

Lead Scoring Model Project

In response to X Education's need for a more efficient lead conversion process, this project aimed to develop a predictive lead scoring model. The objective was to identify high-potential leads and prioritize them for the sales team. The process encompassed exploratory data analysis (EDA), feature selection, addressing multicollinearity, and fine-tuning a decision tree model to achieve a recall of 0.8.

Key steps to summarize:

- Identifying and treating missing values
- EDA to understand top features that could possibly impact the conversion
- Scaling for numerical features
- Dummy variable creation for categorical features
- Creating based model using RFE and VIF for feature selection
- Fine tuned the model using hyper param tuning to arrive at model with 0.8 recall and 0.84 accuracy
- Convert output probability to lead score between 0 and 100
- We can use this model to target new leads to improve conversion

The initial phase involved thorough EDA, employing both univariate and bivariate analyses. Univariate analysis highlighted feature distributions, offering insights into potential outliers and patterns. Bivariate analysis explored relationships between features and the target variable, laying the groundwork for subsequent modeling.

Recursive Feature Elimination (RFE) was employed for feature selection, systematically ranking and eliminating less influential features. This process refined the feature set, enhancing the model's predictive capabilities. Subsequently, Variance Inflation Factor (VIF) analysis was conducted to address multicollinearity concerns. Features with high VIF values were removed to improve model robustness.

A logistic regression model was built using the selected features, providing insights into the relationship between these variables and the probability of lead conversion. Coefficients were interpreted to understand each variable's impact on conversion likelihood. This logistic regression model served as a baseline for further development.

The decision tree model was chosen for its interpretability and flexibility. Iterative fine-tuning of hyperparameters aimed to strike a balance between precision and recall.

The model was refined to achieve a recall of 0.8, ensuring the correct identification of a significant proportion of convertible leads.

Practical implications for the sales team are significant. The final decision tree model with a recall of 0.8 can be implemented as a lead scoring system. By prioritizing leads with higher scores, the sales team can focus efforts on prospects more likely to convert. This targeted approach aligns with the CEO's target lead conversion rate of 80%, offering a practical solution to improve efficiency.