# Summary Report

upGrad
upGrad & IIITB | Data Science Program - January 2023

## Leading Scoring Case Study

**Team Members:** Gandla Akhil, Garima Panjwani , Shaik Ghouse Moin Uddin

## Summary:

1) We began by comprehending the problem statement, desired outcome, and data dictionary.

2) Data understanding and treatment of missing values: Starting with a thorough understanding of the data is crucial to address missing values, as they could otherwise lead to skewed results and a weak model.

- Variables with the level 'Select' were replaced with null values, indicating that no option was selected by the individuals.
- Variables with missing values exceeding 40% were deleted.
- Columns with relatively high missing values and predominantly one unique value were also dropped.
- Columns with very few missing values were imputed using the mode.
- Since the selected specializations are evenly distributed, dropping them is not a suitable choice. Instead, a new category called 'Others' was created to replace the missing values.
-

3) Outlier analysis was performed on the numerical columns 'TotalVisits' and 'Page Views Per Visit.' Both variables were found to contain outliers, and these outliers were treated and removed.

4) Once the data was cleaned, we conducted Univariate and Bivariate Analysis on categorical columns using bar plots, scatter plots, and heatmaps. This analysis provided valuable insights.

5) Data Preparation:
Before implementing Logistic Regression, the data was prepared as follows:

- Dummy variables were created for the categorical columns, and the original categorical variables were dropped to eliminate redundancy.
- The given data was split into training and testing sets, with proportions of 70% and 30%, respectively.
- Feature scaling was performed using Standard Scaler.

6) Model Building:

- For feature selection, Recursive Feature Elimination (RFE) was used to select the top 15 variables.

- The initial Logistic Regression model was built using these 15 variables.
- Insignificant variables with p-values greater than 0.05 were progressively removed, resulting in the final model 4. Model 4 exhibited VIF values below 5, indicating low multicollinearity.

7) Model Evaluation:

- The predicted values for the training dataset were obtained using model 4.
- The confusion matrix was calculated using a threshold probability of 0.5. Based on this, accuracy, sensitivity, and specificity were determined. A low sensitivity and high specificity suggested the need to find the optimal cutoff value.
- A ROC curve was plotted, resulting in an AUC of 0.88, which is considered to be a very good value.
- To find the optimal cutoff, accuracy, sensitivity, and specificity were plotted for various probabilities. The intersection point was found to be at 0.345, indicating the optimal cutoff value.
- The predicted values were obtained again for the training dataset using model 4 and a cutoff probability of 0.345. Lead scores were calculated for each lead in the training dataset, and a confusion matrix was obtained. Based on this, accuracy, sensitivity, and specificity were determined and found to be approximately 80%, indicating a good model.
- Model 4 was applied to the test dataset for predicting the target variable, and lead scores were calculated for each lead. A confusion matrix was obtained for the test dataset, and accuracy, sensitivity, and specificity were found to be near 80%.
- The close values of the evaluation metrics from the training and test datasets suggest that this is a very good model.

# Thank You