Lead Scoring Analysis Summary

Project Overview

This project involves building a lead scoring model to predict the likelihood of lead conversion using machine learning techniques.

Key Analysis Steps

1. Data Preprocessing

- Handled missing values by dropping columns with >40% missing data
- Filled remaining missing values with mode (categorical) and median (numerical)
- Converted binary categorical columns to numerical (0/1)
- Created dummy variables for categorical columns
- Combined infrequent categories for lead sources

2. Feature Selection

- Used Chi-square test to identify statistically significant features
- Applied Recursive Feature Elimination (RFE) to select top 20 features
- Performed multicollinearity check using VIF (Variance Inflation Factor)
- Standardized numerical features using StandardScaler

3. Model Building

- Used Logistic Regression as the primary modeling technique
- Split data: 70% training, 30% testing
- Applied statsmodels GLM for detailed statistical analysis

4. Model Optimization

- Performed threshold optimization
- Identified optimal threshold of 0.41 for classification
- Balanced precision-recall tradeoff

5. Model Performance

Final metrics at optimal threshold:

- Accuracy: ~0.86
- Precision: ~0.80
- Recall: ~0.83
- F1-Score: ~0.82

- ROC-AUC: 0.92
- Confusion Matrix:

  [[1492 212]

  [176 892]]

Key Findings

1. The model shows strong predictive performance with an AUC of 0.92
2. The optimal threshold of 0.41 provides the best balance between precision and recall
3. The model demonstrates good discrimination between converted and non-converted leads
4. Feature importance analysis revealed key predictors of lead conversion

Model Validation

- Used confusion matrix to validate predictions
- Performed ROC curve analysis
- Conducted precision-recall tradeoff analysis
- Verified predictions against actual values

Conclusion

The model demonstrates robust performance in predicting lead conversions, making it a reliable tool for lead scoring. The high AUC score of 0.92 indicates excellent discriminative ability between converting and non-converting leads.