# Summary :

# Lead Scoring Case Study :

# Steps Followed

1. Importing necessary libraries
2. Importing the provided dataset
3. Data Wrangling
4. Exploratory Data Analysis (Variables Inspection)
5. Data Preparation
6. Model Building (Logistic Regression)
7. Model Evaluation (Logistic Regression Metrics)
8. Model Testing
9. Model Inference
10. Conclusion based on our results.

## Data Wrangling:

1. Import dataset
2. Go through the entire dataset and make key observations.
3. Check overall dimensions of the dataset.
4. Check column formats and correct any irregularities found in dataset.
5. Check for any NULL values present in the dataset.
6. Deal with NULL values by imputing those rows or replacing with mean or median values.

## Exploratory Data Analysis (EDA):

1. Data imbalance was checked, and the ratio was found to be 1:1.6 (converted to not converted).

2. Univariate and multivariate categorical analysis was made on all features, and count plots were displayed.
3. Columns with high data imbalance were dropped.
4. Univariate and multivariate numerical analysis was carried out on all numerical columns, and a pair plot and heatmap were plotted.
5. Boxplot analysis was made to handle and treat outliers present.

# Data Preparation:

1. Binary level categorical columns were already mapped to 1 / 0 in previous steps
2. Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source , Do not email, Last Activity, Specialization, Current occupation, Tags ,City , A free copy of Mastering the Interview, Last Notable activity.
3. Splitting Train & Test Sets - 70:30 % ratio was chosen form the split
4. Feature scaling - Standardization method was used to scale the features
5. Checking the correlations - Predictor variables which were highly correlated with each other were dropped.

# Model Building:

1. The dataset has many dimensions and features.
2. This can reduce model performance and increase computation time.
3. **Recursive Feature Elimination** (RFE) is important to select only important columns.
4. Manual Feature Reduction was used by dropping variables with p-value greater than 0.05.
5. Model 3 looks stable after 3 iterations with significant p-values within the threshold (pvalues < 0.05)
6. There is no sign of **multicollinearity** with VIFs less than 5.
7. Model 3 will be the final model used for Model Evaluation and predictions.

# Model Evaluation:

1. The final trained model had an accuracy score of 91%, Precision score of 89%, F1 score of 93%, and ROC curve area of 97% after choosing the optimal cut off at 0.35 from the graph of accuracy, sensitivity, and specificity.
2. lead score was assigned for the trained data.

| Metrics | Scores |
|---|---|
| Accuracy Score | 0.915 |
| F1-Score | 0.931 |
| Precision Score | 0.891 |
| Recall Score | 0.915 |

## Model Testing:

1. The built model was then tested on the test data where we got an accuracy score of 82%, sensitivity of 80%, and an F1 Score of 77%. Hence the model was stable.
2. Lead score was then assigned to the tested data.

# Conclusion:

1. Increased conversions are observed with Landing Page Submissions and Lead Add Form submissions.
2. Leads originating from Google, Organic Search, Direct Traffic, and Referrals exhibit higher conversion rates.
3. Conversions are more frequent among leads generated through SMS and Email marketing efforts.
4. Sectors such as Finance, HR, Marketing, Operations, and Banking demonstrate higher conversion rates.
5. Opting for "Better Career Prospects" as a career outcome choice correlates with increased conversions.
6. Leads spending extended periods on the website show a propensity for conversion.
7. Enhancing website engagement and reducing bounce rates can elevate customer engagement time and, subsequently, conversions.
8. Utilizing the Lead Add Form across strategic channels yields qualifying leads and enhances conversion rates.
9. Targeting working professionals is advised by the sales team for optimizing conversion rates.
10. Leads with a Lead Score exceeding 0.35 tend to exhibit higher conversion rates, supported by a model accuracy score of 91%.