LEAD SCORE CASE STUDY

PROBLEM STATEMENT

- X Education, an online course provider for industry professionals, faces a challenge with a low lead conversion rate.
- Although they generate numerous leads, only a fraction convert into paying customers.
- To improve this, the company seeks to identify and prioritize 'Hot Leads' with a higher likelihood of conversion.
- The objective is to build a lead scoring model that assigns scores to leads, helping the sales team focus on potential customers.
- The CEO aims for an 80% lead conversion rate. A dataset with 9000 data points, including attributes like Lead Source, Total Time Spent on Website, and Last Activity, is provided.
- The target variable is 'Converted' (1 for converted, 0 for not converted). Handling 'Select' levels in categorical variables is also crucial.

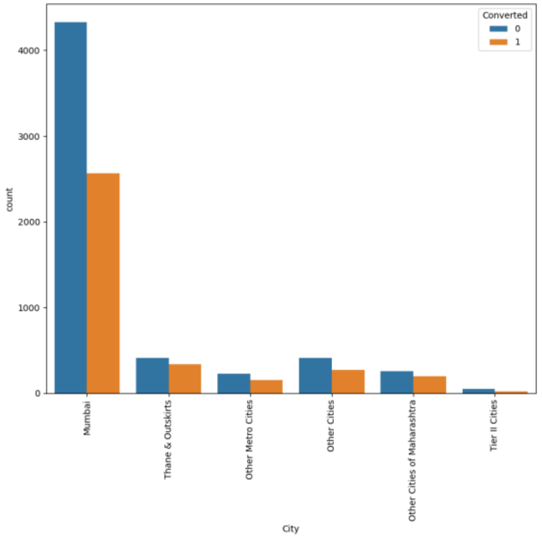
Business Goals

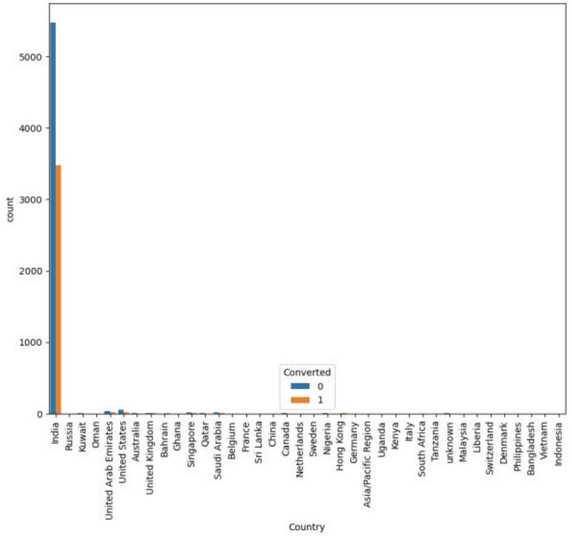
- The primary business goal of this case study is to develop a logistic regression model that assigns a lead score ranging from 0 to 100 to each prospective lead, enabling the company to effectively target potential leads.
- A higher lead score indicates a 'hot' lead with a higher likelihood of conversion, while a lower score signifies a 'cold' lead less likely to convert.
- Additionally, the model should exhibit adaptability to address future changes in the company's requirements and challenges.
- These potential adjustments are documented separately and will be incorporated into the logistic regression model's recommendations, ensuring ongoing effectiveness and relevance in lead conversion strategies.

STRATEGY

- Importing and reading data
- Cleaning and preparing data
- Exploratory data analysis
- Train test spilt and standardizing the data for modelling
- scaling the datamodel building
- Plotting ROC curve
- Test set prediction
- Final observation and recommendations

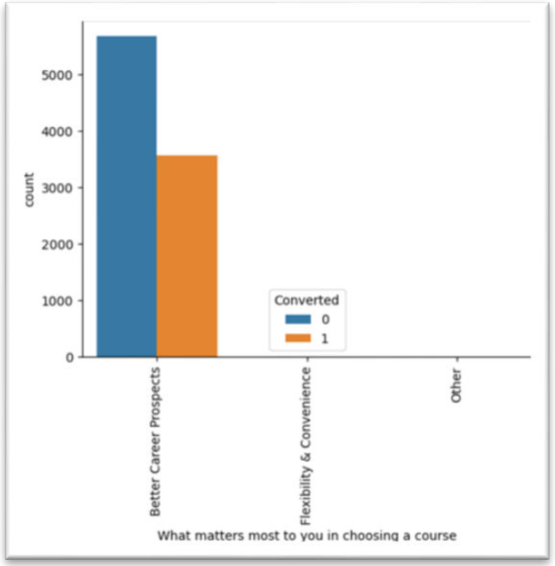
Country and city vs converted

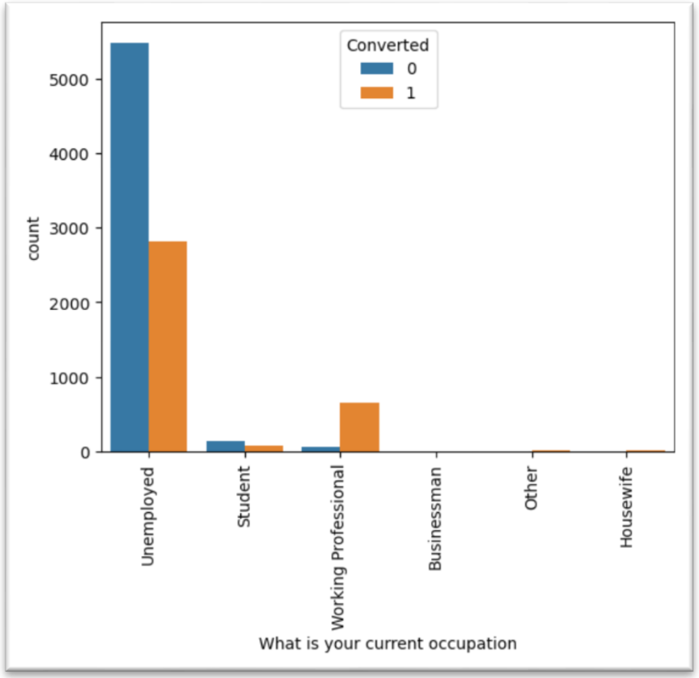




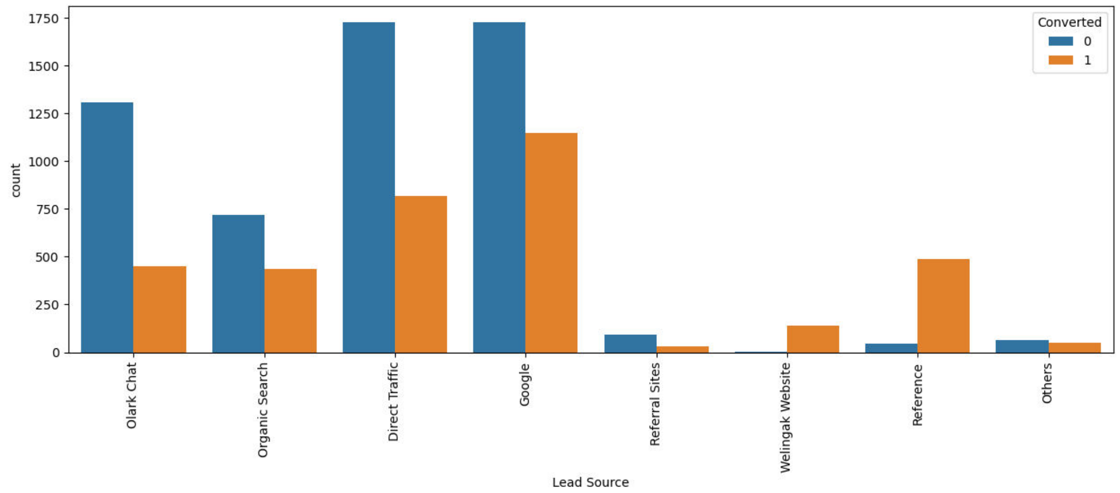
- Most of the conversion are done from india followed by united state and UAE.
- In India highest city conversion are done from Mumbai.

Occupation and What matters most to you in choosing a course



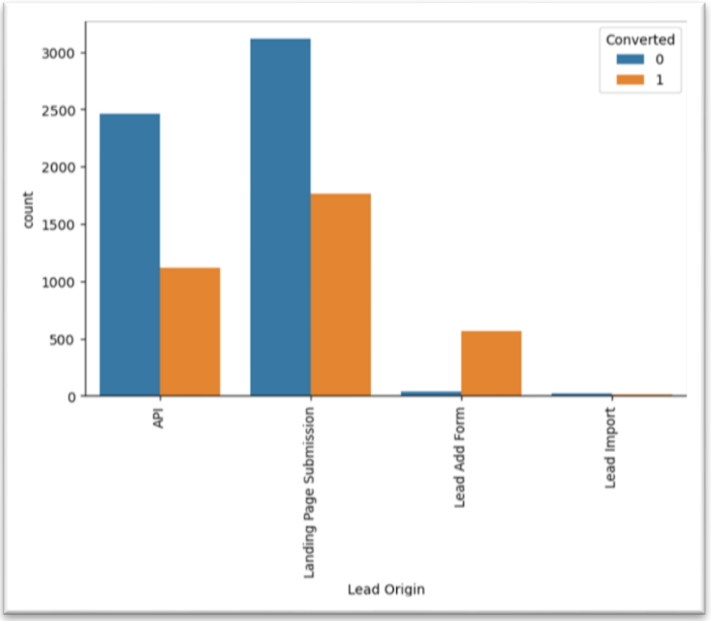


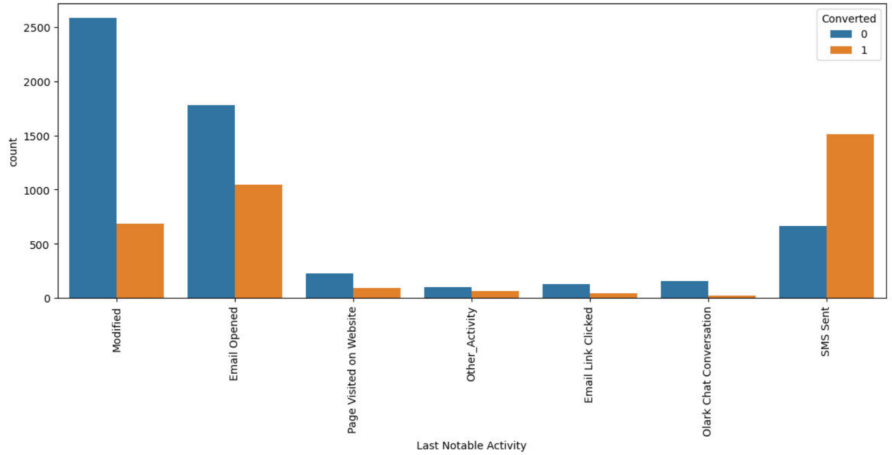
- Most of the conversion are done by unemployed.
- The employed people switch for this course for better career prospect



Lead source and tags

- Maximum number of leads are identified by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, X education should focus on conversion rate of olark chat, organic search as these are generating high number of leads but they are not getting converted



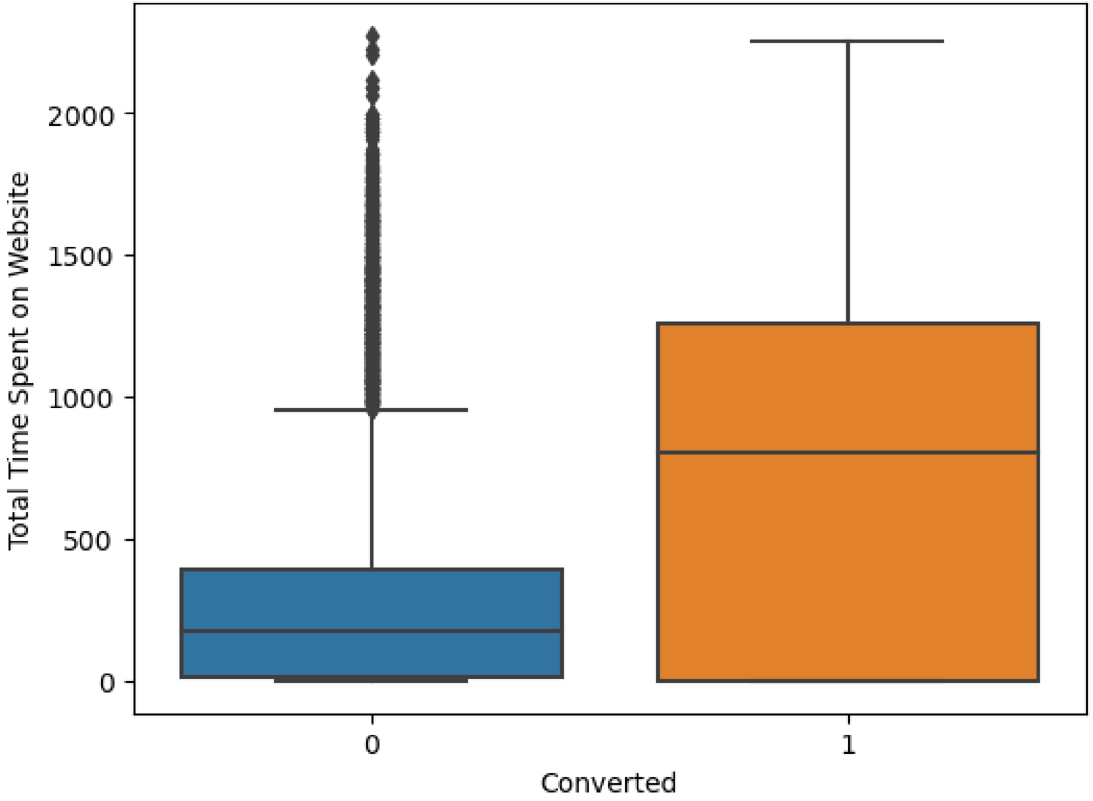


Lead origin converted and last notable activity.

- API and Landing Page Submission generate higher number of leads and many leads do get converted as well.
- Lead Add Form has a very good conversion rate but number of leads generated is not very high.
- Lead Import and Quick Add Form generate very few leads.
- In order to improve overall lead conversion rate, X education should target to improve lead conversion of API and Landing Page Submission origin and should try to generate more leads from Lead Add Form.

- Last notable activity was sms sent.

Total time spent on website vs converted.



- Leads who spend more time on the website are more likely to get converted

Model Building – Preparing The Data

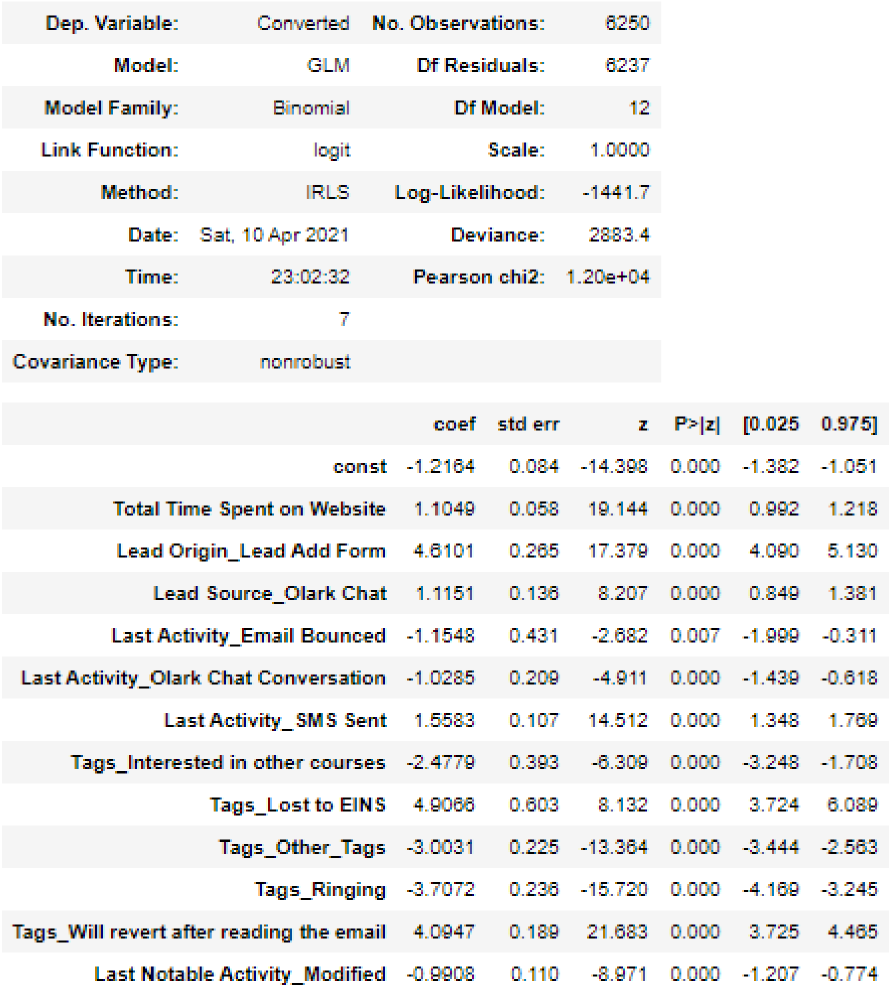
- The data set has been split into Test and Train dataset using the
- train_test_split utility
- The X and y variables for our model are defined as
  - y = Converted (Target Variable)
  - X = All other columns in dataset except Converted
- Data has been scaled using the StandardScaler utility class from
- sklearn.preprocessing module

Model Building

- Model Building has been done using Stats Model & RFE
- First elimination of a few features is done using Recursive Feature Elimination (RFE), and once a small set of variables to work with was obtained, then manual feature elimination (i.e. manually eliminating features based on observing the p- values and VIFs) was used to further fine tune the model
- 15 best features out of 35+ variables were chosen by RFE, which were further fine
- tuned by manual elimination based on the p value(<0.05) and VIF (<5)

Model Evaluation – Train Set Statistics

- Few of the statistics of the model of the train set are as below:
    - Accuracy = 91.02%
    - Sensitivity = 84.89%
    - Specificity = 94.85%
    - False Positive Rate = 5.14%
    - Positive predictive value = 91.15%
    - Negative predictive value = 90.95%

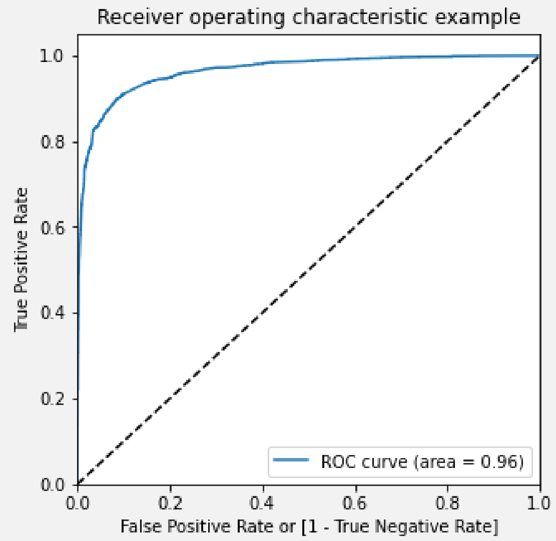


Dep. Variable:	Converted	No. Observations:	6250
Model:	GLM	Df Residuals:	6237
Model Family:	Binomial	Df Model:	12
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1441.7
Date:	Sat, 10 Apr 2021	Deviance:	2883.4
Time:	23:02:32	Pearson chi2:	1.20e+04
No. Iterations:	7		
Covariance Type:	nonrobust		

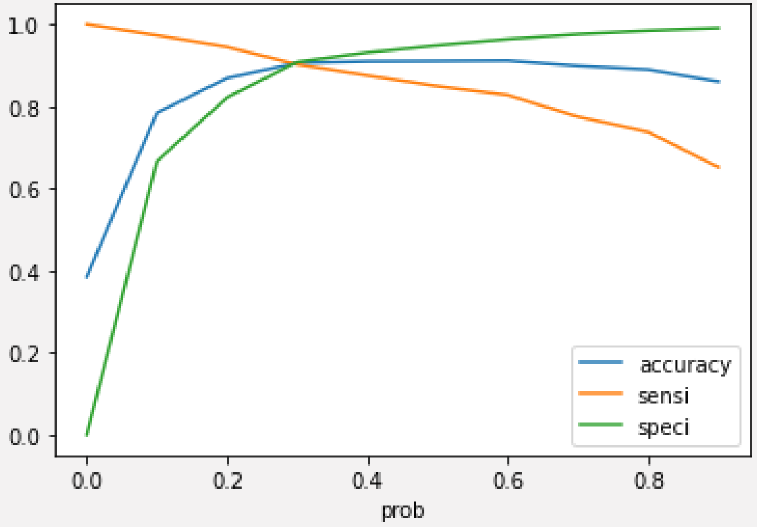
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2164	0.084	-14.398	0.000	-1.382	-1.051
Total Time Spent on Website	1.1049	0.058	19.144	0.000	0.992	1.218
Lead Origin_Lead Add Form	4.6101	0.265	17.379	0.000	4.090	5.130
Lead Source_Olark Chat	1.1151	0.136	8.207	0.000	0.849	1.381
Last Activity_Email Bounced	-1.1548	0.431	-2.682	0.007	-1.999	-0.311
Last Activity_Olark Chat Conversation	-1.0285	0.209	-4.911	0.000	-1.439	-0.618
Last Activity_SMS Sent	1.5583	0.107	14.512	0.000	1.348	1.769
Tags_Interested in other courses	-2.4779	0.393	-6.309	0.000	-3.248	-1.708
Tags_Lost to EINS	4.9066	0.603	8.132	0.000	3.724	6.089
Tags_Other_Tags	-3.0031	0.225	-13.364	0.000	-3.444	-2.563
Tags_Ringing	-3.7072	0.236	-15.720	0.000	-4.169	-3.245
Tags_Will revert after reading the email	4.0947	0.189	21.683	0.000	3.725	4.465
Last Notable Activity_Modified	-0.9908	0.110	-8.971	0.000	-1.207	-0.774

Model Evaluation: ROC Curve & Optimal Cut-Off Point

- Area under the ROC curve = 0.96

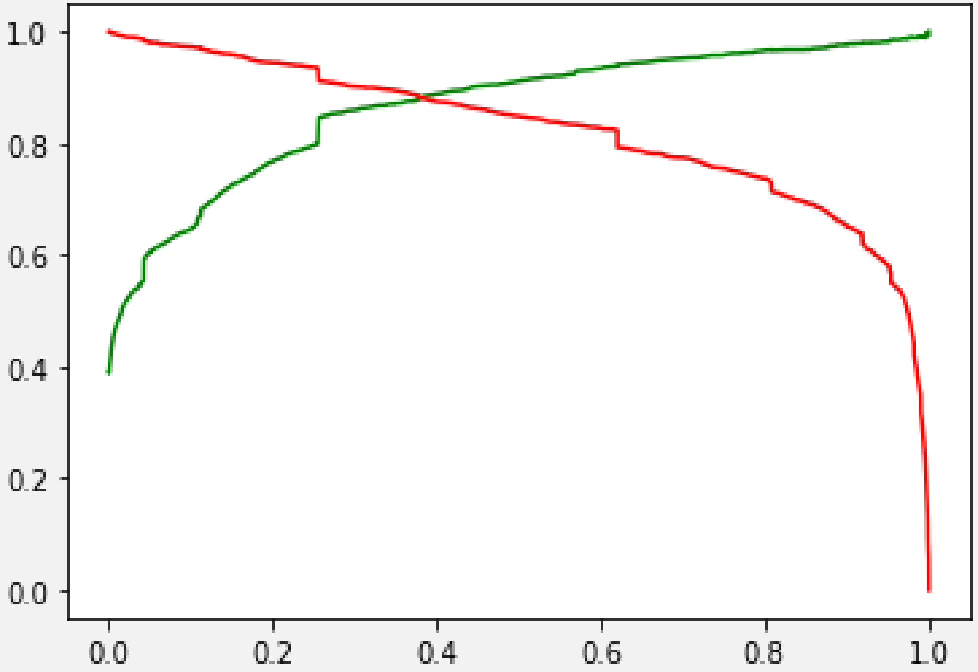


- The optimal cut-off point is at 0.3 based on the accuracy, sensitivity and specificity cross over



Model Evaluation: Train set metric with new cut-off

- Based on new cut-off of 0.3, the new model evaluation stats are as below:
  - Accuracy = 90.56%
  - Sensitivity = 90.17%
  - Specificity = 90.79%
  - Precision = 85.95%
  - Recall = 90.17%
- Using the new cut off, the predicted probability of a lead getting converted was found and then the probability was converted into lead score



Model Evaluation:

- **Test set metrics**
  - Accuracy = 90.29%
  - Sensitivity = 91.64%
  - Specificity = 89.51%
  - Precision = 83.48%
  - Recall = 91.64%

Inferences And Conclusions

	Train Set	Test Set
Accuracy	90.56%	90.29%
Sensitivity	90.17%	91.64%
Specificity	90.79%	89.51%

- The model has good accuracy, sensitivity and specificity for both test and train data set.

- Top three variables that contribute most towards the probability of a lead getting converted are:
    - Tags_Will revert after reading the email
    - Tags_Lost to EINS
    - Lead Origin_Lead Add Form

Inferences And Conclusions

	Train Set	Test Set
Accuracy	90.56%	90.29%
Sensitivity	90.17%	91.64%
Specificity	90.79%	89.51%

- The model has good accuracy, sensitivity and specificity for both test and train data set.

- Top three variables that contribute most towards the probability of a lead getting converted are:
  - Tags_Will revert after reading the email
  - Tags_Lost to EINS
  - Lead Origin_Lead Add Form

Inferences And Conclusions

- We should target those leads which are originating from Lead Add Forms
- We should target leads for which the Tag value is “Will revert after reading
- the email”
- We should target such leads for whom Last Activity is SMS Sent
- We can go for a lower cut-off threshold value so that we can target more
- and more “Hot Leads” and adopt a more aggressive strategy