Lead Scoring Analysis for X Education

Optimizing Lead Conversion Rate with Logistic Regression

Pranay Bhal

&

Ishita Gupta

Problem Statement

Goal: Improve the lead conversion rate from the current 30% to 80% by identifying "Hot Leads" (those with high chances of conversion).

Challenges:

•X Education acquires numerous leads, but only a small portion gets converted.

•The aim is to optimize the sales process by focusing on high-potential leads.

Data Overview

- ~9000 leads with various attributes like Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
- Target Variable: 'Converted' (1 = Converted, 0 = Not Converted)
- Handled missing values, duplicates, and categorical variables.

Data Preprocessing

Actions Taken:

- Replaced 'Select' values with NaN and imputed missing values.
- Removed duplicates to ensure data quality.
- Converted categorical variables to dummy variables for modeling.
- Scaled numerical features for better model performance.

Model Building Approach

- **Model Used:** Logistic Regression**Why Logistic Regression?**
  - Suitable for binary classification (Converted or Not Converted).
  - Provides interpretability via feature importance.
- **Hyperparameter Tuning:**
  - Used GridSearchCV to find the best model parameters (C and Solver).
  - Optimized model for accuracy using 5-fold cross-validation

Key Metrics

- **Best Accuracy:** 92%
- **Precision:** 94%
- **Recall:** 95%
- **F1-Score:** 95%
- **AUC-ROC Score:** 0.95

Explanation:

The high precision and recall values indicate that the model is good at identifying hot leads with minimal false positives.

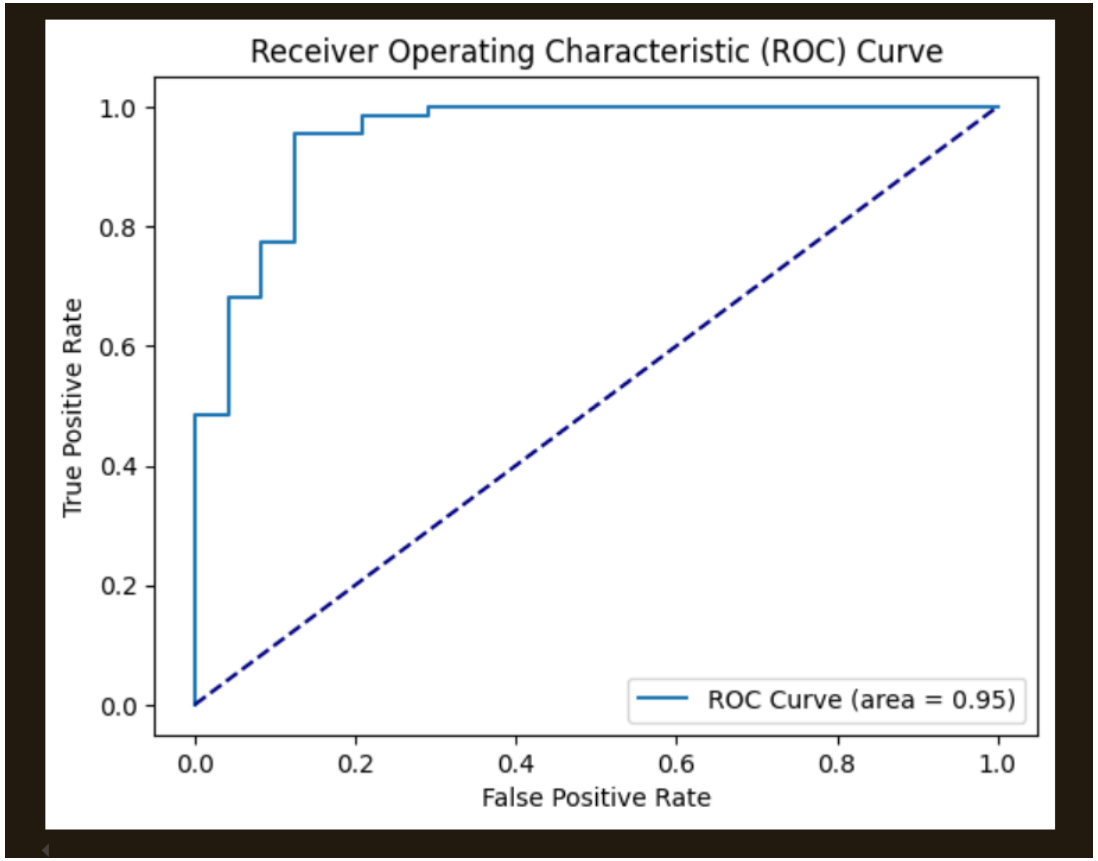
Model Performance

```
Best Accuracy: 0.92
Confusion Matrix:
[[20  4]
 [ 3 63]]
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.83      0.85        24
           1       0.94      0.95      0.95        66

    accuracy                           0.92        90
   macro avg       0.90      0.89      0.90        90
weighted avg       0.92      0.92      0.92        90

AUC-ROC Score: 0.95
```



CODES FOR MODEL BUILDING

```
# Confusion matrix visualization
conf_matrix = confusion_matrix(y_test, y_pred_best)
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues", cbar=False)
plt.title("Confusion Matrix")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()
```

ROC Curve Code

```
fpr, tpr, thresholds = roc_curve(y_test, grid_search.best_estimator_.predict_proba(X_test)[:,1])
plt.plot(fpr, tpr, label=f"ROC Curve (AUC = {roc_auc:.2f})")
plt.plot([0, 1], [0, 1], color='navy', linestyle='--')
plt.title("Receiver Operating Characteristic (ROC) Curve")
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.legend(loc="lower right")
plt.show()
```

```python
# Import necessary metrics at the beginning of your code
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score

# Evaluate the best model's performance

# Accuracy
accuracy = accuracy_score(y_test, y_pred_best)
print(f"Accuracy: {accuracy:.2f}")

# Precision
precision = precision_score(y_test, y_pred_best)
print(f"Precision: {precision:.2f}")

# Recall
recall = recall_score(y_test, y_pred_best)
print(f"Recall: {recall:.2f}")

# F1-Score
f1 = f1_score(y_test, y_pred_best)
print(f"F1-Score: {f1:.2f}")

# AUC-ROC score
roc_auc = roc_auc_score(y_test, grid_search.best_estimator_.predict_proba(X_test)[:, 1])
print(f"AUC-ROC Score: {roc_auc:.2f}")
```

Top Features Influencing Conversion

- **Total Time Spent on Website:** Higher engagement on the website correlates with higher conversion.

- **Lead Source – Google:** Leads coming from Google have a higher likelihood of conversion.

- **Last Activity – Email Opened:** Indicates strong interest and potential for conversion.

Categorical Variables to Focus On

- **Lead Source (Google, Direct Traffic):** Optimize marketing efforts on these sources.

- **Last Activity (Email Opened, SMS Sent):** Focus on leads who engage via emails and SMS.

- **Lead Quality (High, Medium):** Segment and prioritize leads based on quality rating.

Aggressive Conversion Strategy (During Intern Period)

Goal: Maximize lead conversion during periods with extra resources (interns).

Strategy:

- Focus on leads with predicted conversion probabilities near 1
- Prioritize leads from high-converting sources (e.g., Google, Email Opened).
- Increase personalized communication efforts (calls, emails, SMS) for these hot leads

Low Activity Period Strategy

Goal: Minimize unnecessary calls during periods when the target is already achieved.

Strategy:

•Focus only on leads with high confidence in conversion (predicted conversion probabilities close to 1).

•Reduce outreach to leads with low probability scores.

•Shift resources to other productive activities, such as new lead generation or customer retention strategies.

Recommendations

For Business:

•Invest in marketing efforts targeting high-conversion lead sources.

•Focus on improving website engagement (Total Time Spent) as it is a strong indicator of lead conversion.

•Continue using logistic regression with periodic retraining to adjust to market dynamics.

For Sales Team:

•Prioritize communication with leads predicted to convert, especially from Google and Direct Traffic.

•Use insights from the model to focus on high-quality leads and optimize communication strategies

Summary

- **Outcome:** By implementing the logistic regression model, X Education can improve its lead conversion process by targeting potential hot leads more effectively.

- **Next Steps:** Deploy the model and implement the proposed strategies during high and low sales activity periods