SUMMARY

AUTHOR:JAIMINKUMAR

DATE:14/01/2024

Summary

The goal of this analysis was to develop a predictive lead scoring model to identify high value leads using logistic regression. The model was trained on a dataset of over 9,240 leads with 37 features including lead origin, source, demographics, and web activity.

Key findings:

- The final model achieves an accuracy of 82% in predicting converted leads, with good balance between precision and recall.
- The most important predictors of lead conversion are lead origin, source, time on website, and specialization. Landing page submissions and organic search leads have the highest conversion rates.
- Lead scores were generated as probabilities from 0 to 100 for each lead to enable prioritization. The distribution is right skewed with most leads having low scores but a long tail of higher scores.
- Data cleaning involved handling missing values, removing outliers, encoding categorical variables, and feature engineering. This improved model performance.

Data Understanding

The raw dataset contained over 9,240 leads and 37 features. Initial exploration showed high missing value rates for some features like lead quality, tags, and city. These were removed. Other categorical variables were encoded into dummy variables.

Data Preparation

Several steps were taken to prepare the data for modelling:

- Missing values were imputed with median or mode where appropriate. Remaining high missing rate features were removed.
- Outliers in web activity were clipped to 99th percentile to reduce impact on model.
- Categorical variables were encoded as dummy variables to enable use in modelling.
- New feature 'lead origin' was created by consolidating sources with low volumes.

Model Development

A logistic regression model was trained to predict the binary converted/not converted target. This algorithm was chosen due to interpretability and good performance for binary classification.

The dataset was split 70/30 into train and test sets. Input features were normalized before modelling. Hyperparameter tuning was not performed for this prototype model.

Model Evaluation

The final model achieved an accuracy of 0.82 on a held-out test set. Precision and recall were strong for both positive and negative classes, with an F1-score of 0.85 for not converted leads, and 0.76 for converted leads.

The most important features were lead source and origin, time on website, and specialization, aligning with domain knowledge.

Conclusions and Recommendations

The lead scoring model meets the initial business requirements for predicting conversion probability. Further work could explore tuning model hyperparameters, incorporating additional features, and putting into production use for lead prioritization.

Going forward, continuously monitoring model performance and retraining on a regular basis is recommended to account for changing lead dynamics over time. The model can also be extended to predict revenue potential for prioritizing higher value leads.

Overall, this analysis provides a strong starting point for data-driven lead qualification, using proven machine learning techniques. Adoption of lead scoring has the potential to significantly optimize the sales and marketing process.

In summary, the key highlights are:

Achieved 82% accuracy in predicting converted leads.
Lead origin, source, web activity are top predictors.
Generated lead scores from 0-100 for prioritization
Recommend ongoing model evaluation and enhancement.