
Title: Lead Conversion Analysis Using Logistic Regression

Introduction
This analysis aimed to predict the likelihood of lead conversion using historical data. By leveraging logistic regression, the study sought to enable better resource allocation and improve decision-making in lead management.

Methodology
The process began with data understanding and preparation:

- Missing values were addressed using domain-specific imputation techniques (e.g., mode, median, or "Not Available" placeholders).
- Redundant Asymmetrique-related columns were excluded due to excessive nulls and limited value addition.
- Categorical variables were transformed into dummy variables.
- Duplicates were checked and data consistency was ensured.

Exploratory Data Analysis (EDA)
EDA provided critical insights into the data:

1. Visualized the distribution of converted and non-converted leads. The conversion rate stood at approximately **38.5%** (3561 converted out of 9240 leads).
2. Generated a correlation heatmap, highlighting the weak correlation between numerical features and conversion.

Feature Engineering

1. Converted binary columns into numeric form (e.g., Yes = 1, No = 0).
2. One-hot encoded categorical variables for compatibility with the logistic regression model.

Model Building
A logistic regression model was built using features derived from EDA and preprocessing:

- The dataset was split into training and test sets (70:30).
- Hyperparameter tuning optimized model performance.
- Class imbalance was mitigated using weighted loss functions.

Lead Score Calculation

To meet the business objective of assigning a lead score between 0 and 100, the model's predicted probabilities (y_pred_prob) were scaled to this range. This scaling enables the company to target high-potential leads. The final lead score was calculated as follows

lead_scores = y_pred_prob * 100 # Scale probabilities to a range of 0-100

The Lead Score was then added to the test dataset, allowing the company to rank and prioritize leads based on the likelihood of conversion.

Model Evaluation
The model achieved robust performance:

- **Accuracy**: 91%
- **ROC-AUC Score**: 0.958
- **Classification Metrics**:
    - Precision (Converted Leads): 88%
    - Recall (Converted Leads): 89%
    - F1 Score (Converted Leads): 88%

Confusion Matrix:

	Predicted: No	**Predicted: Yes**
Actual: No	1571	133
Actual: Yes	118	950

Key Insights

1. **Lead sources** significantly influenced conversion rates. Investing in high-performing sources could yield better returns.
2. **Occupation** and **specialization** emerged as strong predictors of lead conversion.
3. Features like **total time spent on the website** showed a positive correlation with conversion probability, emphasizing the importance of user engagement.

Business Implications
The model enables organizations to:

- Prioritize leads with high conversion potential, optimizing sales efforts.
- Refine marketing strategies by focusing on effective channels.

- Enhance customer engagement by understanding key conversion drivers.

Learnings
The analysis reinforced the importance of handling missing values and scaling numerical data. The iterative process of feature engineering and model evaluation highlighted the significance of domain knowledge in data science projects.

Conclusion
The logistic regression model provided actionable insights for improving lead management. By scaling the predicted probabilities to a range of 0-100, the model assigns a lead score to each lead, enabling targeted marketing efforts. Future work could explore more advanced algorithms like gradient boosting for potentially higher accuracy. Regular updates with new data will ensure the model remains relevant and accurate.