Lead Scoring Summary

The primary goal of this logistic regression model is to identify **potential leads** from the dataset based on various customer features. So that sales team could prioritize the potential leads and contact them.

This is the summary of the approach we followed to build the model:

1. Data Understanding :
The dataset includes key customer attributes (e.g., demographics, past interactions, engagement levels) and target variable which is a binary. The dataset was analysed for imbalanced classes (e.g., fewer positive cases than negative cases), which is common in lead conversion datasets.

2. Data Cleaning
We applied EDA process where we dropped columns having **40%** of the null values, replaced **'Select'** value with NaN, computed missing numerical values using **median** and categorical null values replaced with **'Unknown'** . We handled data outliers by dropping values greater than 95% percentile.

3. Univariate Analysis
We performed univariate analysis on Numerical columns, Categorical columns and summarize the observations. Using this, we identified columns having same value for **95%** of the records. We dropped those columns as this does not add value for model prediction.

4. Bivariate Analysis
We performed bivariate analysis on Numerical columns, Categorical columns and summarized the observations. We got to know the hidden patterns or trends against the target variable, we have summarized the analysis in the Notebook.

5. Data Preparation
We learnt some Categorical columns has more values which has low frequency. So we merged those column values under the category '**Others'** , this will reduce the number of columns post creating the dummy variables.

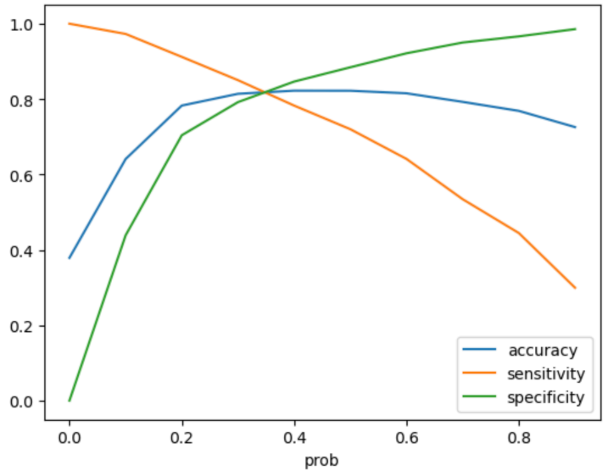
We applied feature scaling (Standard scalar) to scale the numerical columns. Finally we split Train dataset and Test dataset for the model building.

6. Model Building
We build model with below steps:

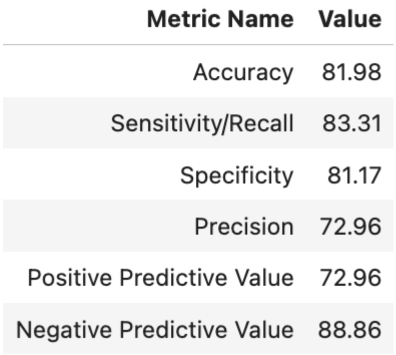
- Selected top features using **RFE** (Recursive Feature Elimination), this returned top columns having feature names under Unknown Category, **ex**: *'city_Unknown', 'course_purpose_Unknown', 'curr_occupation_Unknown'* these columns were not business interpretable and hence we need to drop those columns and rebuilt the model.
- Logistic Regression was selected due to its simplicity, interpretability, and ability to provide probabilities for binary classification. We built the model with 22 features initially. By evaluating the p-value and VIF recursively, we finally arrived at 17 top features for the model.

7. Model Evaluation

We evaluated model using confusion matrix, sensitivity, specificity and accuracy. Sensitivity / Recall seemed to be less compared to Accuracy, so we started finding optimal cut-off using ROC curve, Precision-Recall curve and found **0.34** as optimal cut-off.



Post applying 0.34 as optimal cut-off value, we evaluated the model and calculated the matrix.

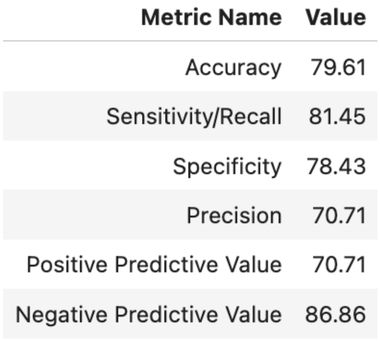


Metric Name	Value
Accuracy	81.98
Sensitivity/Recall	83.31
Specificity	81.17
Precision	72.96
Positive Predictive Value	72.96
Negative Predictive Value	88.86

8. Prediction on Test Dataset

We applied model on the test dataset and predicting the conversion rate. Below is the metrics calculated on the test dataset.

The metrics are almost matching with the Trained dataset. This confirmed that the model is stable.

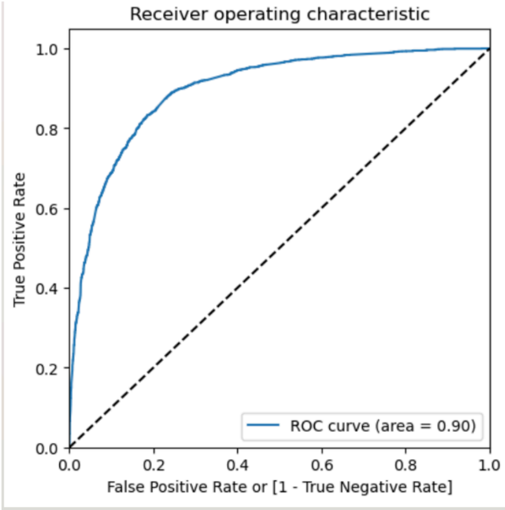


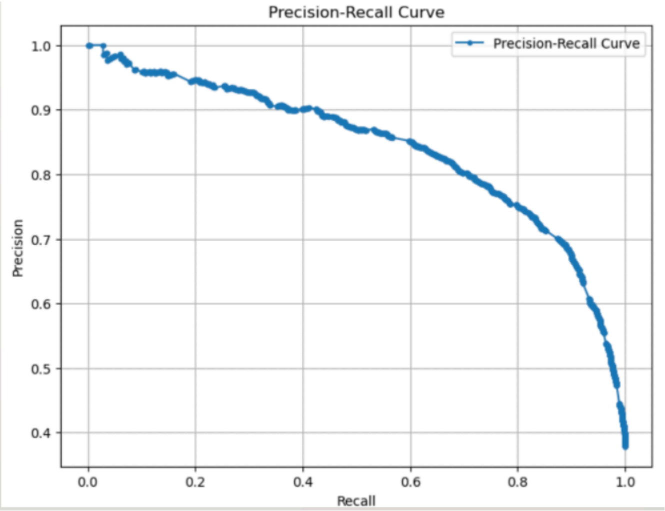
Metric Name	Value
Accuracy	79.61
Sensitivity/Recall	81.45
Specificity	78.43
Precision	70.71
Positive Predictive Value	70.71
Negative Predictive Value	86.86

9. Summary

The primary goal of this logistic regression model is to identify **potential leads** from a dataset based on various customer features. By predicting the probability of a lead converting, the model enabled the sales team to prioritize high-quality leads effectively.

We evaluated the model performance of the model using the ROC curve and Precision-Recall Curve. The AUC for both curves are almost 90%, this indicates that the model is accurate and stable.





10. Recommendations

Below are the actionable recommendations to Sales team to increase the revenue.

- Engage with *Working Professionals* as they need to upskill time on time and there is a high conversion rate.
- Analyze lead behavior who originated from *Lead Add Form* and who *spends more time on website*.
- Understand the *Student's* or *Unemployed* people's background and recommend the course which adds value to them. Offer free counselling with mentors which convinces customer to opt for course.
- Automate *e-mail and SMS subscription* for the leads by sending the course brochure and curriculum. This interests customer to opt for the course.
- Target customers having *Lead Score above 80* by rolling out special discounts or encouraging them to opt for higher end products.