Summary report

Initially, we understood the problem statement and focused on the main objective of this lead score case study. The objective of this study is to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert and provide actionable insights to optimize conversion rates.

Analysis approach:

- To achieve lead score around 80%, we built a logistic regression model after exploratory data analysis (EDA)
- Initially there is a need of data import, inspection and cleaning of data, handling of categorical variables
- Next step is exploratory data analysis (EDA) which includes univariate analysis of both categorical and numerical variables using count plots and hist plots respectively.
- Identification of correlation of numerical variables using heat maps projected the top most three variables helps in conversion of leads to hot leads.
- From the heap map, we can conclude that there is a positive correlation for total visits, total time spent on website and page views per visits
- Dummy variable creation is necessary in regression models when dealing with categorical variables because these models can only process numerical inputs. So all the necessary categorical variables got converted to dummy variables
- Dropping of original categorical columns after dummification of categorical variables is required.
- Building regression models using Scikit-learn and statsmodel libraries is major step for the predicative models results

Model Development:

A logistic regression model was built using the generalized linear model (GLM) framework. Logistic regression was chosen as the data represents binary classification (converted/not converted).

The model included key numerical and dummy-encoded categorical variables. Backward elimination was used to remove variables with high p-values (insignificant predictors).

Model Evaluation:

The model's performance was evaluated using metrics such as accuracy (0.82) and Pseudo R-squared (0.42). The high accuracy indicates that the model effectively predicts lead conversions.

Presenters: Abhishek Goel, Abhishek Dhabal, Bhavani Jagu

Summary report

Coefficients, p-values, and z-scores from the GLM summary were analyzed to determine the impact of each variable on conversion probabilities.

Insights Derived:

Strong predictors like Total Time Spent on Website (positive), Page Views Per Visit (negative), and Lead Add Form emerged.

Categorical variables like Working Professionals, Welingak Website, and Lead Add Form had a significant positive influence, while sources like Direct Traffic and Do Not Email negatively impacted conversions.

Key improvements:

- Invest in improving *Lead Add Form* and *Welingak Website* channels to boost conversions.
- Focus marketing campaigns on *Working Professionals* and optimize email strategies to reduce opt-outs.
- Leverage insights into lead behavior (e.g., website time) for targeted interventions.

Presenters: Abhishek Goel, Abhishek Dhabal, Bhavani Jagu