Lead Scoring Case Study

SUMMARY REPORT

Build a model in order to increase the lead conversion rate to around 80% i,e: get in touch with those leads who are more likely to be converted into a customer. In order to achieve the objective, we need to build a logistic regression model which can assign a score from 0 to 100 next to the lead . A higher score means a lead who is most likely to convert or a hot lead. Whereas, a lower score means a cold lead which is less likely to be converted

- Predict a Lead Conversion Probability for each lead
- Decide the cutoff above which a lead will be predicted as converted
- From Lead Conversion Probability calculate Lead Score for each Lead

Steps conducted in this analysis are listed below:-

1. Understanding the data frame by conducting EDA and removed the non-required dataset as well as imputing missing value
2. Split the data into Train & Test set and scale the features
3. Run Logistic Regression Model and use RFE and remove columns with high p-value and VIF
4. Evaluate the model with various metrics like Accuracy, Sensitivity, Specificity, Precision, Recall etc.
5. Find the Optimal Cutoff point and predict the dependent variable based on probability threshold value
6. Use the model on the test dataset and perform the model evaluation

   Originating features of dummy variables are removed

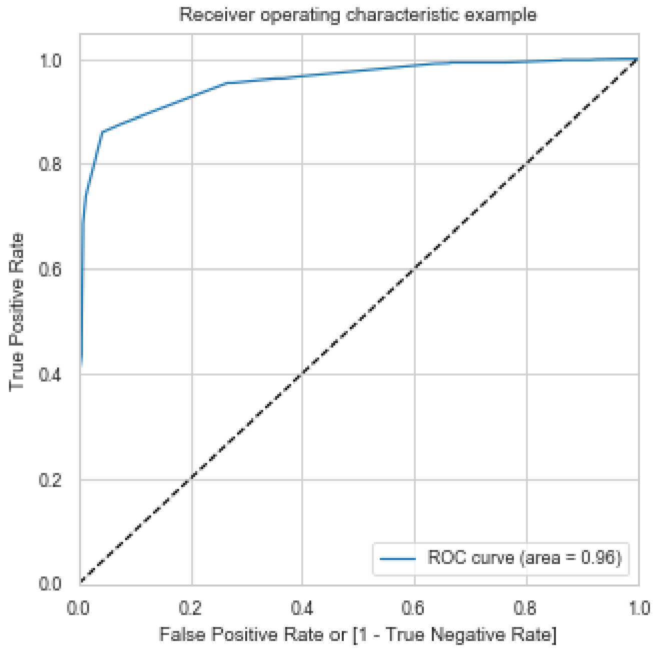
Modeling:

The data was first partitioned into 70-30 split. 70% of the data was used as a training set and 30% of the data was used a validation set. By portioning the data, we were able to see the performance of the model on the unseen data set.

- The dataset is divided into training and test dataset by 70:30 ratio.
- Training dataset is used to build the model whereas the test dataset is used to test the model.
- Scaling is done for all the features to bring all the numeric features into same scale
- Perform Feature Elimination using RFE. RFE is used for 20 features initially and checked the p-value and VIF based on p-value (<0.05) and VIF(<5). Higher p-value features are eliminated from the dataset.

- Find the Optimal threshold and it is required to balance the sensitivity and specificity and hence required a threshold point. Hence, we ran accuracy, sensitivity and specificity for various probability cut-off value to determine the same
- And by using probability threshold value of 0.20 on the test dataset to predict if a lead will convert or not

Receiver Operating Characteristic Curve (ROC Curve)



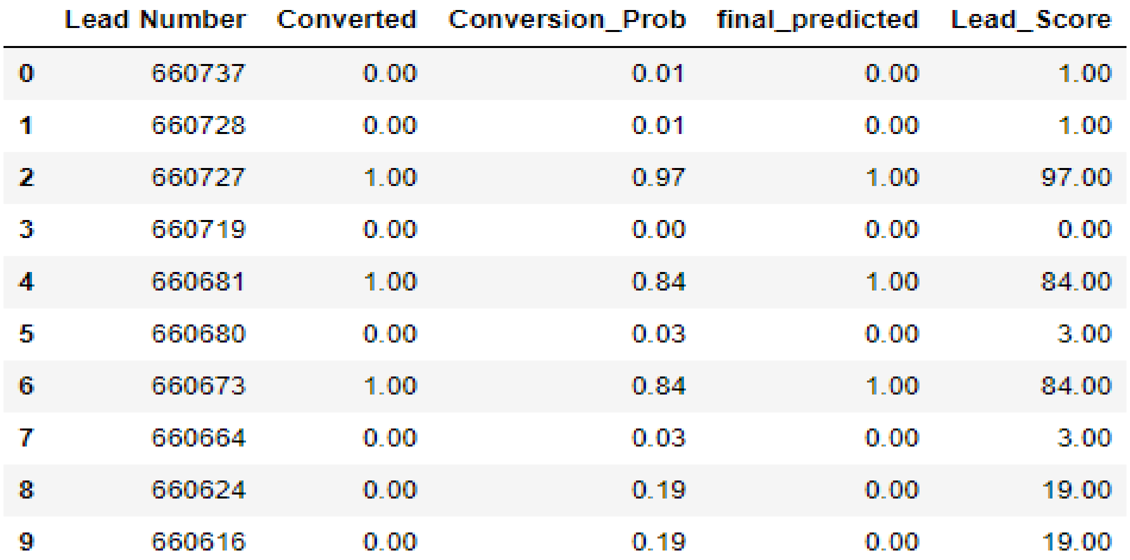
KPIs	value
False Positive Rate **FP/ (TN+FP)**	0.0766
Area Under the Curve**	0.9555

* True Positive Rate value can also found from the formula of sensitivity
**From the area under the curve (AUC) of a ROC curve, one can determine how good the model is. The larger the AUC, the better will be the model.

Lead Score Calculation & Conclusion:

Lead Score Formula: 100*Conversion Probability

- Since, we divided the actual dataset into train and test at the beginning, we append them again to get the entire list of leads
- Conversion probability is multiplied by 100 to get the score
- Higher lead score denotes that the lead is more likely to convert



	Lead Number	Converted	Conversion_Prob	final_predicted	Lead_Score
0	660737	0.00	0.01	0.00	1.00
1	660728	0.00	0.01	0.00	1.00
2	660727	1.00	0.97	1.00	97.00
3	660719	0.00	0.00	0.00	0.00
4	660681	1.00	0.84	1.00	84.00
5	660680	0.00	0.03	0.00	3.00
6	660673	1.00	0.84	1.00	84.00
7	660664	0.00	0.03	0.00	3.00
8	660624	0.00	0.19	0.00	19.00
9	660616	0.00	0.19	0.00	19.00

**Lead score with >=20 will have a final prediction of 1 as we consider the threshold value of 0.20*