Lead Scoring Case Study

Presentation

- CHANDANA DEEKSHA PULAVARTHI
- SHWETA KUMARI
- SOUMIK CHAKRABORTY

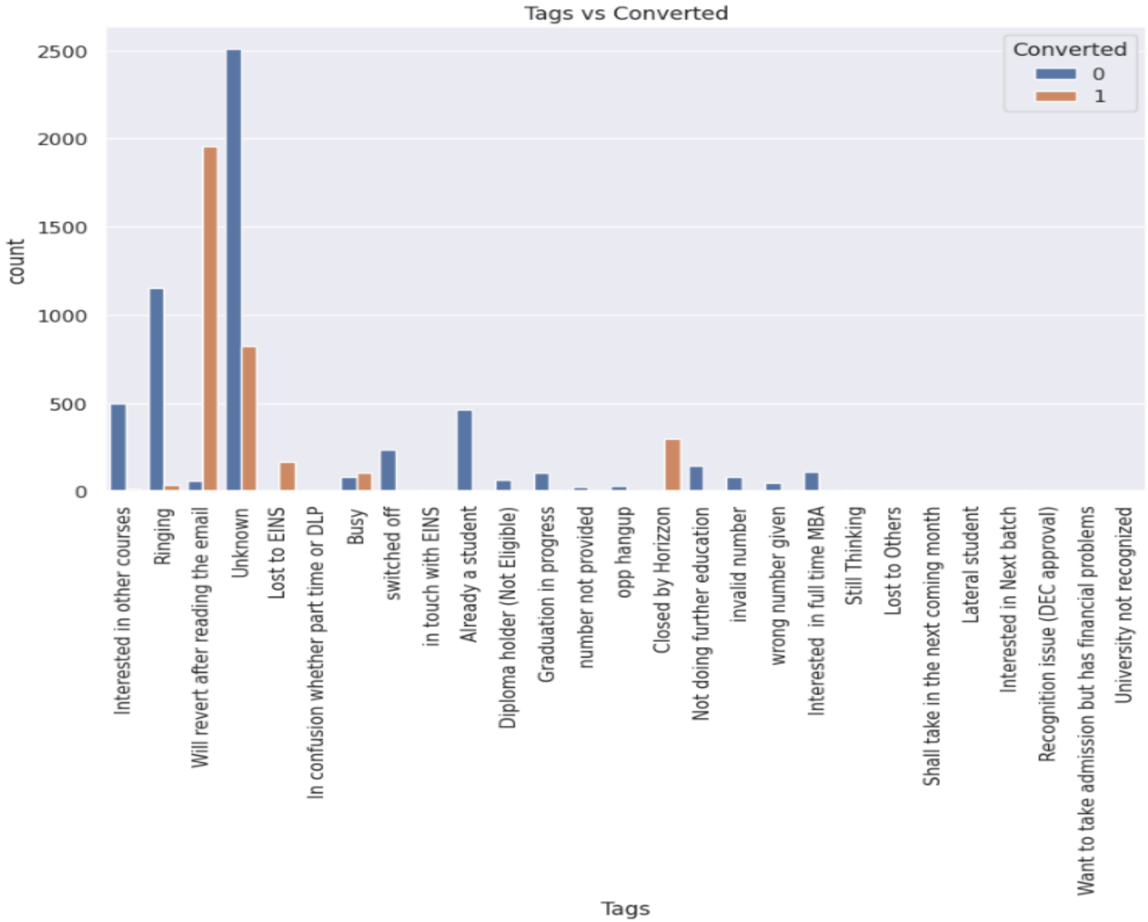
Business Understanding

1. **Challenge:**
   - Current lead conversion rate is very low – 30%.
   - High volume of leads demands a targeted strategy.
2. **Goal:**
   - Identify “Hot Leads” with conversion likelihood $\geq 80\%$.
   - Optimize sales team focus for higher conversion efficiency.
3. **Solution:**
   - Use a logistic regression model trained on historical data to identify “Hot Leads”
4. **Rationale:**
   - Data driven decision making will optimize resource allocation and improve conversion efficiency.
   - Machine learning model will give actionable insights to sharpen strategies.
5. **Expected Impact:**
   - Higher conversion rates by engaging high-potential leads
   - Improved ROI through targeted marketing efforts.

Data Preparation

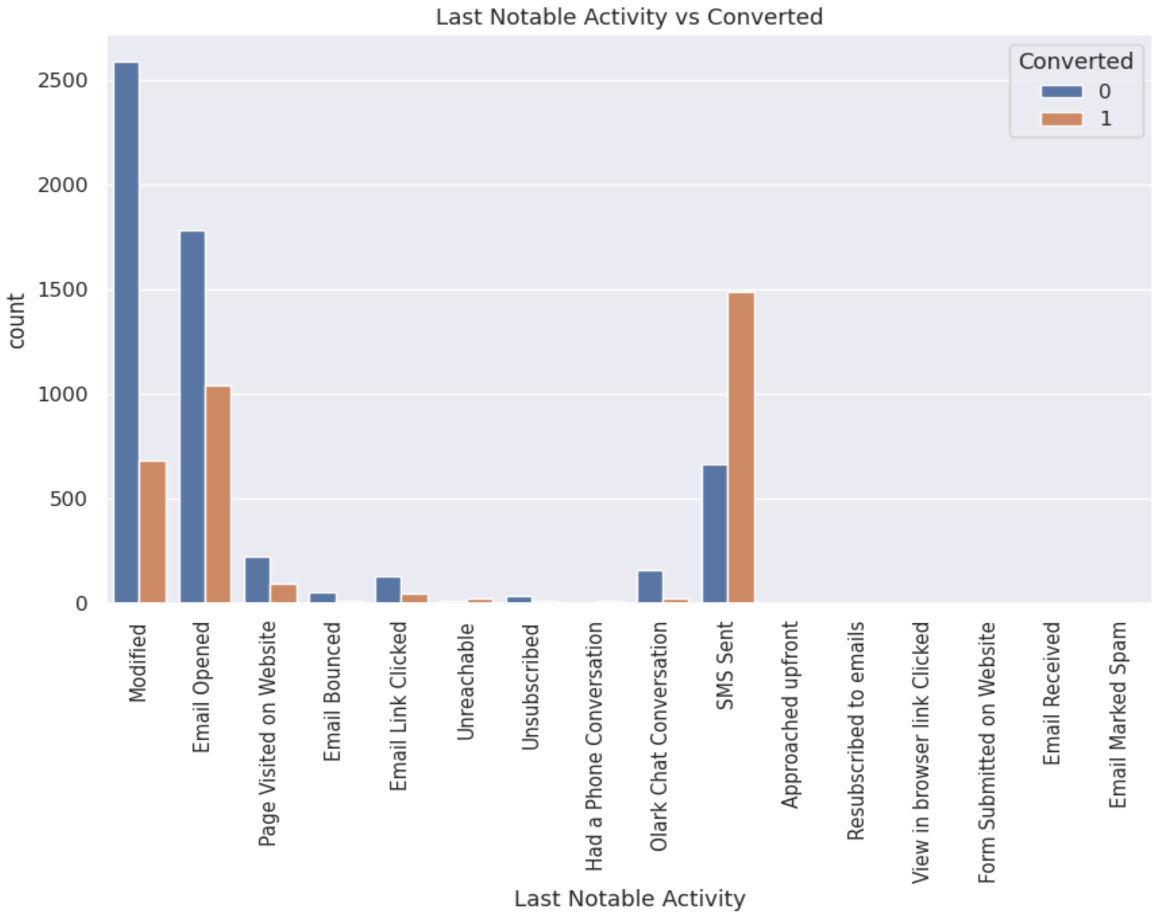
1. **Data Overview:**
   - 9240 leads with 37 attributes.
   - Columns with $\geq 45\%$ missing values dropped to ensure model quality.
2. **Data Imbalance:**
   - Not addressed since final model performed well in predicting minority class.
3. **Categorical columns:**
   - Dropped if one category dominated ($> 90\%$) to avoid overfitting.
4. **Sparse Dataframe:**
   - One-hot encoding led to sparse dataframe due to numerous categories with few records.
5. **Feature Selection:**
   - Initial coarse selection with RFE followed by manual refined selection.
6. **Feature Scaling:**
   - Min-max scaler used to prepare the dataframe for modelling.

Visualizations



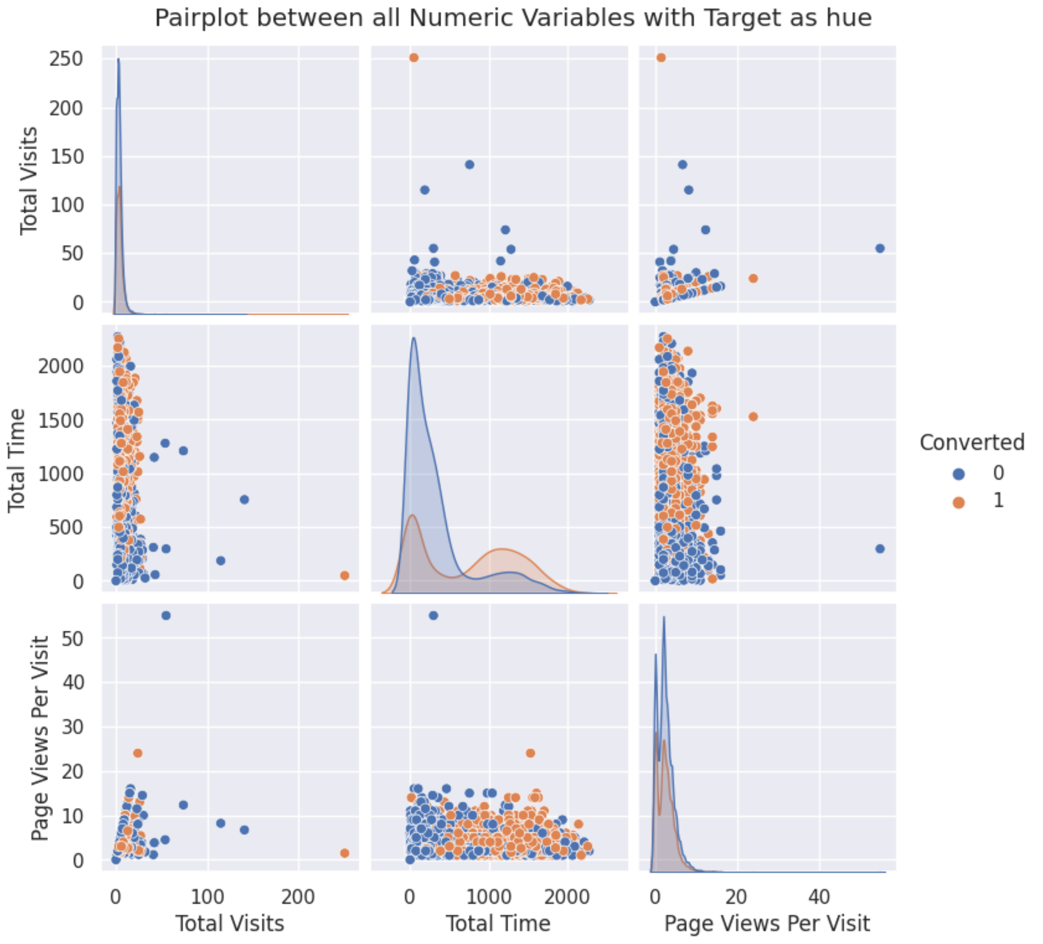
We observe that the leads with Tags marked as "Will revert after reading the email" and "Closed by Horizon" have higher percentage of lead conversions.

Visualizations (Contd.)



Leads with last activity and last notable activity of “SMS Sent” have higher percentage of lead conversions.

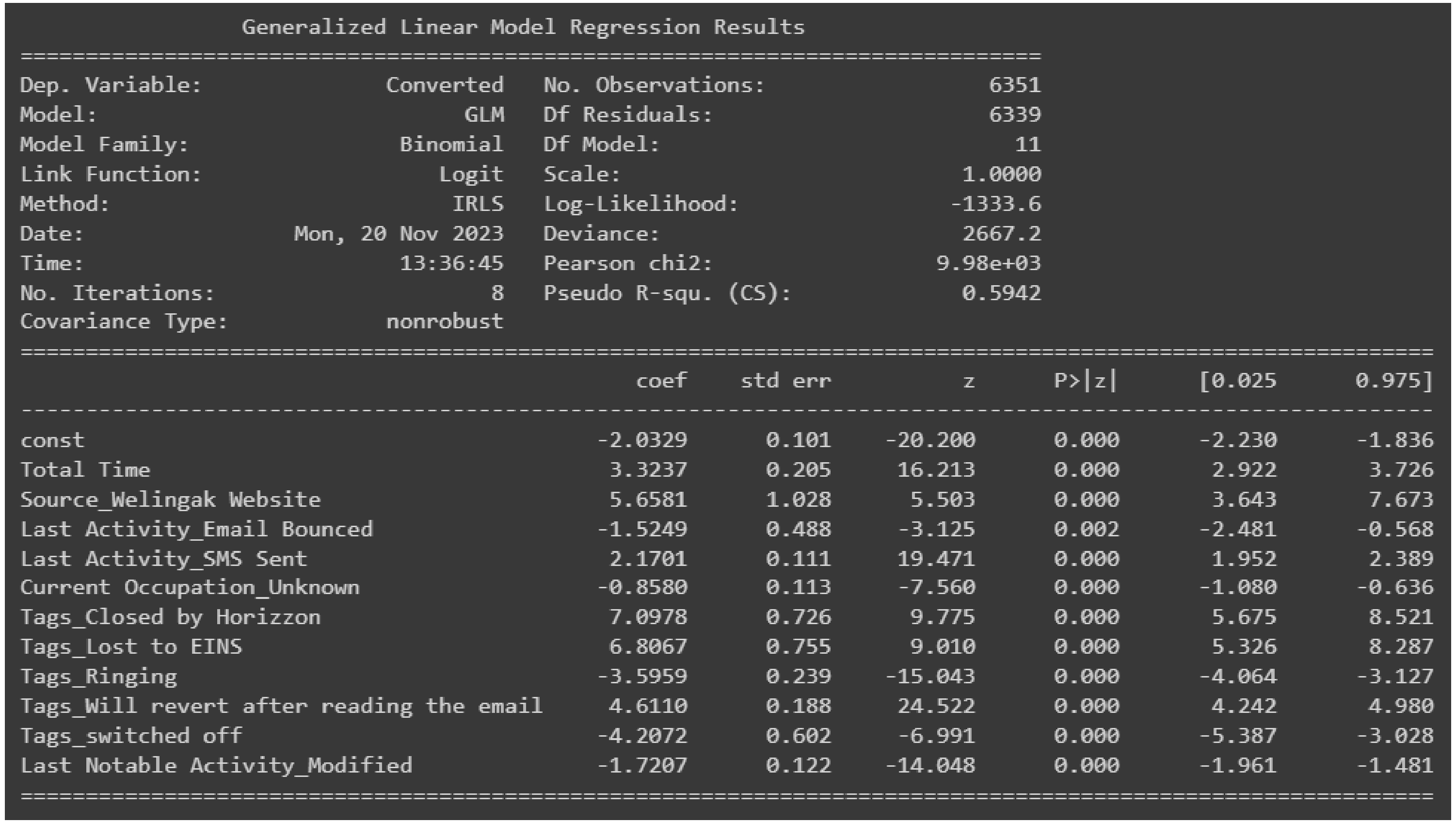
Visualizations (Contd.)



- There isn't much difference in the distributions of converted and not converted leads in terms of "Total Visits" and "Page Views Per Visit".

- Leads who spent longer time on the website seem to show higher conversion rates as per their distribution.

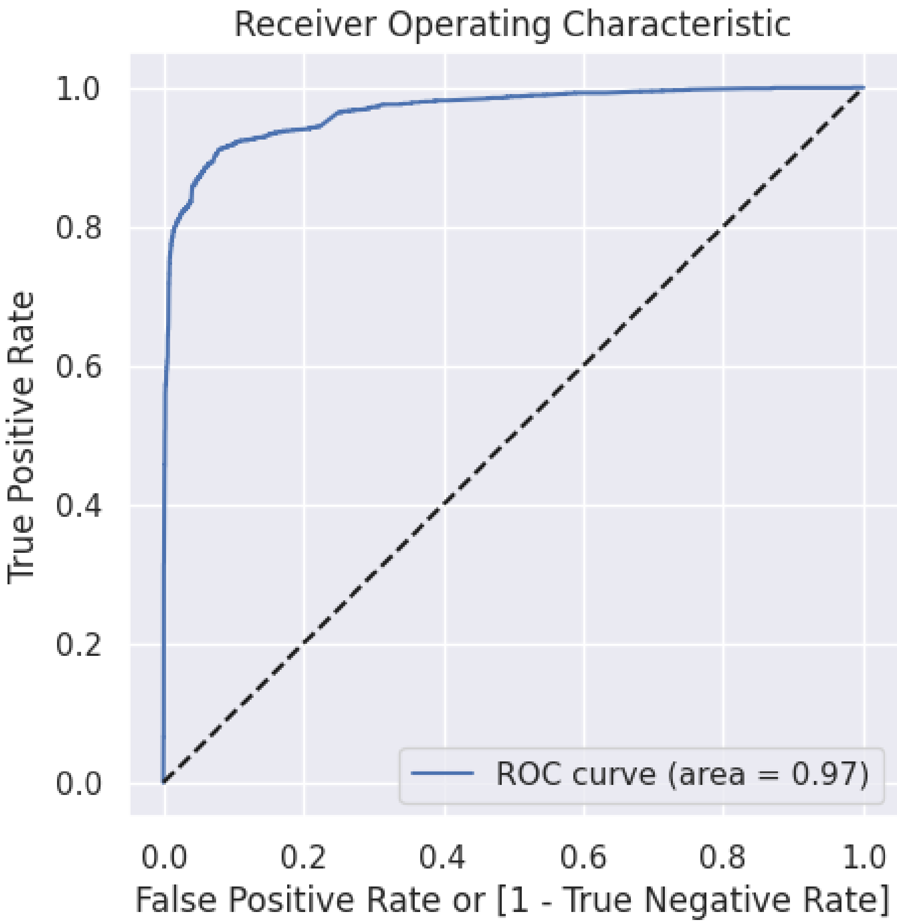
Final Model Summary



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6351
Model:                            GLM   Df Residuals:                     6339
Model Family:                Binomial   Df Model:                           11
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1333.6
Date:                Mon, 20 Nov 2023   Deviance:                       2667.2
Time:                        13:36:45   Pearson chi2:                 9.98e+03
No. Iterations:                     8   Pseudo R-squ. (CS):             0.5942
Covariance Type:            nonrobust
===========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------
const                                   -2.0329      0.101    -20.200      0.000      -2.230      -1.836
Total Time                               3.3237      0.205     16.213      0.000       2.922       3.726
Source_Welingak Website                  5.6581      1.028      5.503      0.000       3.643       7.673
Last Activity_Email Bounced             -1.5249      0.488     -3.125      0.002      -2.481      -0.568
Last Activity_SMS Sent                   2.1701      0.111     19.471      0.000       1.952       2.389
Current Occupation_Unknown              -0.8580      0.113     -7.560      0.000      -1.080      -0.636
Tags_Closed by Horizzon                  7.0978      0.726      9.775      0.000       5.675       8.521
Tags_Lost to EINS                        6.8067      0.755      9.010      0.000       5.326       8.287
Tags_Ringing                            -3.5959      0.239    -15.043      0.000      -4.064      -3.127
Tags_Will revert after reading the email 4.6110      0.188     24.522      0.000       4.242       4.980
Tags_switched off                       -4.2072      0.602     -6.991      0.000      -5.387      -3.028
Last Notable Activity_Modified          -1.7207      0.122    -14.048      0.000      -1.961      -1.481
===========================================================================================================
```

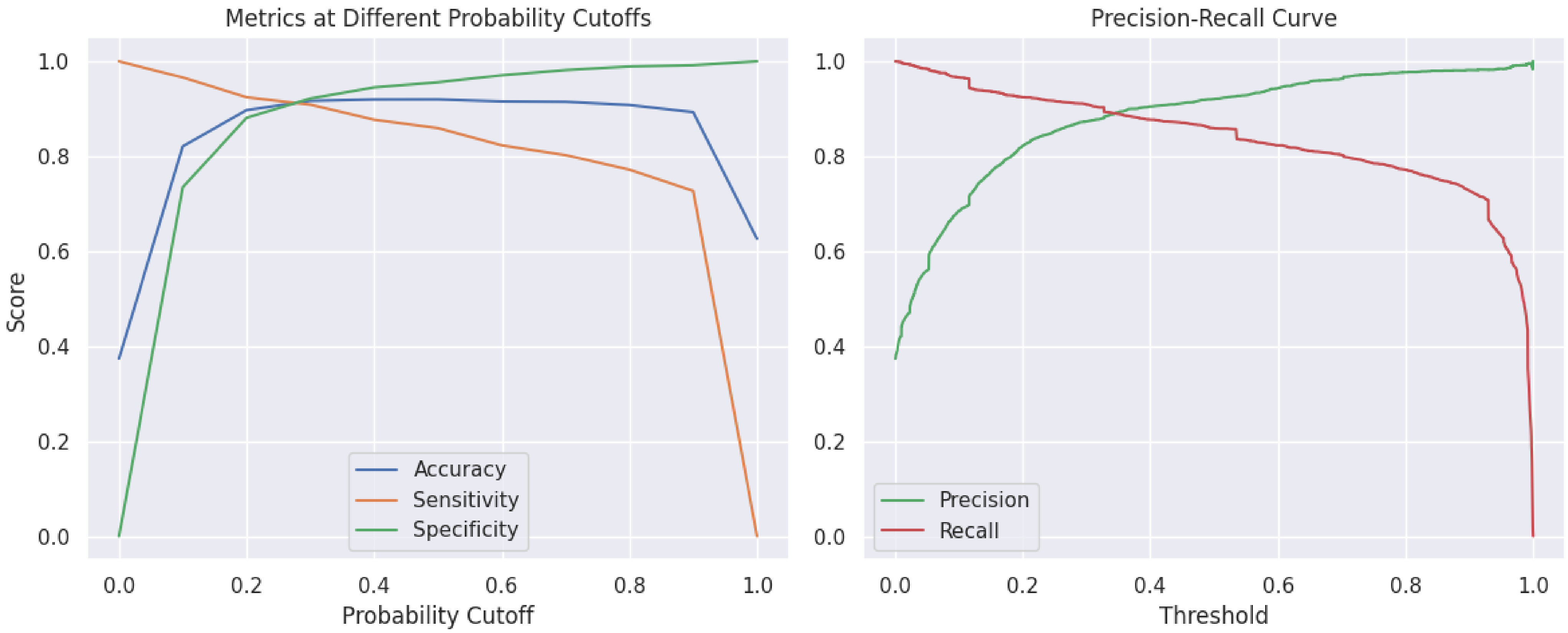
- Logistic Regression model.
- 11 finally chosen features.
- All variables significant (p-value close to 0)

ROC Curve



- Observed AUC = 0.97
- Extremely good model performance on train data

Determining Threshold



Metrics at Different Probability Cutoffs

Precision-Recall Curve

Chosen threshold 0.3 based on Sensitivity-Specificity and Precision-Recall trade-offs.

Model Performance

Metric	Train Data	Test Data
Accuracy	0.9167	0.9108
Sensitivity	0.9086	0.9161
Specificity	0.9215	0.9073
Precision	0.8736	0.8632
Recall	0.9086	0.9161
F1 – Score	0.8908	0.8889

- Model is performing well on both train and test data.
- Model is performing even better on test data based on some of the metrics.
- No underfitting or overfitting issues observed.

Model Summary

1. **High Sensitivity/Recall:**
   - Consistent achievement of over 90% sensitivity/recall on train and test data
   - Showcases robust ability to identify “Hot Leads”
2. **Key Feature Selection:**
   - “Total time spent on website”, “Last Activity” and “Tags” categories are important.
   - These offer actionable insights.
   - Interpretability and simplicity of the logistic regression model further helps.
3. **Lead Score Utility:**
   - Leads are scored on a range of 0 to 100 based on likelihood of conversion.
   - Helpful for practically identifying and prioritizing important leads.
   - Helps optimize resource allocation

Recommendations

1. **Utilize High-Performing Features:**
   - Use insights from key features to devise informed sales strategies.
   - Target on gathering more leads from sources with positive impact.
   - Consider communicating via channels which lead to higher conversions.
2. **Implement Lead Scoring:**
   - Prioritize leads based on their lead scores.
   - Define Lead Score cut-offs based on business strategy followed at the moment.
3. **Regular Model Monitoring and Updates:**
   - Periodic model updates to adapt to evolving lead behaviour.
   - Consider further research in understanding competitor behaviour.