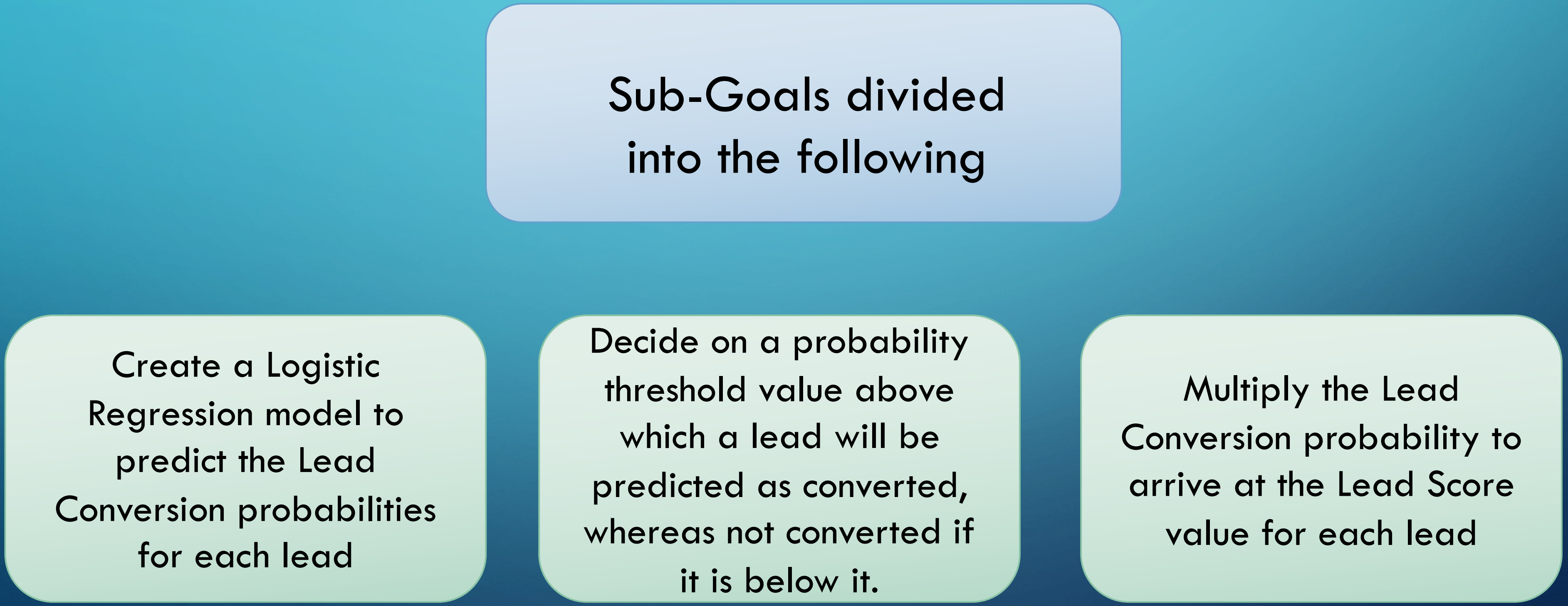
LEAD SCORING CASE STUDY SUBMISSION REPORT

TO CREATE A LOGISTIC REGRESSION MODEL THAT WOULD ALLOW X EDUCATION, A BUSINESS THAT PROVIDES ONLINE LEARNING, TO PREDICT WHETHER A LEAD WOULD BE EFFECTIVELY CONVERTED OR NOT.

Group Members:
M Hemanth Kumar
Ashish
Dhananjay

BUSINESS OBJECTIVE

- helping X Education choose the leads that are most likely to turn into paying customers, or the most promising leads (Hot Leads). To create a logistic regression model that will allow the business to target prospective leads by giving each lead a lead score value between 0 and 100.



Sub-Goals divided into the following

Create a Logistic Regression model to predict the Lead Conversion probabilities for each lead

Decide on a probability threshold value above which a lead will be predicted as converted, whereas not converted if it is below it.

Multiply the Lead Conversion probability to arrive at the Lead Score value for each lead

STEPS FOLLOWED

- Understanding the Data Set & Data Preparation
- Applying Recursive feature elimination to identify the best performing subset of features for building the model.
- Building the model with features selected by RFE. Eliminate all features with high p-values and VIF values and finalize the model
- Use the model for prediction on the test dataset and perform model evaluation for the test set
- Decide on the probability threshold value based on Optimal cutoff point and predict the dependent variable for the training data
- Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc

DATA PREPARATION

The following data preparation steps have been applied

- Remove columns which has only one unique value
- Removing rows where a particular column has high missing values
- Imputing NULL values with Median
- Imputing NULL values with Mode
- Handling 'Select' values in some columns
- Assigning a Unique Category to NULL/SELECT values
- Outlier Treatment
- Binary Encoding
- Dummy Encoding
- Test – Train Split
- Feature Scaling

FEATURE SCALING USING RFE

- Recursive feature elimination is an optimization technique for finding the best performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features.This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated.

BUILDING THE MODEL

- Generalized Linear Models from StatsModels is used to build the Logistic Regression model. •The model is built initially with the 20 variables selected by RFE. •Unwanted features are dropped serially after checking p values (< 0.5) and VIF (< 5) and model is built multiple times. •The final model with 16 features, passes both the significance test and the multi-collinearity test
- A heat map consisting of the final 16 features proves that there is no significant correlation between the independent variables.

FINDING OPTIMAL PROBABILITY THRESHOLD

- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity
- The accuracy sensitivity and specificity was calculated for various values of probability threshold and plotted in the graph to the right.
- From the curve above, 0.33 is found to be the optimum point for cutoff probability.
- At this threshold value, all the 3 metrics - accuracy sensitivity and specificity was found to be well above 80% which is a well acceptable value.

PLOTTING THE ROC CURVE & CALCULATING AUC

- ROC Curve: It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- Area under the Curve (GINI): By determining the Area under the curve (AUC) of the ROC curve,the goodness of the model is determined. Since the ROC curve is more towards the upper-left corner of the graph, it means that the model is very good. The larger the AUC, the better will is the model. • The value of AUC for our model is 0.9678.

EVALUATING THE MODEL ON TRAIN DATASET

Confusion Matrix

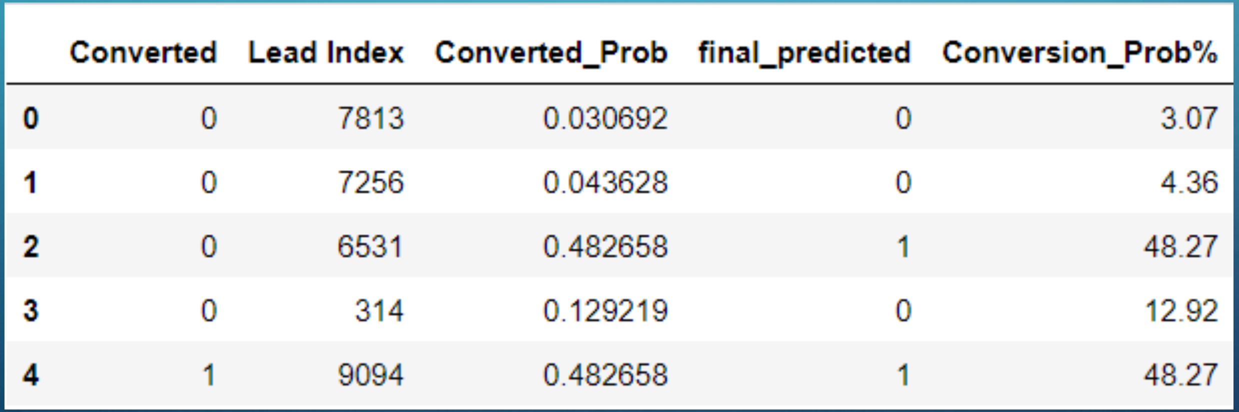
Predicted	Not Converted	Converted
Actual		
Not Converted	3073	432
Converted	215	2211

Probability Threshold = 0.4

Accuracy: 0.89
Sensitivity: 91.14%
Specificity: 87.67%
False Positive Rate: 12.33%
False Negative Rate: 8.86%
Positive Predictive Power: 83.65%
Negative Predictive Power :93.46%

MAKING PREDICTIONS ON THE TEST DATASET

- The final model on the train dataset is used to make predictions for the test dataset •The train data set was scaled using the scaler.transform function that was used to scale the train dataset. •The Predicted probabilities were added to the leads in the test dataframe. •Using the probability threshold value of 0.4, the leads from the test dataset were predicted if they will convert or not.
- The Conversion Matrix was calculated based on the Actual and Predicted 'Converted' columns



	Converted	Lead Index	Converted_Prob	final_predicted	Conversion_Prob%
0	0	7813	0.030692	0	3.07
1	0	7256	0.043628	0	4.36
2	0	6531	0.482658	1	48.27
3	0	314	0.129219	0	12.92
4	1	9094	0.482658	1	48.27

EVALUATING THE MODEL ON TEST DATASET

Accuracy: 89.23%
Sensitivity: 91.56%
Specificity: 87.7%
Precision Score: 82.99%
Recall Score: 91.56%

FINAL FORMULA

- The final formula for this Log Reg model is – ln (p/(1-p)) = -1.9079 + 5.7010 * Tags_Closed by Horizzon + 4.3909 * Lead Source_Welingak Website + 4.3704 * Tags_Lost to EINS + 3.6220 * Tags_Will revert after reading the email + 1.9865 * What is your current occupation_Working Professional + 1.8385 * What is your current occupation_Unemployed - 4.0378 * Tags_Already a student - 3.9105 * Tags_switched off - 3.6396 * Tags_Not doing further education - 3.5416 * Lead Quality_Worst - 3.3832 * Tags_Diploma holder (Not Eligible)-3.3131 * Tags_Ringing - 3.0180 * Tags_Interested in other courses - 2.8539 * Tags_Interested in full time MB