall	0.95	Recall	0.94

Analysis Outcomes - Business Implications

1. The 15 feature variables selected by the final model signify the following:
   a. Positive coefficient of dummy feature variables such as 'Tag_Closed by Horizzon', 'Tags_Lost to EINS' and 'Tags_Will revert after reading the email' signify that:

      Tags assigned to potential customers are good indicators of their willingness to buy a course.

   b. Negative coefficients of all dummy variables created from categorical variable 'Last Activity' and 'Last Notable Activity' signify that:

      Paying attention to a potential customer's activities can pay off. Some activities are strong indicators of a bad lead.

   c. Positive coefficient of dummy variable 'Lead Source_Welingak Website' indicates that:

      Any lead which has source 'Welingak Website' has strong chances of converting.

   d. Company can make good use of other information such as 'Lead Quality' and 'Asymmetrique Activity Index' to identify good or bad leads.

Questions and Answers

X company had posed following business questions. The analysis performed and the LR Model that was built, help in answering these questions:

1.   Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?

Answer: -

   I.   Tags_Closed by Horizzon (coeff. = 6.5189)
  II.   Tags_Lost to EINS (coeff. = 5.8156)
 III.   Tags_Will revert after reading the email (coeff. = 4.6554)

Questions and Answers (Continued)

2.   What are the top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion?

Answer: -

  I.   Tags
 II.   Last Activity
III.   Lead Quality

Questions and Answers (Continued)

3.   X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

Answer: - After building the logistic regression model, we built a cutoff matrix for all levels of probability, which helped in determining the optimal cutoff to strike a good balance between sensitivity and specificity. In this particular case, we found that the optimal cutoff probability is 0.34 (Company can use Lead Score instead, which is equal to Conversion Probability x 100 = 0.34 x 100 = 34), beyond which, any lead can be considered as convertible. However, when the company hires interns and wants to convert all the potential leads by making as many phone calls as possible, a good strategy would be to bring this cutoff down. In other words, the company must focus more on sensitivity of the model, than the specificity. If we bring the cutoff down from 0.34 to, let's say, 0.3 (Lead Score = 30), interns will make phone calls to more potential leads and more leads will probably be converted.

Questions and Answers (Continued)

4.	Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Answer: - As opposed to the strategy employed during the months when interns are available to make as many phone calls as possible, to reduce the number of phone calls, the company must increase the cutoff for the Lead Score. In other words, instead of calling all leads above the Lead Score of 34, the company can decide to call only the leads above Lead Score of 60. This way, the company will focus more on specificity by reducing False Positives in the lead conversion attempt.

Conclusion

1. Logistic regression model can help find significant variables in data pointing to good or bad leads, so that company can focus on customers as situation demands.
2. Based on the situation, company can choose the threshold of Lead Score beyond which a lead can be considered and focused on.
3. Good balance between sensitivity and specificity shows that the ideal level of Lead Score is 34, to avoid too many False Positives or too many True Negatives in lead conversion.
4. The final model built has accuracy level of 84% on both training and test datasets.

Thank you!