Lead Scoring Assignment Summary

- Started with understanding the problem statement and the resources provided which were the dataset and data dictionary.
- Goal of the assignment was to build a Logistic Regression Model and make predictions on the data.
- Imported the dataset into python to understand various dimensions of it such as total number of rows and features.
- EDA was performed on the dataset with the objective of cleaning out the junk data and make the data available for analysis.
- Next step was to prepare the dataset where-in dummy variables were created for categorical variables and binary mapping performed on columns containing flags with two values mainly “Yes’ or “No”
- Dataset was split into Train and Test sets with 70:30 ratio. ’Converted” feature was considered to be the target variable.
- Feature Scaling had to be done on continuous variables so as to bring the features into same scale.
- Correlations amongst the variables had to be checked.
- Initial model was built suing stats model api to understand the metrics
- Arriving at final model by eliminating the features initially using RFE(Recursive Feature Elimination) and later by checking the p-values and VFI. Final model fitted will have the features only having p-values and VFI under control, i.e., p-values < 0.05 and VIF < 5.
- Probabilities have been predicted using the trained model, probabilities are converted to predictions initially using arbitrary cut-off value as 0.5
- Confusion matrix and Accuracy score was calculated
- ROC curve was plotted to check Area under the curve which is the factor for model performance.

- Optimal cutoff value was arrived by plotting the metrics Accuracy, Sensitivity and Specificity against various probabilities between 0 and 1.
- Optimal cut-off was verified by using Precision-Recall tradeoff.
- Final predictions were made on the Test set using the trained model.
- Model evaluation metrics such as Accuracy, Confusion Matrix, Sensitivity, Specificity, Precision and Recall are calculated on test dataset.
- Final trained model showed key metric Accuracy values close in case of both train and test datasets.
- Biz recommendations were prepared based on model fitted.

Key Learnings:

- Understanding the problem statement
- Span across the dataset provided and understand the features using data dictionary.
- Exploratory Data Analysis on the raw dataset.
- Data preparation for building a classification model.
- Knowledge of key libraries to be used for training a Logistic Regression model.
- Splitting the dataset into Train and Test, handling categorical data by creating dummies and dropping original variables, handling continuous variables using Feature Scaling Techniques.
- Fitting a Logistic Regression model.
- Feature Elimination using RFE and by observing p-value and VIF of the features. Arriving at the Final model.
- Make predictions on the trained model.
- Understanding key model evaluation metrics such as Accuracy and Confusion Matrix, also metrics beyond simple accuracy such as Sensitivity, Specificity, Precision, Recall etc.

- ROC curve and Area under the Curve.
- Arriving at Optimal cut-off value using Accuracy, Sensitivity and Specificity metrics also Precision-Recall trade off plots.
- Making predictions on the test data set and evaluating key metrics.
- Providing Biz recommendations based on the model built.