Summary Report: Analysis and Model Creation

Introduction

This case study focused on developing a predictive model for customer conversion for X Education, an online education company. The goal was to create a model that could assign a lead score to each potential customer, predicting their likelihood of converting into a paying student. This summary report outlines the key steps taken and insights gleaned during the analysis.

Data Preparation and Exploration

The initial steps involved comprehensive data preparation and exploration. The dataset was meticulously cleaned, addressing missing values and handling outliers. This ensured data quality and consistency for model building. The dataset contained a mixture of numerical and categorical variables, which were thoroughly examined.

Key insights emerged from the Exploratory Data Analysis (EDA):

- **Lead Origin**: The majority of leads originated from unemployed individuals.
- **Lead Source**: Leads generated through Google and direct traffic were the most common, while the Welingak website exhibited the highest conversion rate.
- **Lead Add Form**: While this lead source demonstrated a high conversion rate, it generated a relatively low volume of leads.
- **Country**: The "Country" column was dominated by "India," potentially introducing bias. This column was removed from the dataset.

Data Transformation

Categorical variables were transformed for model suitability using binary encoding for binary attributes and one-hot encoding for others. Numerical features were standardised to ensure equal contribution to the model.

Feature Selection

The most relevant features were identified using Recursive Feature Elimination (RFE). This iterative process involved ranking features based on their importance within a logistic regression model. Features with high p-values and those contributing to high Variance Inflation Factors (VIF) were systematically removed to mitigate multicollinearity and enhance model interpretability.

Model Building and Evaluation

A Logistic Regression model was built using the selected features. Model performance was evaluated using various metrics:

- **Accuracy**: The model exhibited high accuracy (92.29% on training and 92.78% on testing).
- **Sensitivity**: The model accurately predicted positive cases (91.70% on training and 91.98% on testing).
- **Specificity**: The model effectively identified negative cases (92.65% on training and 93.26% on testing).
- **ROC-AUC**: The ROC curve indicated a good predictive model with an AUC of 0.97.

The confusion matrix further validated the model's effectiveness by providing a detailed breakdown of true and false positives and negatives.

Learnings and Insights

The analysis revealed several key learnings:

- **Data Preprocessing**: Effective handling of missing values and outliers is critical for model accuracy and stability.
- **Feature Selection**: RFE proved a valuable tool for identifying important features, leading to a more interpretable and performant model.
- **Model Evaluation**: Using a combination of metrics like accuracy, sensitivity, specificity, and ROC-AUC provides a comprehensive assessment of model performance.
- **Optimal Cutoff:** Choosing the right cutoff point for classification is crucial for balancing sensitivity and specificity, aligning with business objectives.

Conclusion

This case study successfully demonstrates the power of a robust data analysis process. By combining data preparation, feature selection, and thorough model evaluation, X Education can leverage this model to enhance their lead conversion rate, ultimately driving business growth. Further exploration of advanced algorithms like Random Forests or Gradient Boosting might yield even more insights and enhance model performance.