Summary

This analysis was conducted for X Education, an education company aiming to identify the most potential leads, known as Hot Leads, in an efficient manner. The dataset provided valuable information about professionals interested in the courses, including their site visits, time spent on the site, and referral sources.

The main objective was to build a Logistic Regression model that predicts whether a professional will convert or not, and assign a Lead Score between 0-100 based on their conversion chances. The following steps were undertaken to achieve this goal:

Data Understanding & Feature Selection:

- The initial steps involved loading the dataset, checking its shape, and exploring numeric feature descriptions.
- Redundant features that held a unique value for each customer were dropped, while feature names were standardised for better readability.
- Features with only one or two values, as well as highly skewed features, were removed.
- Null values were identified and features with null values exceeding 30% were dropped. Imputation techniques were applied to handle remaining missing values.
- Collinearity among numeric columns was assessed using a heatmap, and outliers were addressed accordingly.
- Visualisation techniques were employed to assess the relevance of features for model building, leading to the final feature selection.

Data Transformation:

- Binary variables (Yes/No) were converted to numeric values (0/1).
- Features with multiple categories were transformed using one-hot encoding.
- The dataset was split into training and testing sets, with a 70:30 ratio. The training set was used for model building, while the testing set was reserved for evaluation.
- Numeric features were scaled using the Standard Scaler.

Model Building & Evaluation:

- A combination of automated and manual approaches was adopted for model building.
- The Scikit-Learn library was utilised to identify the top 15 relevant features. The model was initially built using these features and further refined by removing insignificant features based on p-values and VIF (Variance Inflation Factor).
- Model evaluation involved obtaining the probabilities for each data point and determining the cutoff for identifying hot leads using the Sensitivity-Specificity approach and ROC curve analysis.
- The final model achieved an accuracy of 81%, with a Sensitivity of 80% and a Specificity of 82%.

- The model was further evaluated using the test dataset, resulting in similar performance metrics.

Conclusion:

- The final model demonstrated good accuracy and stability, without overfitting issues.
- It is considered a reliable model that can adapt to future changes in business requirements.
- The analysis identified key features associated with a higher conversion rate, including features such as "Had a Phone Conversation," "Welingak Website" as the lead source, and "Lead Add Form" as the lead organisation.

Overall, the analysis successfully addressed the business objective of identifying potential leads and provided actionable steps that X Education can implement to meet its requirements.

PDF Converter

Only two pages were converted.

Please **Sign Up** to convert the full document.

www.freepdfconvert.com/membership