# Summary report

* In this assignment, the primary objective was to develop a predictive model to identify potential leads that are likely to convert for a company engaged in education services. The dataset provided consisted of various features related to leads, including numerical and categorical variables. The assignment involved several key steps, including data exploration, preprocessing, model building, and evaluation. Below is a summary of the approach and key learnings from each stage.
* Data Exploration and Preprocessing:
* The initial step involved exploring the dataset to understand its structure, identify missing values, and gain insights into the distribution of key variables. The dataset was then split into training and testing sets for model development and evaluation.
* Preprocessing steps included handling missing values, scaling numerical features, and creating dummy variables for categorical features. The goal was to ensure that the data is in a suitable format for training a logistic regression model.
* Logistic Regression Model Building:
* Logistic regression was chosen as the predictive modeling technique due to its suitability for binary classification problems, such as lead conversion prediction. The logistic regression model was trained using the training dataset, with the 'Converted' variable as the target variable.
* Feature selection was performed to identify the most significant variables impacting lead conversion. The model's performance was evaluated using metrics such as accuracy, precision, recall, and the area under the ROC curve (AUC-ROC).
* Threshold Optimization, Model Evaluation and Precision-Recall View and Final Model:
* To enhance the model's performance, the threshold for predicting conversions was optimized. The Receiver Operating Characteristic (ROC) curve was plotted, and the Area Under the Curve (AUC) was examined to assess the model's discriminatory power.
* Additionally, sensitivity and specificity trade-offs were analyzed at different probability cutoff points. The optimal cutoff was selected based on a balance between sensitivity and specificity.
* A precision-recall view was explored to gain insights into the trade-off between precision and recall at different probability thresholds. The final logistic regression model was tuned based on this analysis to achieve a balance suitable for the business context.
* Insights and Recommendations:
* Top Variables: The analysis revealed the top variables contributing to lead conversion. Understanding these variables is crucial for targeted marketing efforts.
* Categorical Focus: Certain categorical variables had a significant impact on lead conversion. Focusing on these variables could improve conversion rates.
* Strategies for Different Phases: Tailored strategies were suggested for periods when the company aims for aggressive lead conversion or aims to minimize unnecessary phone calls.
* Model Deployment and Testing:
* The final trained logistic regression model was deployed on the test dataset to make predictions. The model's performance on the test set was evaluated using accuracy, precision, recall, and confusion matrix metrics.
* Key Learnings:
* Feature Importance: Understanding the importance of features in a logistic regression model is crucial for interpreting the factors influencing lead conversion.
* Threshold Optimization: The choice of probability threshold significantly impacts model performance. It involves a trade-off between false positives and false negatives.