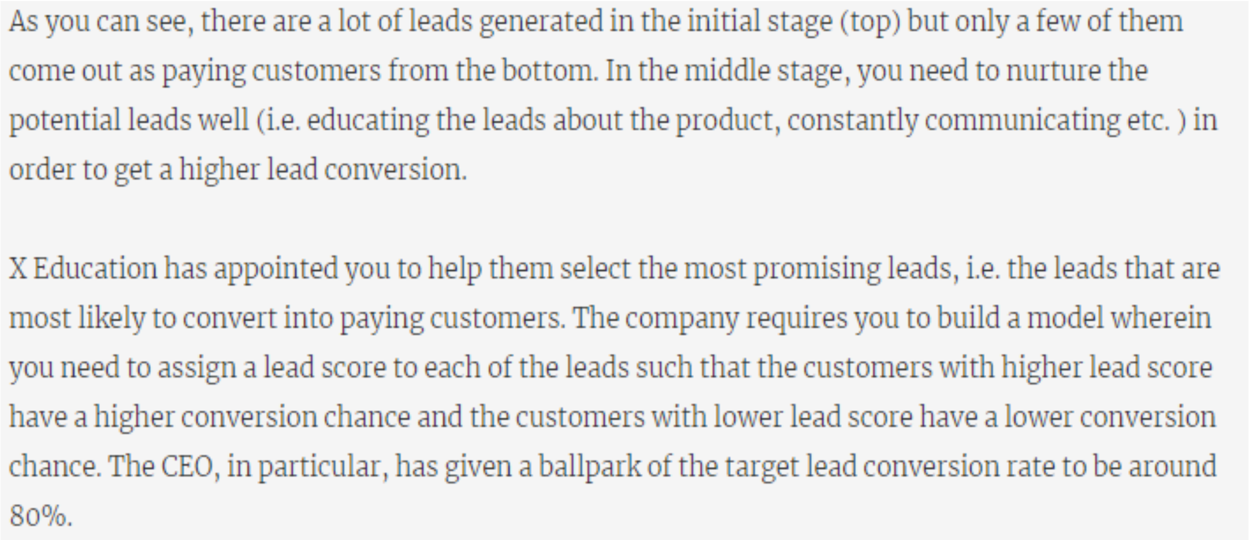
Summary – Lead Score Case Study

Problem Statement:



As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Summary:

Firstly, we started by importing the dataset and understanding the data provided. Understood the dataset after careful examination of the columns, categories of columns, numerical values, we concluded that there are null values in the data and we decided to conduct EDA on the each of the column, which involves:

1. Removed columns, which consists of more than 40% null-values, and are not useful in further analysis.
2. Other variables, which has below 40% null values, are Imputed using calculations mean/median/mode or with 'no-value.

The outlier decision:

By careful examination manually and visually of numeric columns, we have identified the quantile's for detecting outliers as 90% and 99%. That is values less that and greater than 99 % quantiles removed from the data set for further analysis as they identified as the outliers.

Label Encoder- Label Conversion of data:

Secondly, We used 'label encoder' to encode the categorical variables data and made them numeric so that all the variables were in numeric and be used for further analysis.

In addition, we have dropped few columns at this stage as they are not required for analysis.

Summary – Lead Score Case Study

HEAT MAP & Value counts:

To check the correlations of the variables selected, we have plotted a heat map to check the correlations and dropped the variables based on this for further analysis.

More importantly, we have checked the value counts of each column having only one value, which does not contribute anything for modeling perfection. So, removed the same.

Test-Train Split & Scaling:

We have applied test train split on the data and we performed scaling on our dataset to bring all the variables in the dataset to one scale using standard scaler.

At this stage we've decided to make Logistic regression with and without PCA and proceeded further with the model building.

LOGISTIC REGRESSION WITHOUT PCA :

Imported the logistic regression model and instantiated the logistic regression model using RFE 12 columns from 23.After that checked the RFE ranking support columns and took them for model building.

Training set

Performing logistic regression model using the function GLM() (Generalized Linear Models) under stats-model library.

First model building done and executed the Y-predicted values and made a dataframe with conversion probabilities.

Next, Checking the P-values and VIF's we've dropped couple of columns which have high VIF or multi-collinear and settled with good VIF score and concluded the model.

Summary – Lead Score Case Study

Now, created confusion matrix and calculated the Sensitivity, specificity and accuracy of the model.

Plotting the ROC Curve:

An ROC curve shows the trade-off between sensitivity and specificity. We got good space under the curve by plotting the ROC curve and assumed our model is performing well.

After that, we performed the most important step of the analysis, which is PCA on the dataset. The goal of PCA is to capture the maximum variance of the columns, and based on the maximum variance keep the columns for analysis.

Finding optimal cut-off:

Next, we moved on with finding optimal cut-off. By executing the code we got cut offs where we have low, medium and high sensitivities. Plotted the curves between them. We chose high accuracy as we are given an objective where we need the precision to be 80%. To satisfy that condition we have to opt for cut-off that has high accuracy.

Precision and recall trade off:

We also plotted a curve with precision, recall tradeoff, and checked the cutoff value.

Making predictions on TEST data:

Started by applying transform from the scaler and the proceeded with the constant assignment and making predictions on the test data.

We have taken the cut-off from the precision and recall curve plotted and run the model on test set and we have observed slightly reduced accuracy, sensitivity and specificity scores. In addition, this is same with precision and recall.

Furthermore, we have confirmed the metrics sensitivity, specificity and accuracy by making confusion matrix as well as accuracy scores metrics.

Final Result:

Summary – Lead Score Case Study

Finally, we have computed the precision and recall from the test data and achieved the 80% lead conversion rate which is one of the objective CEO specified.

LOGISTIC REGRESSION WITH PCA :

After executing the model without doing PCA on the data we also decided to do it with PCA.

Initiate PCA on the test train split data:

Firstly, we initiated the PCA module from skitlearn library with random state and fit the X train dataset.

Explained variance and Scree plot:

Carefully examined the PCA components and the variance explained by the pca componets.

From scree plot, we analyzed below:

1. Features are contributing equal amount of variance in data
2. That is why, we are having gradual increasing curve
3. Considering, 16 features are showing 90% of data variance.
4. Finally considering 16 features for analysis.

PCA with 16 components

After careful examination we decided to run the PCA with 16 features as they have exhibited 90% of the variance.

In addition, we checked the correlation matrix for any multi-collinearity and found none.

Scalling:

Applied scaling on the training dataset only.

Applying Logistic regression over PCA'd data Train dataset:

Summary – Lead Score Case Study

Net step, applied logistic regression on the PCA’ed train dataset and got the PCA train predicted probabilities.

Made a dataframe out of it and plotted the ROC curve to check the AUC. As per above value, we are concluding, False Positive Rate is very less and True Positive Rate is very high such that we are having a good amount of AUC value as **85.8%**

Finding Optimal Cutoff prob threshold for Train Data values:

Next, we moved on with finding optimal cut-off. By executing the code we got cut offs where we have low, medium and high sensitivities. Plotted the curves between them. We chose high accuracy as we are given an objective where we need the precision to be 80%. To satisfy that condition we have to opt for cut-off that has high accuracy.

Finding precision and recall score:

We have computed the precision and recall score and plotted the curve to check the trade-off between both the curves.

1. Using Accuracy, Sensitivity and Specificity method, we are having 0.365 as probability threshold.
2. Whereas using, Precision & Recall method, we have 0.43 as probability threshold for Conversion analysis
3. **Finally, we are considering 0.365 as the probaility of threshold for further analysis**

Applying Logistic regression over PCA’d data Test dataset:

After applying scaler transform on the test dataset,we apply pca transform on the test dataset and predict the probabilities using the logistic model.

Next, we make a dataframe using the new test probabilities and converted values.and check the accuracy score, precision and recall.

We have taken the cut-off from the precision and recall curve plotted and run the model on test set and we have observed slightly reduced accuracy, sensitivity and specificity scores. In addition, this is same with precision and recall.

Summary – Lead Score Case Study

Final Result:

Finally, we have computed the precision and recall from the PCA’ed test data and achieved the 80% lead conversion rate which is one of the objective CEO specified.