LEAD SCORING CASE STUDY

Logistic Regression Assignment Submission

Student Name: Tarriq Ferrose Khan

Batch : DSC 68

upGrad

upGrad & IIITB | Data Science Program - May 2024

Problem Statement

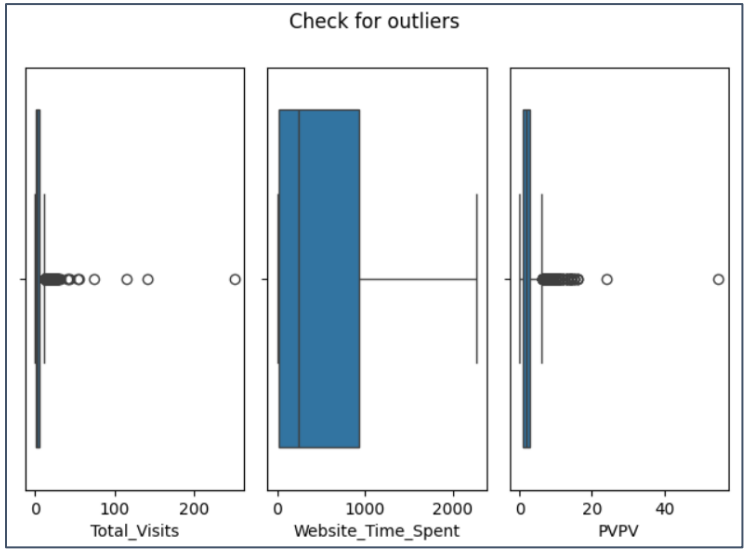
An education company named **X Education** sells online courses to industry professionals, marketing on their own website, several other websites and search engines like Google. Currently **Leads** are generated by people filling up a form on the website (providing email address or phone number), and through past referrals as well. Sales team starts reaching out to these Leads and try to get them converted to as **Prospects** that would but the courses, for which the current **Lead Conversion Rate is around 30%**. For Example, if 100 Leads are acquired, 30 are converted. The Company wishes to identify the most Potential Leads also know as Hot Leads , so that the Lead Conversion Rate should go up as the sales team will now be communicating with the potential leads rather than reaching everyone.
The CEO, in particular, has given the ballpark of the target lead conversion rate to be around 80%.

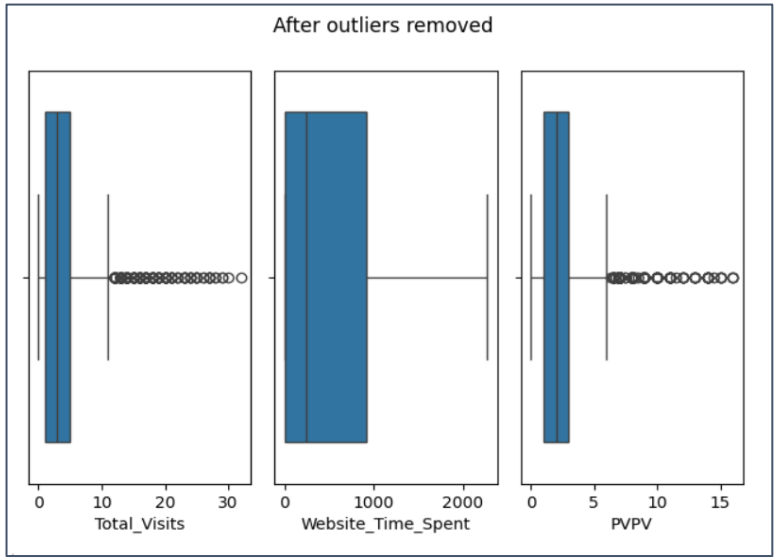
Goal: A logistic Regression Model needs to built to assign a Lead Score between 0 to 100 , that higher the score the customer is likely to Convert (Hot Lead) and vice versa holds true.

Approach Note

Data Cleaning and Preparation:

- **Understanding**: Source Data (Lead.csv) contained 9000+records and 37 columns, just dropna reduced to 1000+ records which will not be sufficient for modelling
- **Column Rename:** Renamed lengthy columns with Meaning full names (For eg., : “How did you hear about X Education “ as “Marketing_Source”).
- **‘Select’ Columns**:Handle Columns with 'Select' Values, replace with null as ‘Select’ is the default value in the User form which is as good as NULL
- **Null Columns:** Dropped columns with nullvalue counts>3000 and other columns not needed.
- **Yes/No Columns:** Replace Columns having ‘Yes/No’ columns with ‘1/0’ respectively.
- **Only Zero columns :**Dropped those columns where it has only 0 as the value and few more columns as it will not be needed for any correlation.
- **Dummy Variables** : Get all Category Columns and Create Dummy Variables using pd.get_dummies function , with drop_first=True
- **Drop _nan columns**
- **Outlier Handling:** For Continuous Variables Columns ['Total_Visits', 'Website_Time_Spent', 'PVPV’], checked using user defined function and removed rows above certain threshold values.

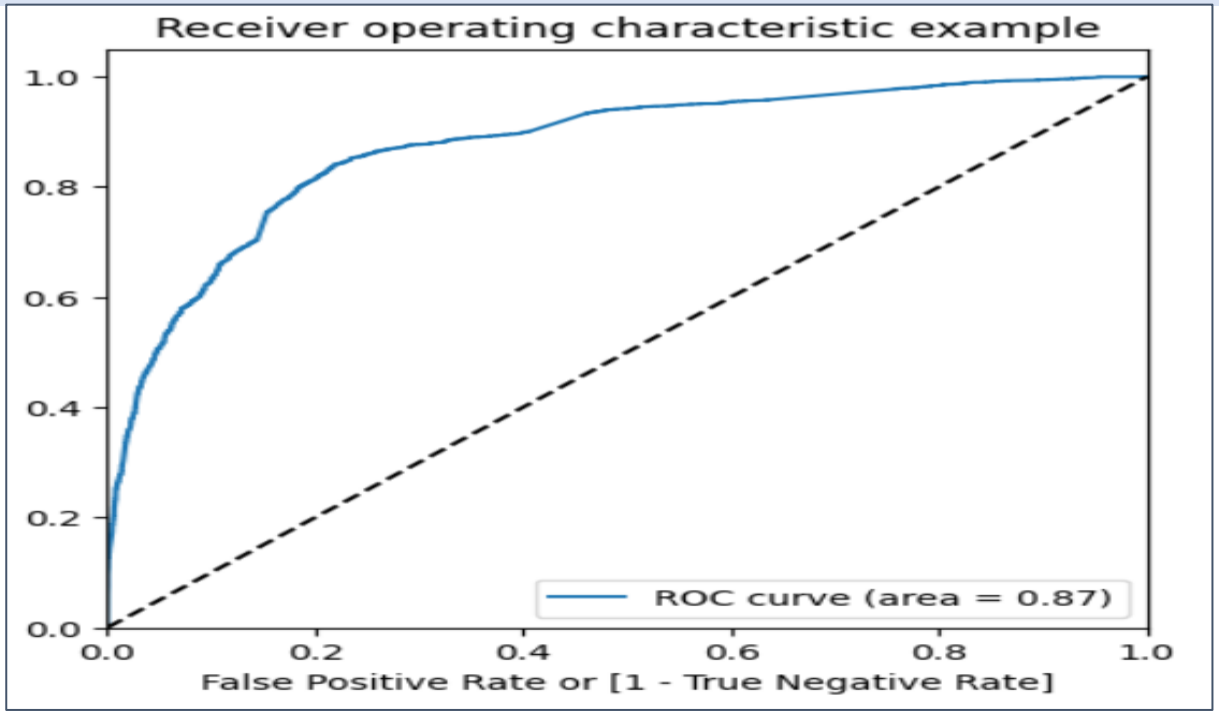


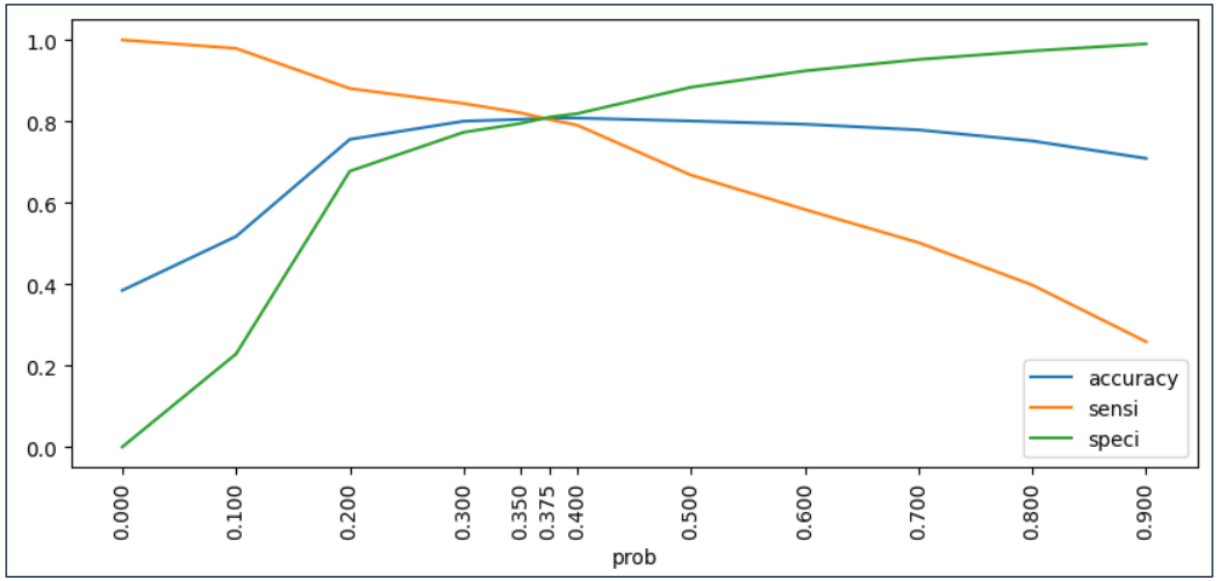


Logistic Regression Model Building, Assessment and Evaluation.

- **Train /Test Data creation**: Train (70%) and Test(30%) data created using train_test_split from sklearn
- **Scaling**: **Min Max Scaling was used to scale continuous variables : ['Total_Visits', 'Website_Time_Spent', 'PVPV']**
- **Model creation and Feature Selection** : Using LogisticRegression module and REF- Automated Feature Selection , selected 15 features for model building
- **Model Assessment**: Used Generalized Linear Models (binomial) from statsmodel library to check the P-values and VIF to eliminate columns that are highly correlated and p-value >0.05
- **Model Evaluation- ROC:** Using ROC (assuming 0.5 as cut off)
- **Model Evaluation-** Used Confusion Matrix to check the trade-off of Accuracy Sensitivity and Specificity metrics and the cut-off point where they all meet(0.36).
- ROC = 0.87 which indicates good model.

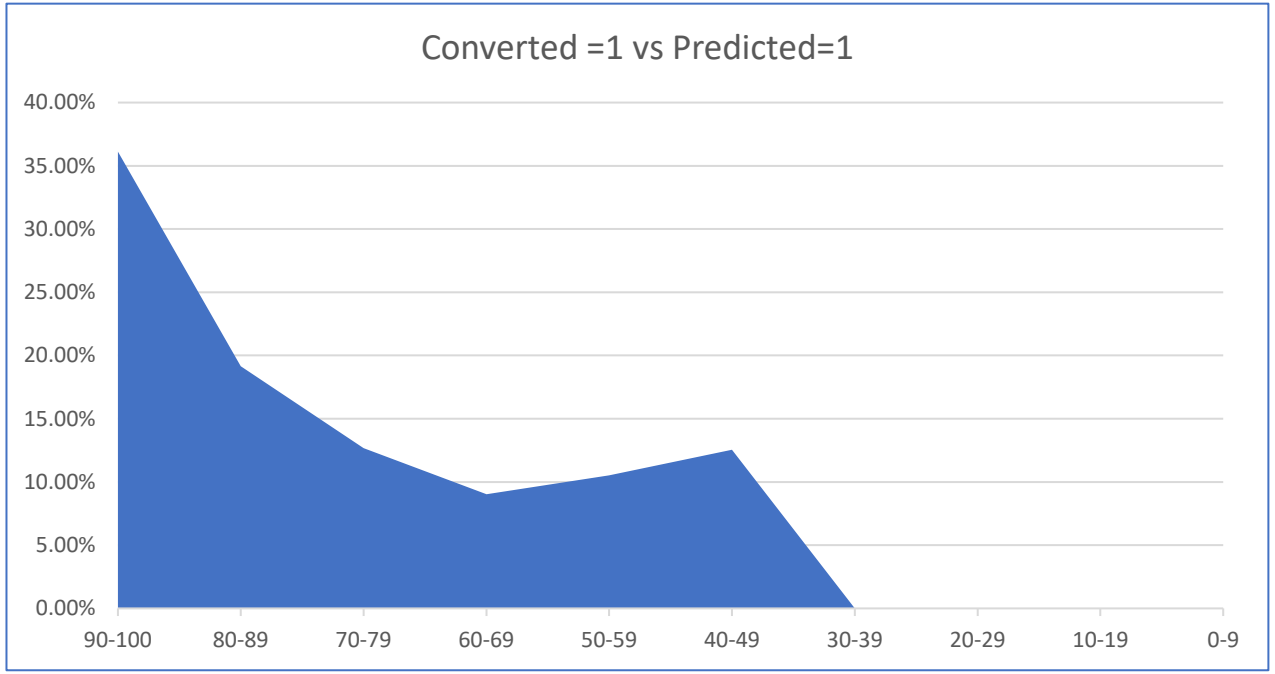
Data	Train Data	Test Data
Metric	Value	Value
Accuracy	0.8054988216810683	0.8035190615835777
Specificity	0.8838396732193005	0.8083867210250437
Sensitivity	0.6678921568627451	0.7952522255192879





Proposed -Hot Lead Scoring, basis the model built.

- **Merged the Test(X) and Predicted (y) by ProspectID column created at runtime using the data frame index to check if the Conversion probability and prediction makes sense.**
- **For instance : 36% of the Predicted =1 falls in 90-100 % of Conversion probability which indicates the prospects within this range can be tagged as "Hot Lead".**
- Basis this understanding the Conversion_probability is binned and proposed a Scoring table is furnished with 10 as better prospect Lead.
- Please refer to dfAnalysis.csv file (please save as Excel and Pivot)
- **Potential leads fall above the cut off predicted by the model (~40%)**



Sum of final_predicted	Converted Vs Predicted	Proposed Scoring
Conversion_Pro	Predicted=1	
95-100	25%	10
90-95	9%	9
85-90	13%	8
80-85	6%	7
75-80	8%	6
70-75	4%	5
65-70	6%	4
60-65	4%	3
55-60	6%	2
50-55	5%	1
45-50	6%	
40-45	7%	
35-40	0%	
30-35	0%	
25-30	0%	
20-25	0%	
15-20	0%	
10-15	0%	
5-10	0%	
0-5	0%	

Proposed -‘Potential Leads’ acquisition strategies basis the model built.

- Taking Conversion probability >40% and Predicted=1, the Categorical variables indicate:
  - Target Working professionals
  - Reach Via Email , giving priority to the prospects who showed interest over the web form on the website.
  - SEO and Content Marketing initiatives could be done, through which more traffic can be attracted to the respective landing pages on which prompt the prospect with web forms to sign up or show interests.

Do_Not_Email	Sum of Total_Visits	Sum of Website_Time_Spent
0	91.74%	93.16%
1	8.26%	6.84%

Categorical Variables(converted=1 and conversion_prob>40%)	Predicted=1
LST_NTB_ACT_Email Opened	203
OCCU_Working Professional	188
LD_ORIGIN_Lead Add Form	161
LST_NTB_ACT_Modified	103
LD_SRC_Welingak Website	37
LST_NTB_ACT_Page Visited on Website	13
LST_NTB_ACT_Email Link Clicked	5
LST_NTB_ACT_Olark Chat Conversation	3
LST_ACT_Converted to Lead	3
LST_ACT_Email Bounced	0

THANK YOU