Lead Score Prediction
- Edtech Company Dataset

Logistic Regression

Swikriti Singhal

Introduction

Problem Statement: Identifying potential leads for conversion in edtech business scenario.

Analysis Approach: A comprehensive data cleaning and analysis pipeline involving missing value imputation, feature engineering, and model building.

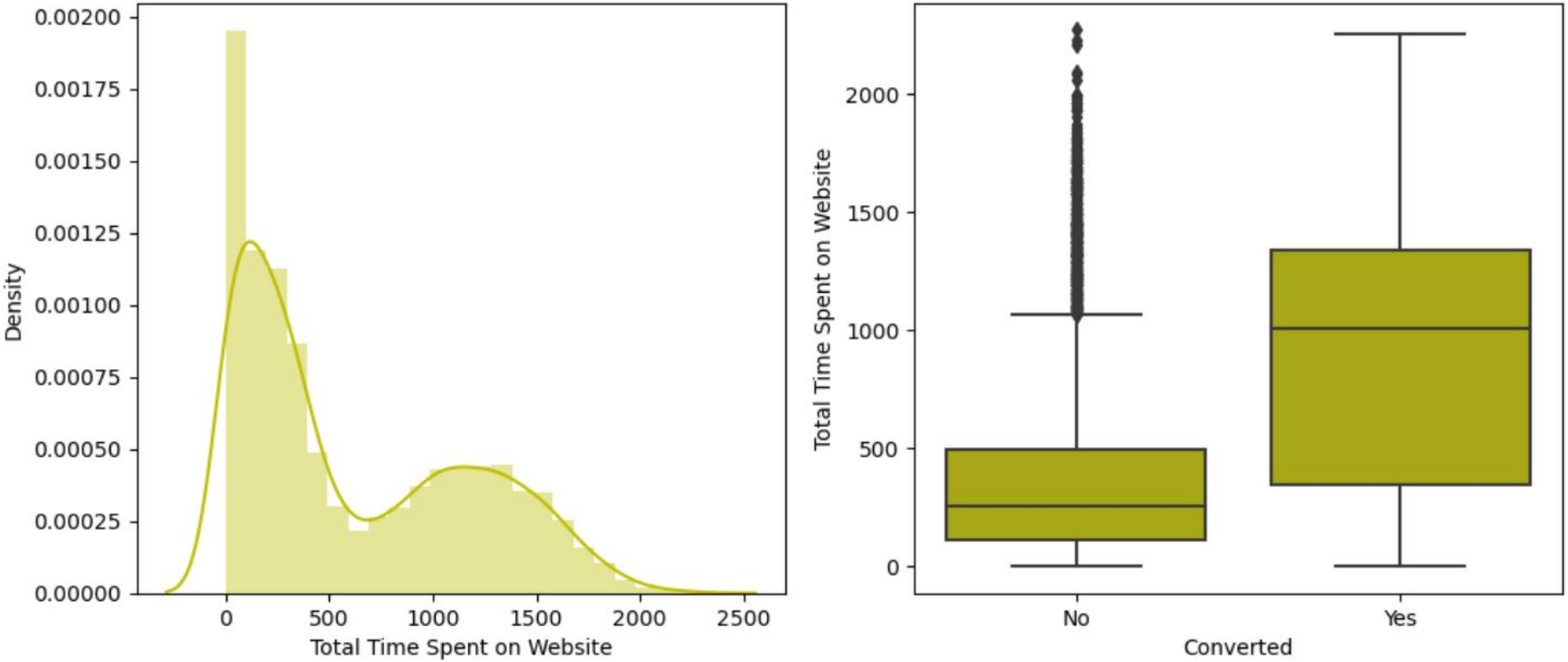
Data Cleaning

Handling Missing Values: Removed features with 40% or more missing values, imputed others using mode.

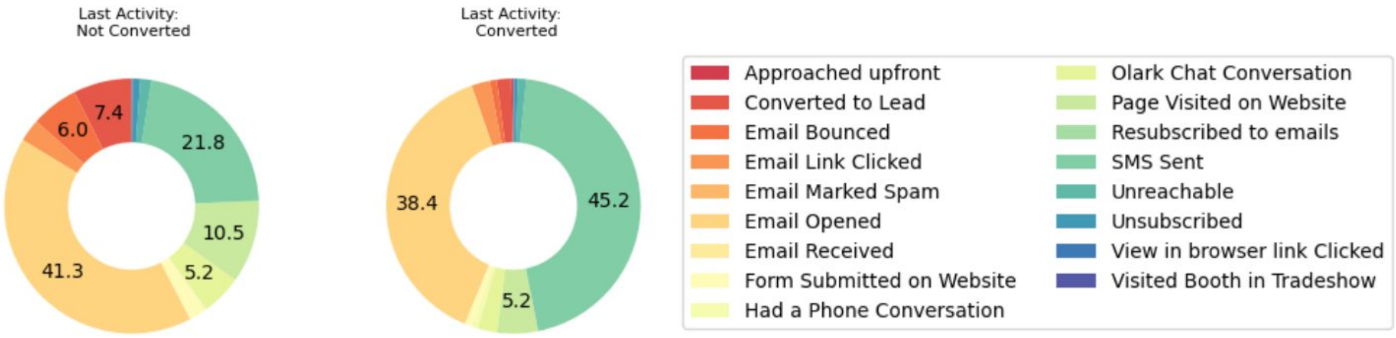
Duplicates Removal: Eliminated duplicate rows introduced during imputation.

Outlier Treatment: Winsorized outliers based on the Interquartile Range (IQR).

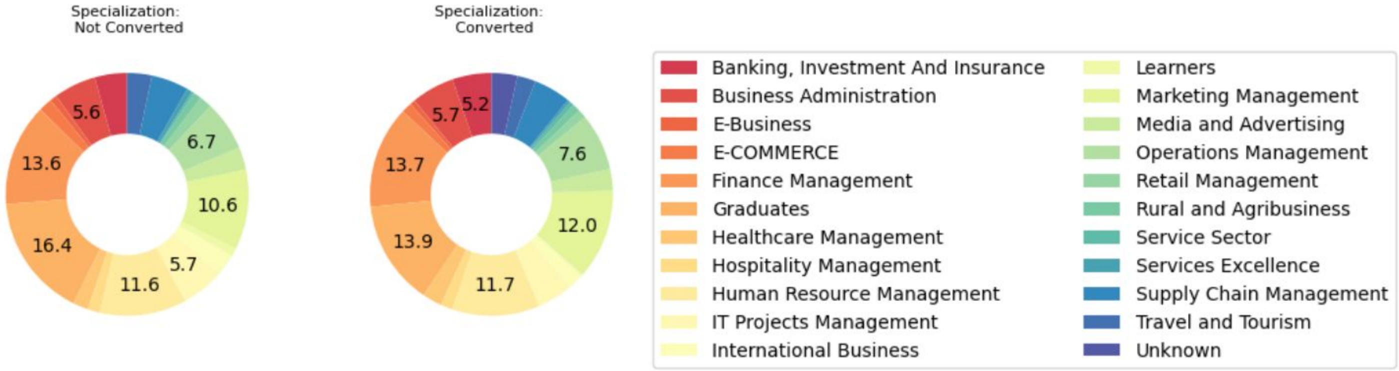
Univariate and Multivariate Analysis: Explored and transformed features.



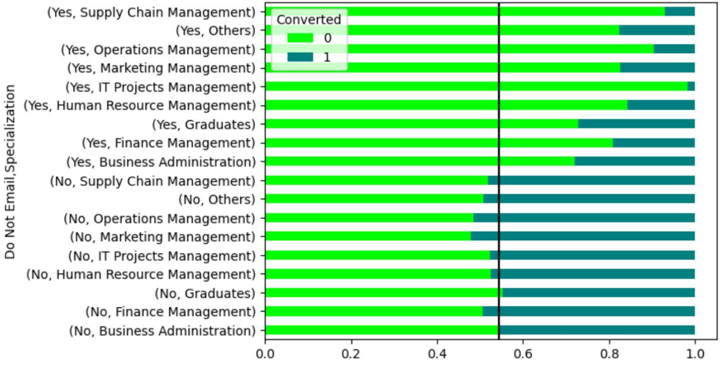
Outlier Winsorized

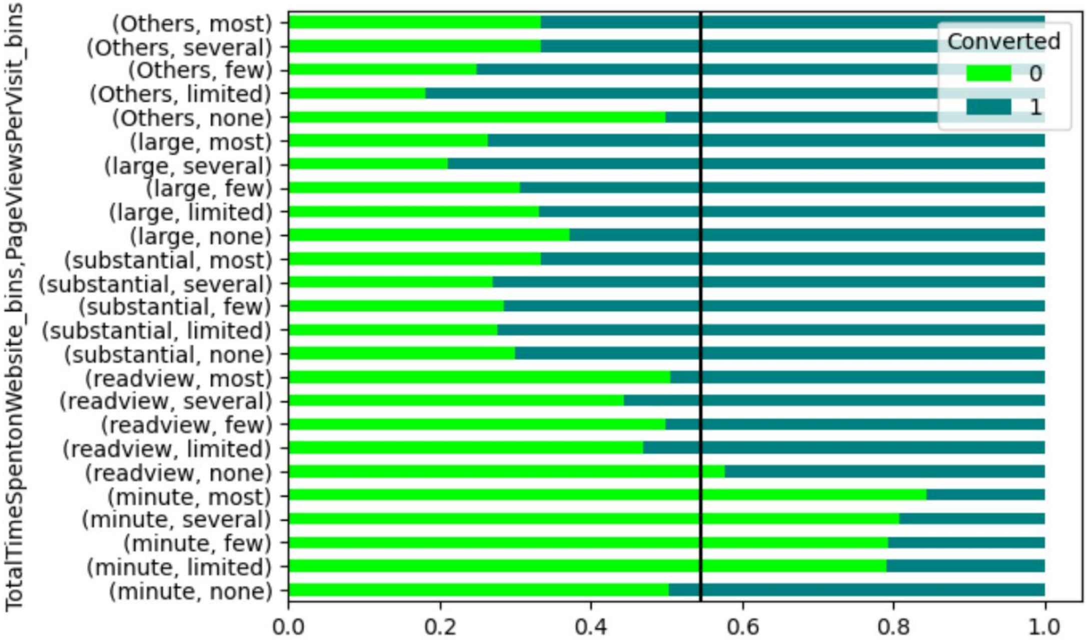


```
<Figure size 900x500 with 0 Axes>
```



Univariate Analysis of Categorical Variables





Multivariate Analysis

Feature Engineering

Binning Numeric Variables: Transformed numeric variables into categorical features.

Handling Mislabeling: Corrected mislabeled entries.

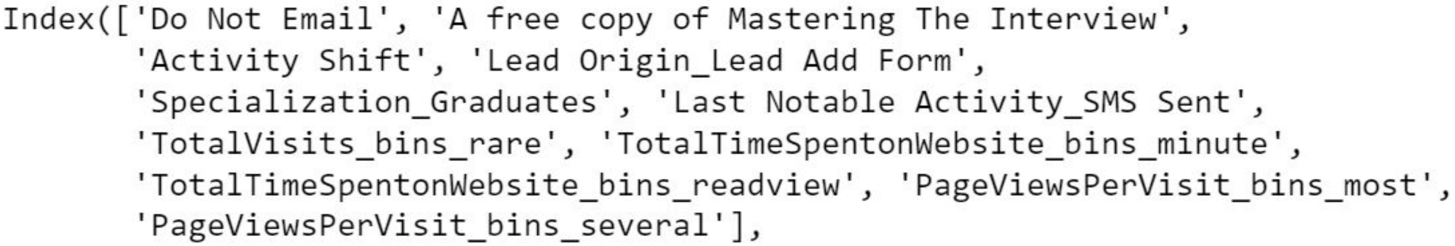
City Labeling for Overseas Accounts: Addressed missing city values for overseas applications.

Stratified Splitting: Ensured a 2:1 ratio in the minority target class in the training set.

Model Building

Encoding: One-hot encoding for non-rank ordered features, label encoding for binary variables.

Correlation Analysis: Selected features based on significance (p-value) and variance inflation factor (VIF).



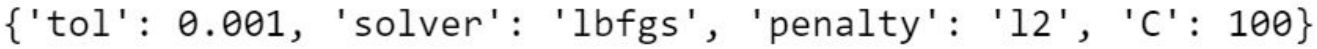
```
Index(['Do Not Email', 'A free copy of Mastering The Interview',
       'Activity Shift', 'Lead Origin_Lead Add Form',
       'Specialization_Graduates', 'Last Notable Activity_SMS Sent',
       'TotalVisits_bins_rare', 'TotalTimeSpentonWebsite_bins_minute',
       'TotalTimeSpentonWebsite_bins_readview', 'PageViewsPerVisit_bins_most',
       'PageViewsPerVisit_bins_several'],
```

Logistic Regression

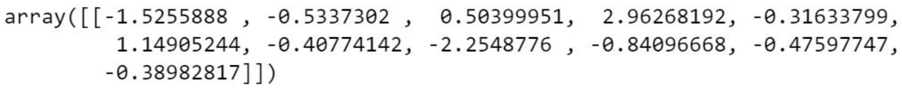
Parameter Tuning: Performed Randomized SearchCV for hyperparameter tuning.

Model Evaluation: Achieved an accuracy of 75.66% on the test data.

Feature Coefficients: Coefficients and corresponding features for the logistic regression model.



```
{'tol': 0.001, 'solver': 'lbfgs', 'penalty': 'l2', 'C': 100}
```



```
array([[-1.5255888 , -0.5337302 ,  0.50399951,  2.96268192, -0.31633799,
         1.14905244, -0.40774142, -2.2548776 , -0.84096668, -0.47597747,
        -0.38982817]])
```

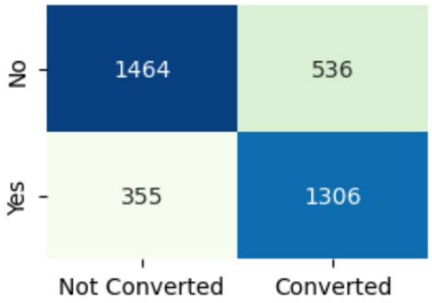
Threshold Tuning

Threshold Tuning: Evaluated different thresholds for binary predictions.

Performance Metrics at Various Thresholds:

Accuracy, Sensitivity, Specificity, ROC AUC, G-Means, Precision AUC.

Identified the optimal threshold for predicting potential leads.



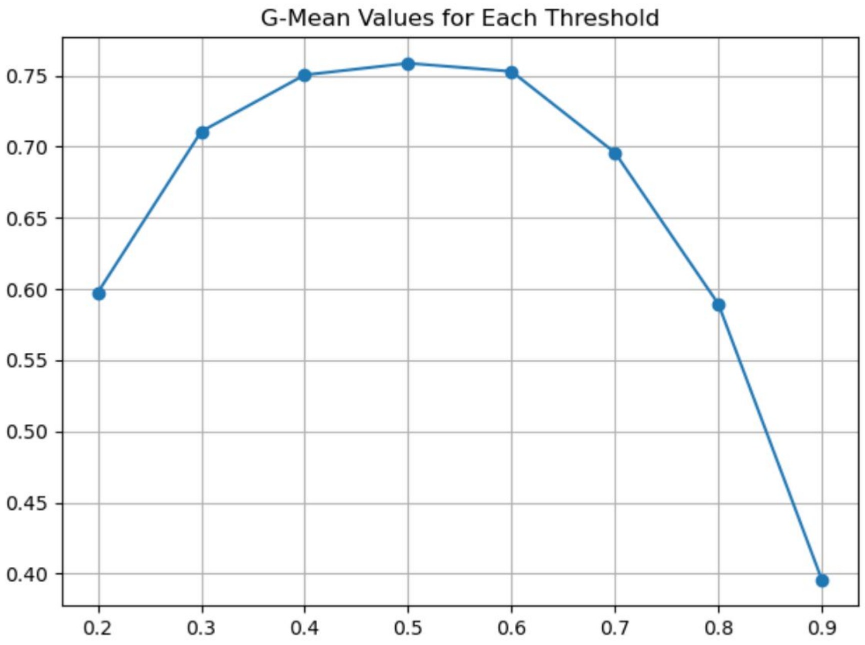
Training Output w/o threshold tuning

Performance Metrics

Threshold Analysis Results:

Optimal threshold of 0.4 for a balanced trade-off.

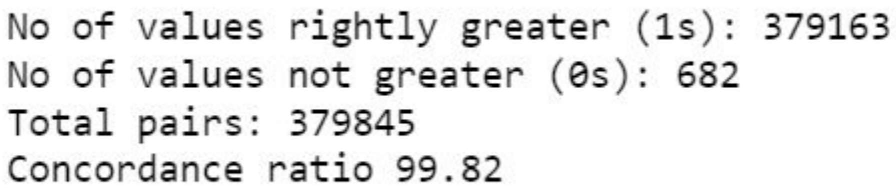
Sensitivity, Specificity, ROC AUC, G-Means, Precision AUC at various thresholds.



Business Implications

Potential Lead Identification: Highlighted the importance of sensitivity in lead identification.

Model Suitability Validation: Concordance ratio, KS Statistic, Gain, and Lift analysis.



```
No of values rightly greater (1s): 379163
No of values not greater (0s): 682
Total pairs: 379845
Concordance ratio 99.82
```

Key Insights

Effective Lead Generation Channels: The analysis highlights that leads generated through the Lead Add Form and those engaged via SMS are more likely to convert. Focusing on these channels can lead to more successful conversions.

Importance of Website Engagement: Total time spent on the website, especially in specific activities like 'Readview,' positively influences conversion. Enhancing website content and user experience can improve lead engagement.

Significance of Last Notable Activity: Last Notable Activity, particularly SMS Sent, plays a crucial role in the conversion process. Continuing and optimizing SMS communication strategies can contribute to higher conversion rates.

Variable Transformations: Binning and transforming numeric variables into categorical ones (e.g., TotalVisits and TotalTimeSpent) proved beneficial in capturing nonlinear relationships and improving model performance.

Recommendation

Lead Nurturing Strategies: Develop personalized lead nurturing strategies, focusing on the preferences identified in the model. Tailor communication methods based on the lead's engagement history and Last Notable Activity.

Optimization of SMS Campaigns: Given the positive impact of SMS on conversions, continually optimize and innovate SMS campaigns. A/B testing different SMS content and timing can uncover more effective engagement strategies.

Enhanced Website Experience: Invest in improving website content and usability to encourage longer engagement periods. Interactive elements and compelling content can enhance the overall user experience.

Next Steps

Continuous Model Monitoring: Regularly monitor the model's performance to ensure its relevance over time. Periodic retraining with updated data can enhance its accuracy and maintain its effectiveness.

Segmentation and Personalization: Explore further segmentation of leads based on characteristics contributing to conversion. Implement personalized communication strategies for different segments to maximize effectiveness.

Feedback Loop Implementation: Establish a feedback loop with the sales team to gather insights on the actual outcomes of leads identified by the model. Incorporate this feedback into model refinement and improvement processes.

Explore Advanced Techniques: Consider exploring advanced machine learning techniques such as ensemble methods or deep learning to capture more intricate patterns in lead behavior and improve prediction accuracy.

Customer Satisfaction Analysis: Integrate customer satisfaction feedback into the model to understand the post-conversion experience. This holistic approach can help in refining strategies for long-term customer retention.