Lead Scoring Case Study Summary

Problem Statement:

An education company selling courses is attempting to reduce their calls made to the leads by the sales team. They need to focus on calls to the potential leads who are more likely to enroll for the course.

The company CEO's expectation is 80% success rate from the calls made.

Goal:

Identify the leads who are more likely, to convert and not convert so that the sales team can make calls appropriately.

Approach:

Logistic Regression approach has been used since the target variable is Categorical. Binomial family from Generalized Linear model is used since the target variable has only 2 values (0 and 1)

Process:

- **Data Reading and Understanding**
- **Data Clean up** – There were lot of null values in the data set. Deletion of columns with more than 33% of null values. Some rows with null values were deleted. Some null values and "Select" were imputed. Finally, we have a data set with 9074 rows and 11 columns.
- **Segmentation** – 2 lists, one with list of numerical columns and one with categorical columns to help in the further processes.
- **EDA and Data Visualization**: Exploratory data analysis was done using various plots which showed us the count, distribution and correlation of the variables.
- **Dummy Variables**: Dummy variables were created since we cannot use categorical values to build the model.
- Target variable and Feature variables were separated into different dataframes.
- **Splitting the data** – Data was split into train and test sets.
- **Scaling**: Numerical variables were scaled to normalize their values using a MinMax scaler.
- **Model Building** – A first Logistic regression model was built with all the variables (82 columns). The features were brought down using RFE

(Recursive Feature Elimination). The correlation with the remaining 25 variables was checked and variables with high correlation were removed. Models were built and features were eliminated till the expected P – values and VIF values were met with the available features.

- Totally **7 models** were built.
- **Prediction** was done on the **train set** and compared with actual data.
- Various metrics such as **accuracy, sensitivity, specificity, precision and recall** were calculated.
- **ROC Curve** and other metrics were plotted and **optimal cut off point** was found.
- Finally, **prediction** in the **test data set** was done and various metrics were calculated.

Learnings:

- Optimal cut off point :0.34

	Train data	**Test data**
Accuracy	77.81	77.85
Sensitivity	84.66	83.75
Specificity	73.61	74.30

- Precision:74.9% Recall:66.3%

Conclusion:

A probability of 0.34 or more for a lead can be considered as a hot lead Working Professionals are more likely to convert. Users who have spent more time on the website and with a greater number of visits to the website are more likely to convert. Leads who have not mentioned their current occupation, who have unsubscribed and whose emails are bouncing are unlikely to convert.

Recommendation:

The sales team should focus on working professionals who visit the website often and spend more time on the website.