SUMMARY

The code model presents a comprehensive analysis aimed at improving lead conversion rates for an online education company, referred to as "X Education." The primary objective of this case study is to develop a logistic regression model that assigns a lead score between 0 and 100, enabling the sales team to identify and prioritize "Hot Leads"—those most likely to convert into paying customers. This initiative is driven by the company's current lead conversion rate of only 30%, highlighting the need for a more efficient targeting strategy.

Data Preparation and Exploration

The initial phase of the analysis involves data preparation, which is crucial for ensuring the accuracy and reliability of the model. The dataset consists of various attributes related to leads, including Lead Source, Total Time Spent on Website, Total Visits, and Last Activity. The first step in the code is to load the dataset and perform exploratory data analysis (EDA) to understand the distribution of features and identify any missing values or anomalies.

Data cleansing is a significant part of the process, where irrelevant columns are removed to streamline the dataset. The code demonstrates the handling of missing values, particularly in categorical variables, where levels such as 'Select' are treated as null values. This meticulous approach to data cleaning results in a data retention rate of 69%, ensuring that the model is built on a robust dataset.

Feature Engineering and Model Building

Following data preparation, the next step involves feature engineering, where the relevant features are selected based on their predictive power regarding lead conversion. The code employs techniques such as one-hot encoding for categorical variables and scaling for numerical features to ensure uniformity across the dataset.

The logistic regression model is then constructed using a 70-30 train-test split. This approach allows for the evaluation of the model's performance on unseen data, which is critical for assessing its generalizability. The model is trained on the training set, and predictions are made on the test set to evaluate its accuracy.

The results indicate that the model achieves an accuracy of 79% with an optimal cutoff threshold of 0.42. This threshold is crucial as it determines the point at which a lead is classified as likely to convert. The code includes detailed evaluation metrics for both training and test datasets, highlighting the model's effectiveness in distinguishing between converted and non-converted leads.

Insights and Conclusions

The analysis concludes with a discussion of the key factors influencing lead conversion. Notably, features such as the total time spent on the website and the source of the lead emerge as significant predictors of conversion likelihood. These insights suggest that X Education can enhance its lead conversion strategy by focusing on leads that exhibit these characteristics.

The structured nature of the code file allows for a clear presentation of each aspect of the study, from problem identification to the conclusions drawn from the analysis. The findings emphasize the importance of data-driven decision-making in optimizing sales strategies. By leveraging the insights gained from the logistic regression model, X Education can significantly improve its chances of converting potential buyers into customers.