Lead Scoring Case Study Summary

The goal of lead scoring case study is to find the leads which are most likely to convert into paying customers. The company required a model, where we assigned 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 goal was to achieve conversion rate around 80%.

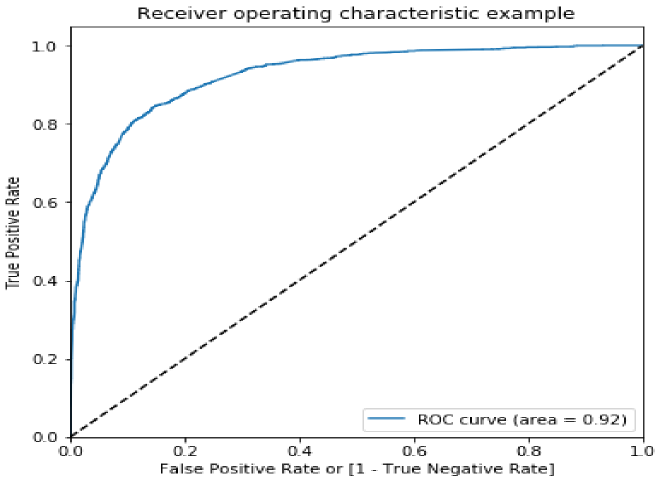
The mathematical tool we have used is Logistic regression model with Recursive Feature Elimination (RFE) method.

Below are the steps followed for required goal:

1. Dropped all the unnecessary columns like country, city, prospect ID etc. as they were found to be not much useful and also columns with higher percentage of missing values were dropped.
2. Imputing missing values: the null values have been replaced either with the mode of the column or the median. the records with less percent of null values, rows have been removed.
3. Univariate analysis for categorical and numerical variables have been done
4. Outlier check and treatment: Inter-Quartile Range (IQR) method is used to remove the outlier present in data for the range (0.05, 0.95).
5. Binary Categorical variables were mapped with 0 and 1 while dummy variables were created for other categorical featured with more than 2 categories.
6. Now the data is ready for modelling and then data was split into training and test dataset.
7. Scaling of the data: Standard Scalar was used to scale the training data for numerical variables.
8. Model building tool: Statsmodels api with RFE (Recursive Feature Elimination) method was used for model building and out of all the features given for training, Top 25 features were selected using RFE.
9. After applying statistics on 25 features using statsmodels api, final 17 significant features were extracted with p-value almost equal to 0 and with no multicollinearity among the features (using VIF method). Thus, we got the final set of more impactful features for lead scoring.
10. Prediction on training data was done with error terms normally distributed.
11. Next, Keeping the threshold probability as 0.5, we calculated the "Converted" flag value from the predicted probabilities. Also, accuracy, sensitivity and specificity using confusion matrix was compared to check the model performance and it came out to be around 77% for sensitivity.

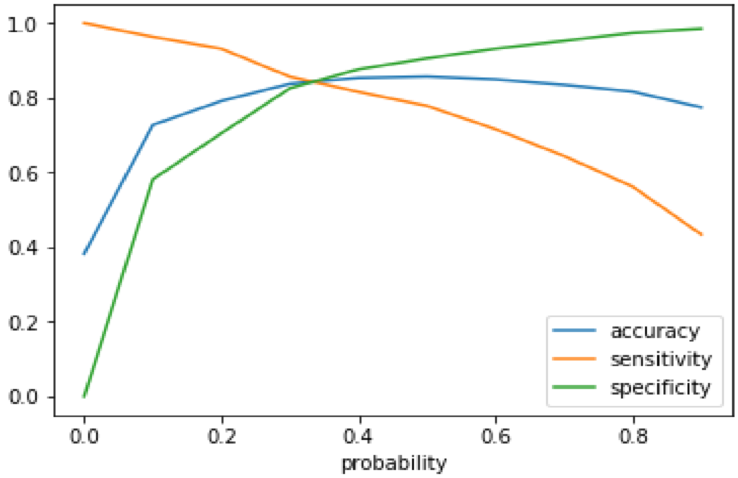
Some important statistical graphs :

ROC Curve:



The Area under curve is around 0.92, which is quite high and a very well representative measure of the model to distinguish between the 2 groups.

12. we built a quiet stable model here. As the CEO expects the conversion rate to be around 80% - Sensitivity around 80% - we found the optimal cut-off point for probabilities using accuracy, sensitivity and specificity line plot as shown below.



The intersection point(0.35) represents where the probability cut-off should be.

13. Now, using the same cut-off of 0.35, we calculated final predicted probabilities again for training data, calculated “Converted” flag value and cross checked the confusion matrix

again to get conversion rate around 80%. This time sensitivity was found to be around 83% as expected.

14. At last, the final model was applied on test data set with sensitivity received as 83% and accuracy score as 84%.
15. Finally, the lead score was calculated for all the leads with range from 1 to 100. (1 being cold lead and 90-100 being hot leads)

16. **Final Conclusion / learning:** After multiple iteration of model-building process, while keeping significance of different variables and multi- collinearity in mind, we concluded that below is the list of original columns/features which impact the conversion rate.
    - Total Time spent on Website
    - Lead origin with add form
    - Lead Source with values
    - Lead Quality
    - Last Activity performed by the user, can be anything like ranging from chat, SMS etc.
    - Current Occupation – Working Professional

Recommendations:

Hence to maximize the chances of converting potential leads into paying customer, company should be focusing more towards Leads activities (chats, SMS etc.), Lead origin, Lead source, Lead occupation, Lead quality and time spent on the website by the Lead. The company should really be focusing into these fields if the CEO wants the conversion rate to be high.