Executive Summary

Lead Scoring Case Study

An education company generates many leads as potential customers interact with its marketing content on various platforms. The company seeks to optimize the identification of the most promising leads filtering out the leads which are least likely to convert to paying customers. This project aims to create a Logistic Regression solution to assign scores to all identified leads, giving a higher score to leads which are most likely to convert. The objective: Present conversion rate is 30%, target conversion rate is 80%.

The first step is to prepare the data so it can be a suitable input to the logistic regression model. This involves first handling the missing values and outliers. Therefore, features having more than 20% null values have been removed and outliers in the numeric variables have also been removed. The next step involves encoding binary variables and creating dummy variables for the non-binary categorical variables. Once the data is sufficiently prepared, EDA is performed by plotting various graphs and visualizations. Furthermore, heatmap is created to identify highly correlated variables and these are removed based on their importance gleaned from the bivariate analysis. From EDA we can conclude that certain features would be more important in predicting the conversion rate than others.

The next step is to scale the data, select the features and build the model. Standardization has been used for scaling to preserve the original distribution of the data. For feature selection three methods have been used: RFE, Univariate selection and Correlation matrix. Features which were common from these three selection methods were chosen as the most relevant features and used for model building. Four models were built in total, in each iteration the most insignificant variable (p-value>0.05) was removed. For the final model multicollinearity was also checked by calculating the VIF and found that all VIF values were less than 5.

To evaluate the model, we plotted the ROC curve and precision-recall curve. From these plots different values of threshold was used to find the most optimal value which would give the highest recall score while not sacrificing too much precision. This value was found to be 0.3 and using this to make predictions on the test set yielded the following results:

	Training Metrics	Test Metrics
Accuracy	0.765	0.779
Precision	0.655	0.662
Recall	0.835	0.853

The model is generalizing well on unseen data and has achieved the target identification rate of 80%. We have focused on the recall metric since it is more relevant to our business case. Follwing this the probabilities were assigned as scores to each lead and a recommendation column was added.

Based on the results, certain driver variables were identified as being more impactful to the conversion rate and actionable recommendations were given based on these drivers.