Lead Scoring - Case Study

Content :

- Data Cleaning
- EDA (Exploratory Data Analysis)
- Data Preparation for Model Building
- Building a Model
- Model Evaluation and Prediction
- Conclusion

Data Cleaning

Importance of Data Cleaning

Data cleaning is crucial before performing data analysis or applying machine learning techniques because:

- **Accuracy**: Ensures the data is accurate and free of errors, leading to more reliable results.
- **Consistency**: Removes inconsistencies and standardizes the data, making it uniform.
- **Completeness**: Fills in missing values and removes incomplete records, ensuring comprehensive analysis.
- **Performance**: Improves the efficiency and performance of algorithms by eliminating irrelevant or redundant data.
- **Insights**: Enhances the quality of insights and decisions derived from the data.

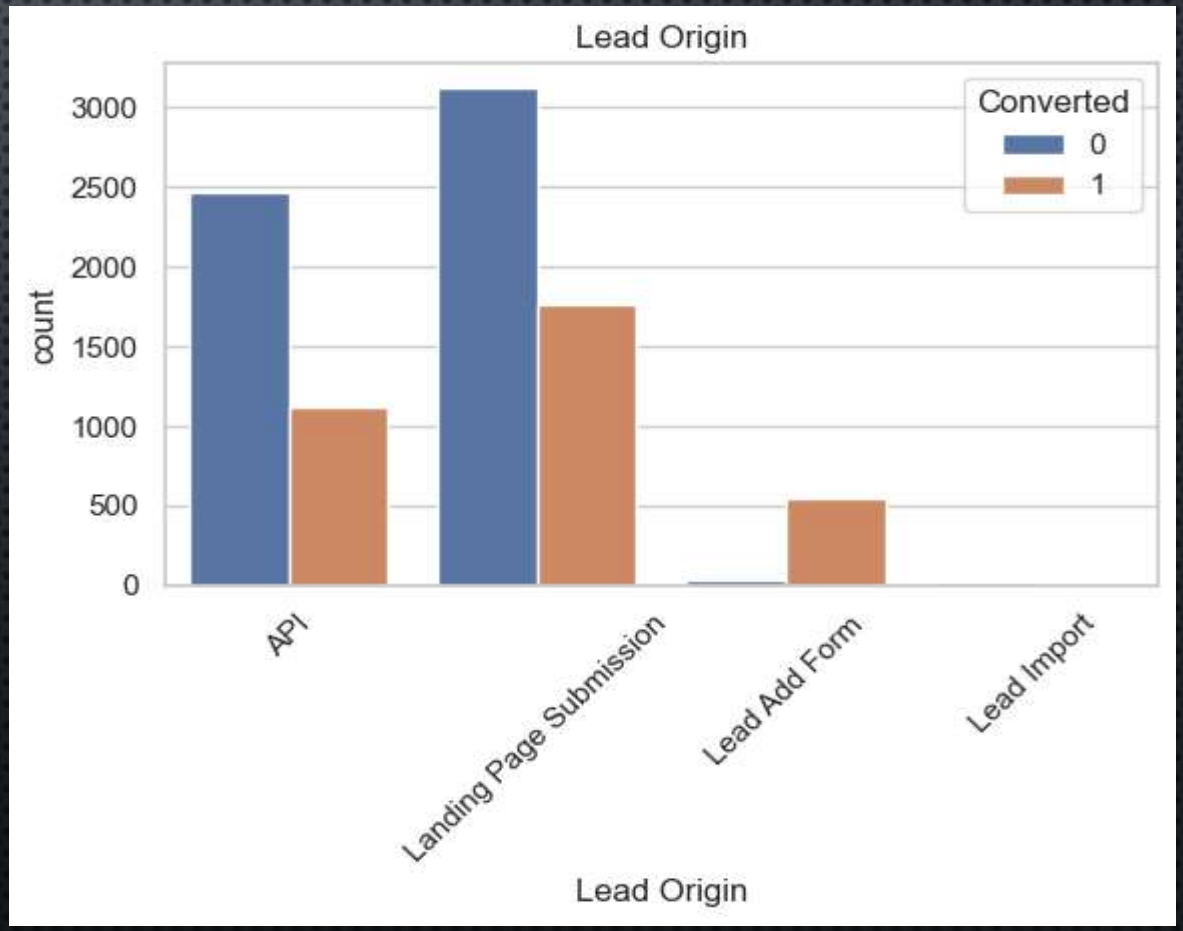
Therefore,
We must Check for null values, Check for missing values, Handling missing values, Dropping unnecessary columns and finally checking for Duplicates.

After Data Cleaning :

So, after performing various methods of cleaning the data, we come to this conclusion that 98% of the data has been retained after cleaning the data, removing the null values and standardizing some of the rows and also there are no duplicates present as well.

Now based on our data cleaning if check with a count plot for Lead origin we can infer the following :

- Landing page submission has both the highest and lowest conversion rate.
- API comes in the 2nd place with highest and lowest conversion rate.
- Lead import has no conversions
- Lead add form has the highest ratio in terms of conversion to no conversion.



EDA

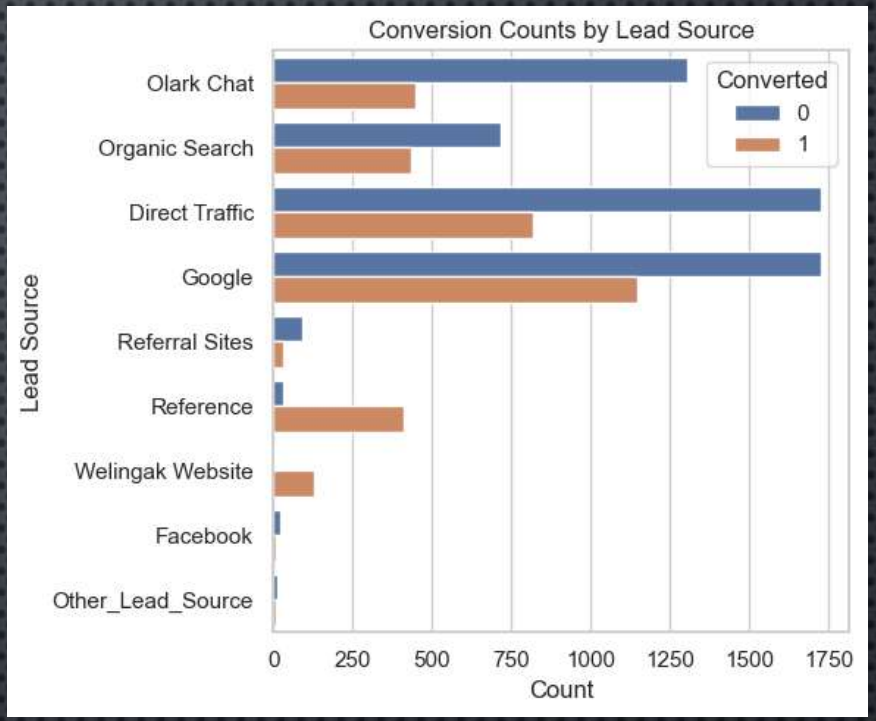
Importance of EDA

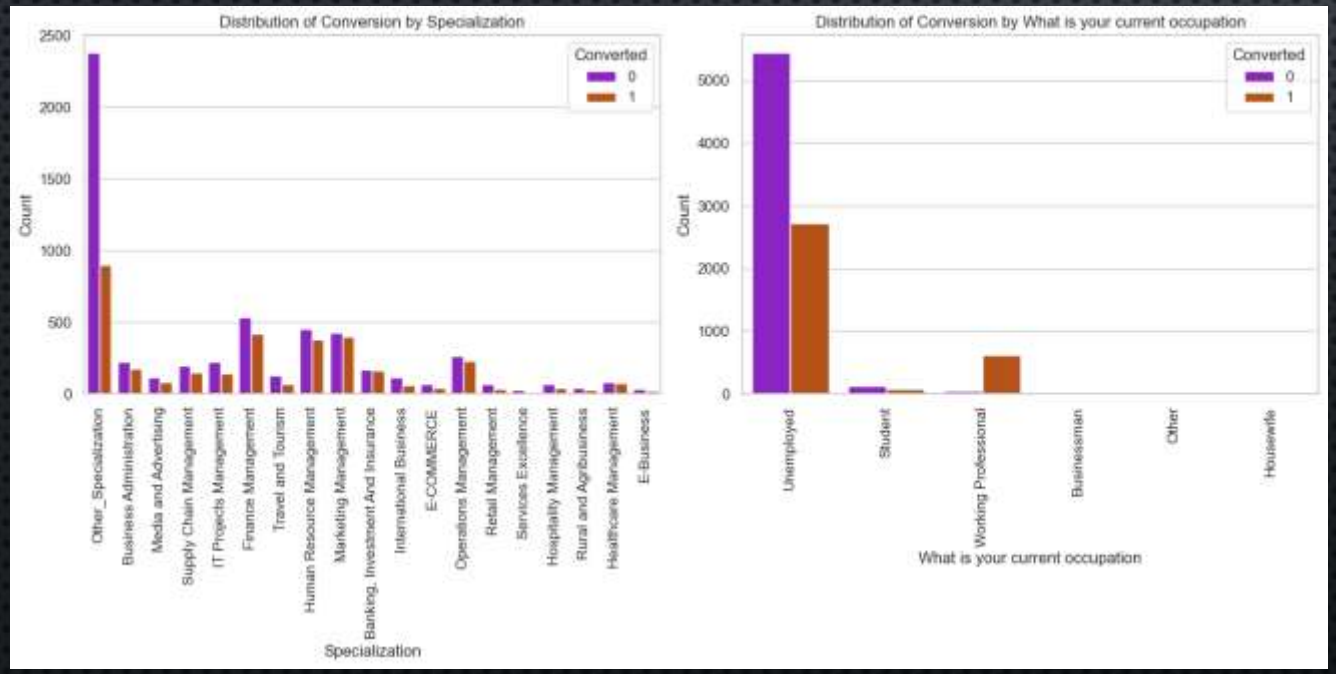
Exploratory Data Analysis (EDA) is important before applying any machine learning techniques because:

- **Understanding Data**: Provides insights into the structure, patterns, and relationships within the data.
- **Identifying Anomalies**: Helps detect outliers, missing values, and errors that need to be addressed.
- **Feature Selection**: Assists in identifying relevant features and reducing dimensionality for better model performance.
- **Hypothesis Testing**: Allows you to test assumptions and hypotheses about the data.
- **Data Distribution**: Helps understand the distribution of variables, which informs the choice of algorithms and preprocessing steps.
- **Improving Models**: Guides the preprocessing steps and feature engineering, leading to more accurate and efficient models.

Now based on the conversion counts of various lead sources. We can see that ,

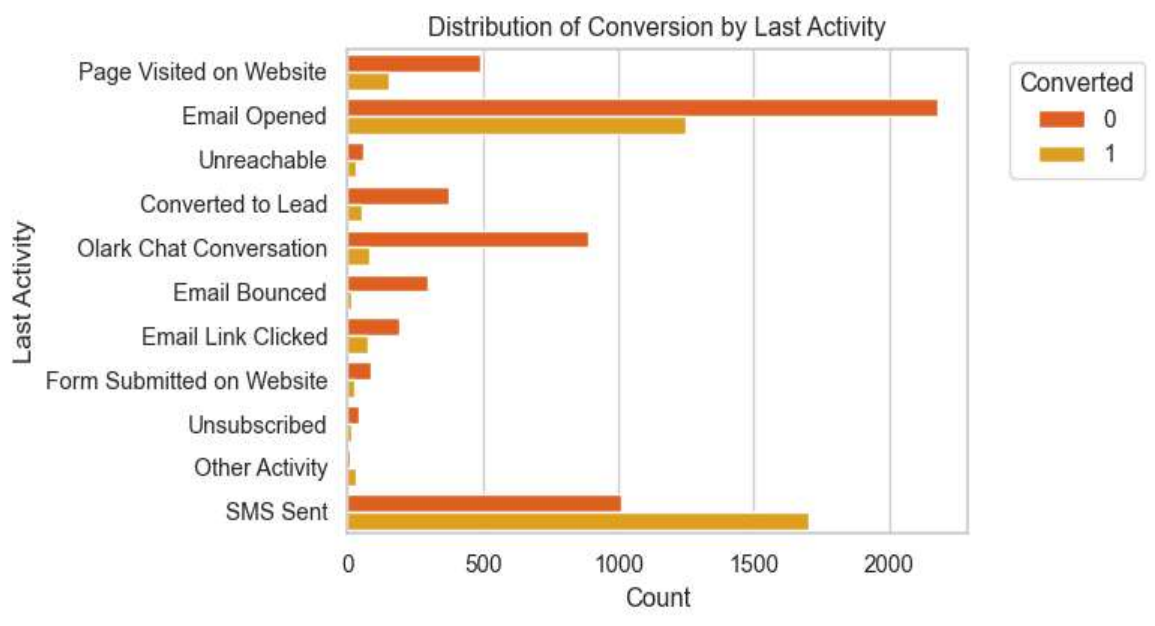
- Google has the highest count of conversions of leads.
- In the 2nd place is Direct traffic.
- Quite surprisingly Facebook has a low conversion rate, given the stature of the medium it seems a better decision not advertise on Facebook.





For Specialization we can see that **Others and Finance management** have highest conversion rates.

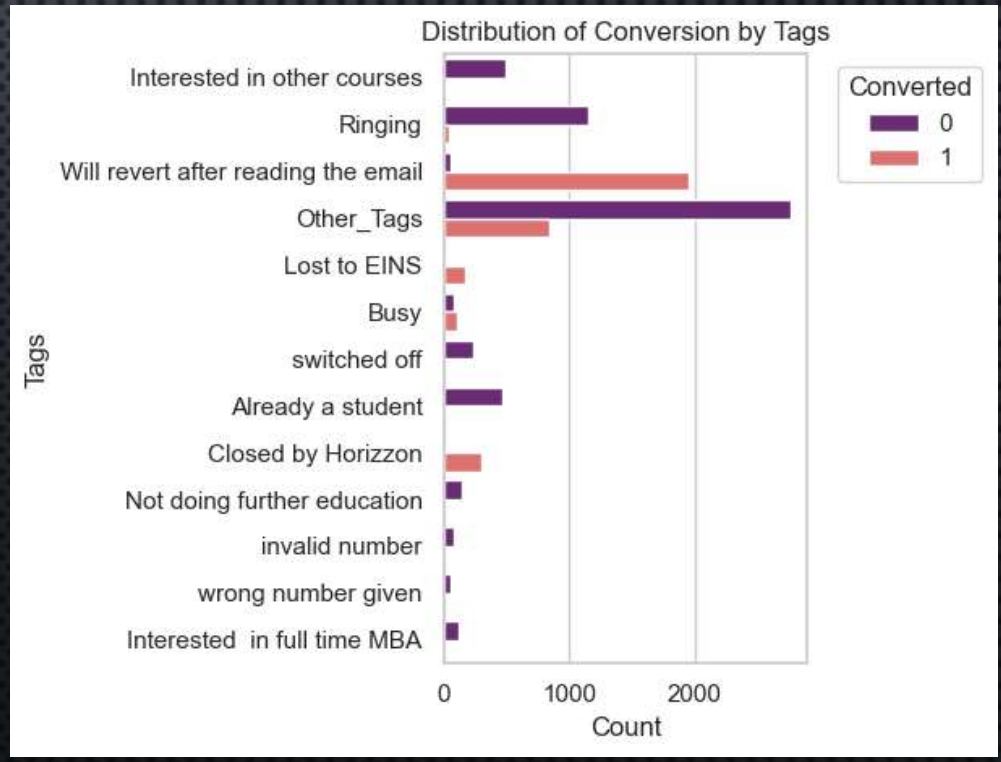
In terms of occupation **Unemployed** have the highest non-conversion rate



Conversion based on the last activity shows that SMS sent and Email Opened have the highest conversion rates.
Also we can see that Email Opened has the highest rate of non conversion as well.

Here based on the Tags we can see that other tags have highest rate of non-conversion.

Whereas, reverting back after reading the email has the highest rate of conversion.

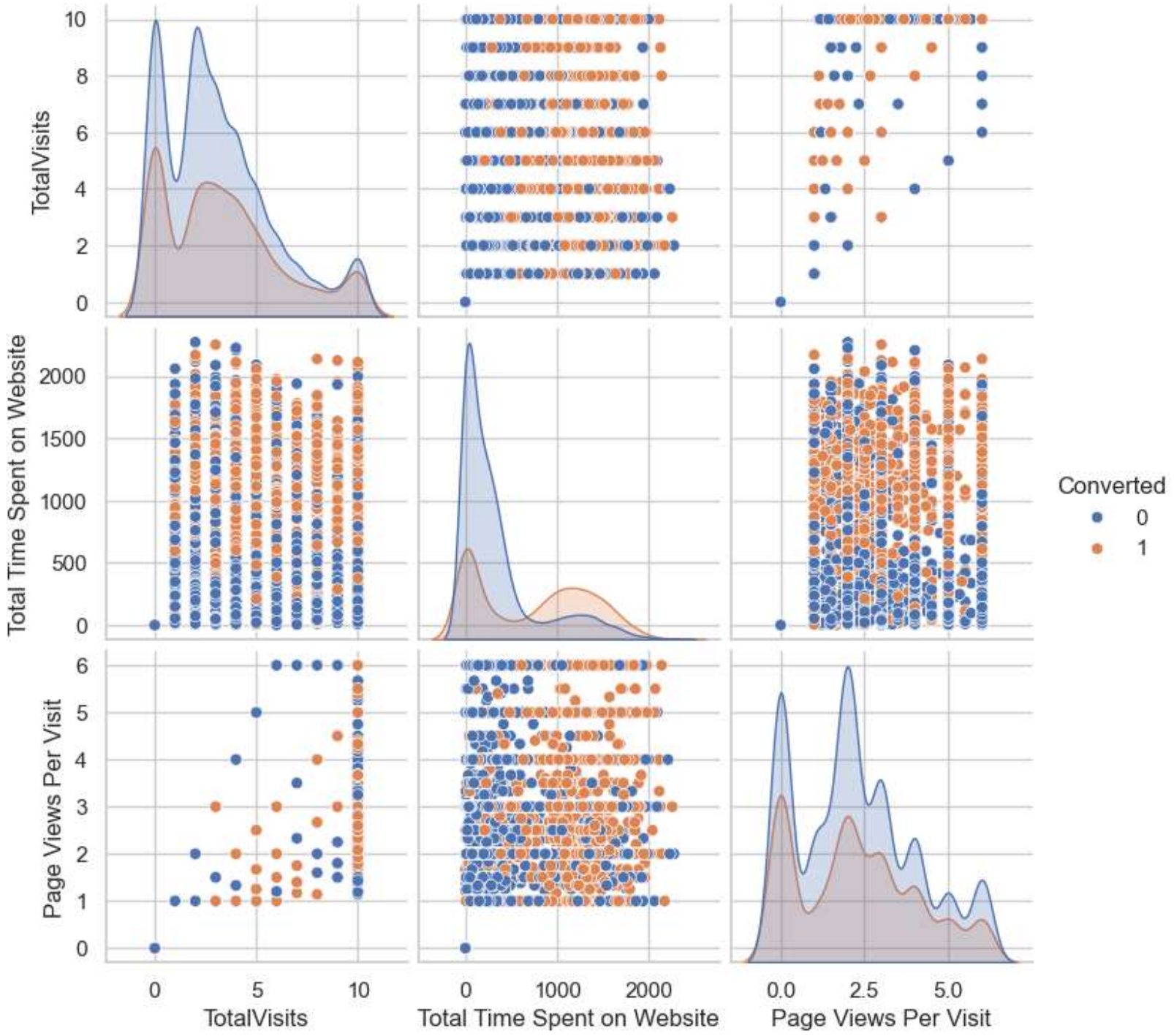


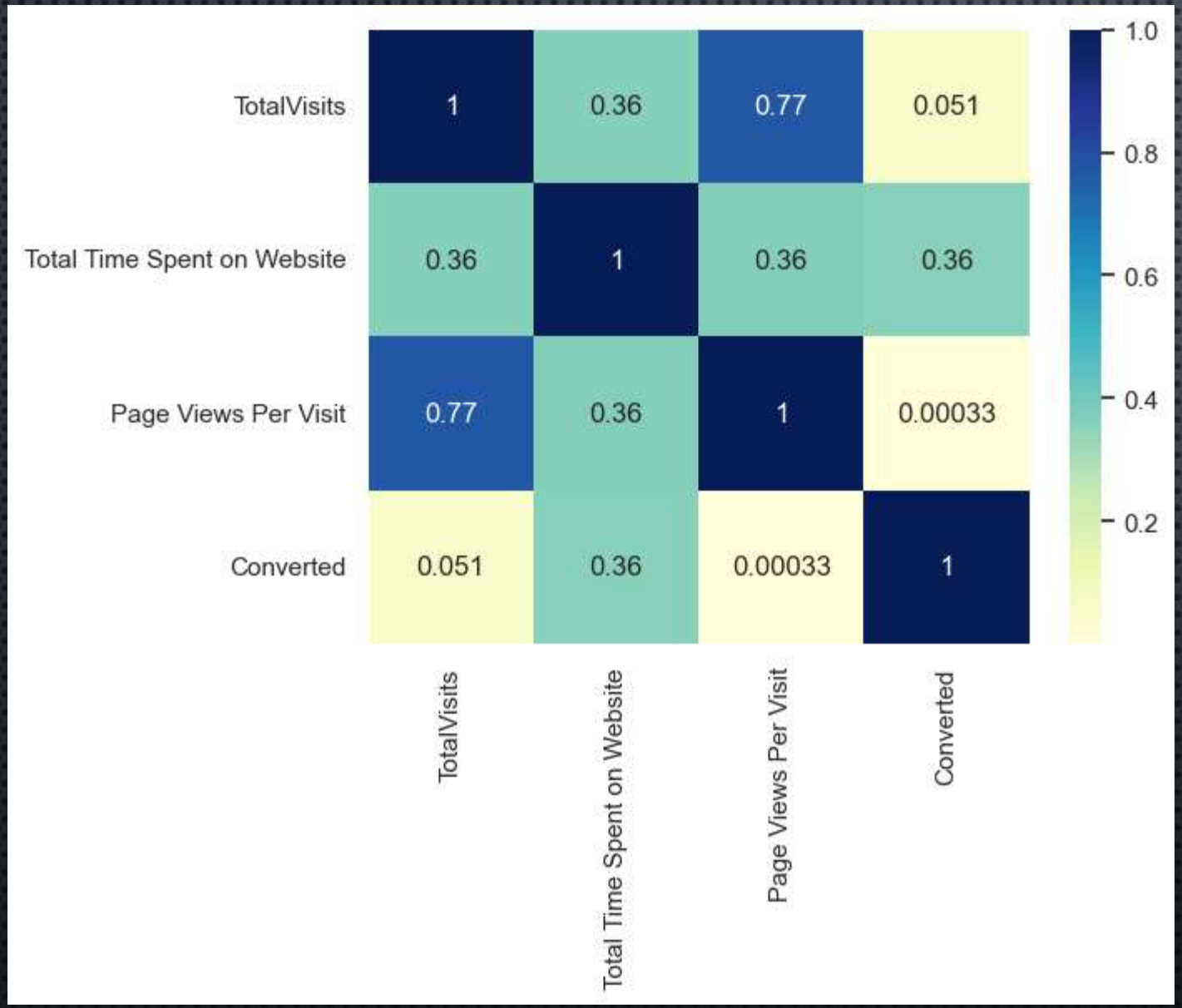
Now based on the Amount of time Spent on the website we can make the following inferences:

Total Time Spent on Website: Shows a positive correlation with conversions

Page Views Per Visit: Also indicates a positive correlation

Total Visits: Does not show a clear trend with conversions





Now based on the Amount of time Spent on the website we can make the following inferences:

Total Time Spent on Website: Has moderately low relation with conversion.

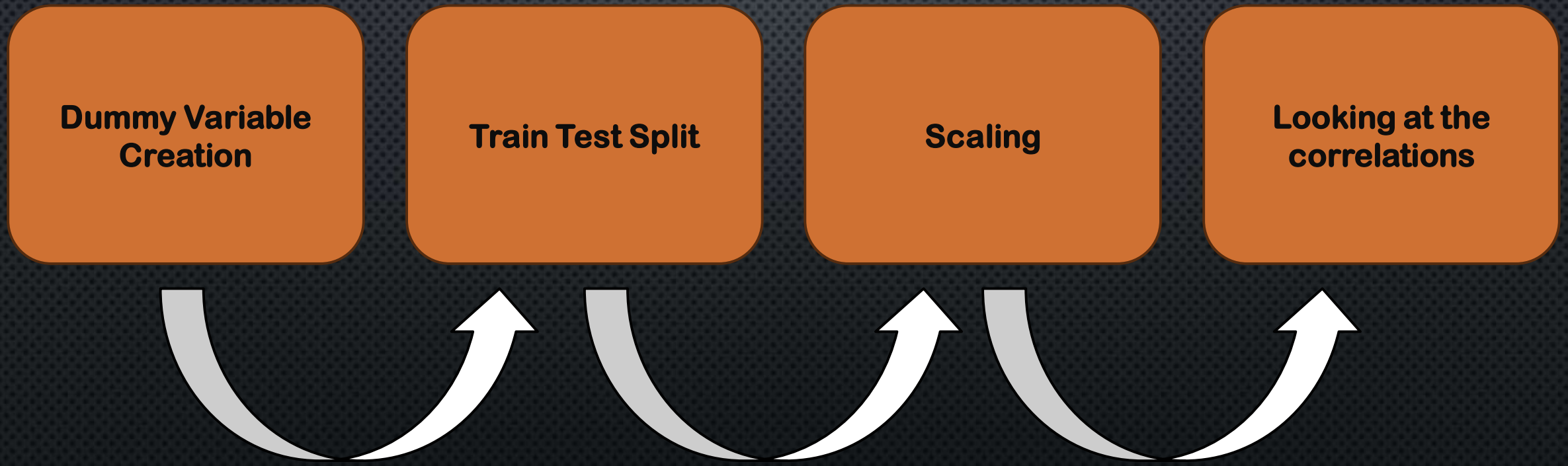
Page Views Per Visit: Has a very low relation with conversion.

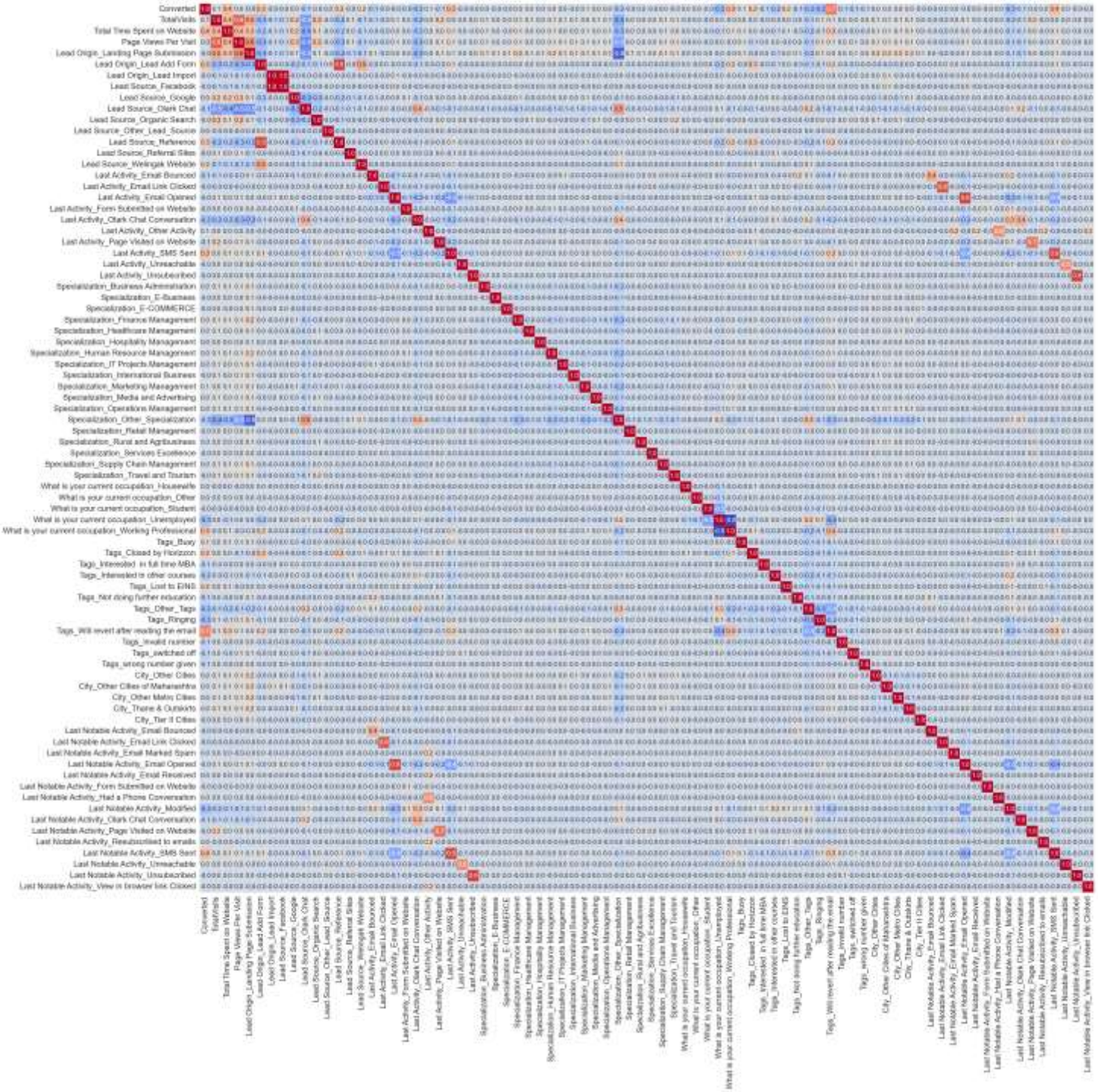
Total Visits: Has moderate relation to conversion.

Data Preparation for model building

Data Preparation

Once our data cleaning is done and we have dropped the necessary columns and did our Univariate analysis now we start preparing our dataset for model building. Going through the following sequential steps.





Now based on the heatmap we have generated we would not be able to see clearly because of the fonts but based on our previous Data visualization we can see a few clear pointers related to Conversion.

Total visits

Total time spent on the website.

Total page views.

SMS sent

Email received

Email opened etc.

Just to name a few are the places on the heatmap we can see have generated more heat in relation to conversion.

Model building

For our Model building –

As we can see that there are a lot of variables present in the dataset which we cannot deal with. So the best way to approach this is to select a small set of features from this pool of variables using RFE.

What is RFE?

Recursive Feature Elimination (RFE)

We are then going to use the RFE method :

RFE Approach

Recursive Feature Elimination (RFE) is a technique used for feature selection in linear regression. It recursively removes the least significant features and builds the model using the remaining attributes. By ranking and eliminating features based on their contribution to the model, RFE helps in creating a more efficient and accurate model with better generalization.

For our Model building we also need to check for the VIF values.

What is VIF?

The Variance Inflation Factor (VIF) is a measure used to detect the presence and severity of multicollinearity in a regression model. Multicollinearity occurs when two or more predictor variables in the model are highly correlated, making it difficult to determine the individual effect of each predictor on the dependent variable.

Hence, Understanding and managing VIF is crucial for building robust regression models that provide reliable and interpretable results.

Features	Variance Inflation Factor
Lead Origin_Landing Page Submission	2.18
Tags_Will revert after reading the email	1.80
Tags_Other_Tags	1.76
Lead Origin_Lead Add Form	1.74
Last Activity_SMS Sent	1.71
Last Notable Activity_Modified	1.47
Lead Source_Welingak Website	1.36
Tags_Ringing	1.35
Tags_Closed by Horizzon	1.22
Tags_Busy	1.09
Tags_switched off	1.08
Tags_Lost to EINS	1.07
Last Notable Activity_Olark Chat Conversation	1.05
Tags_invalid number	1.03
Tags_wrong number given	1.02

Going ahead as check for VIF values we can see that –

All variables have a good value of VIF. But we observed earlier that the column Tags_invalid number and Tags_wrong number given has high p-value and hence we will drop this column and remake the model.

Going ahead we will drop one or two variables at times based on their P-values and keep rebuilding a model until we get a proper model exhibiting proper P-values.

Now as go further we will keep dropping certain variables as we have already dropped **Tags_invalid number and Tags_wrong number**

Now we will also drop the following variables :

- Lead Origin_Lead Add Form
- Tags_Switched off

Now by this time we are going into our 5th model.
We can see that our P-values are fairly up to the mark.

And we have an accuracy score of 92.59%.

Features	Variance Inflation Factor
Lead Origin_Landing Page Submission	1.90
Tags_Other_Tags	1.68
Last Activity_SMS Sent	1.63
Tags_Will revert after reading the email	1.54
Last Notable Activity_Modified	1.45
Tags_Ringing	1.30
Tags_Busy	1.08
Tags_Closed by Horizzon	1.07
Lead Source_Welingak Website	1.06
Tags_Lost to EINS	1.05
Last Notable Activity_Olark Chat Conversation	1.04

All variables exhibit satisfactory VIF and p-values. Therefore, there's no need to eliminate any more variables, and we can continue to make predictions using this model as it is.

Model Evaluation and Prediction

Plotting ROC curve.

What is an ROC curve?

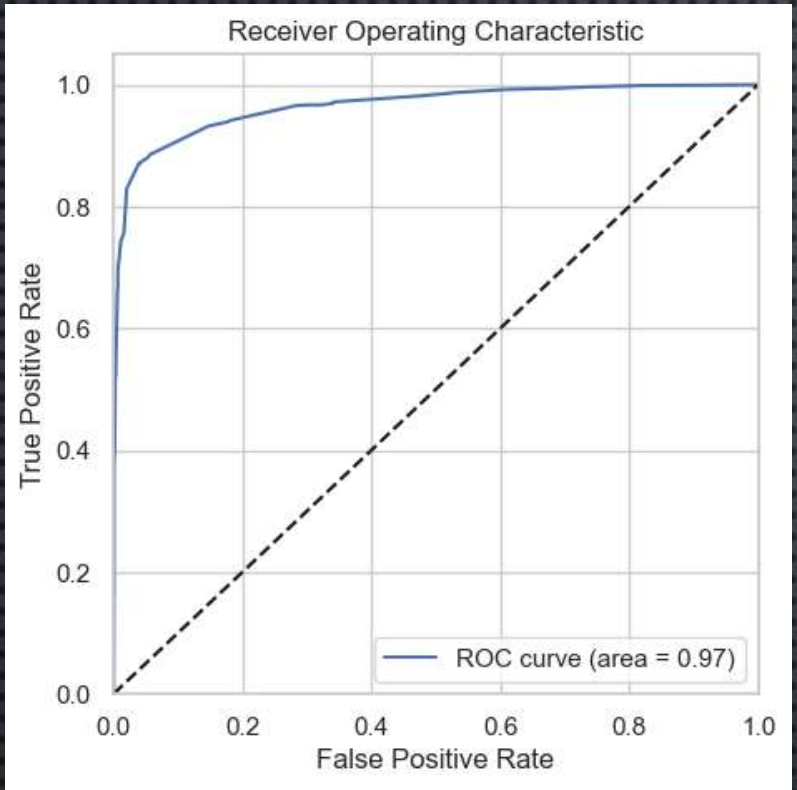
A Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of a binary classification model. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

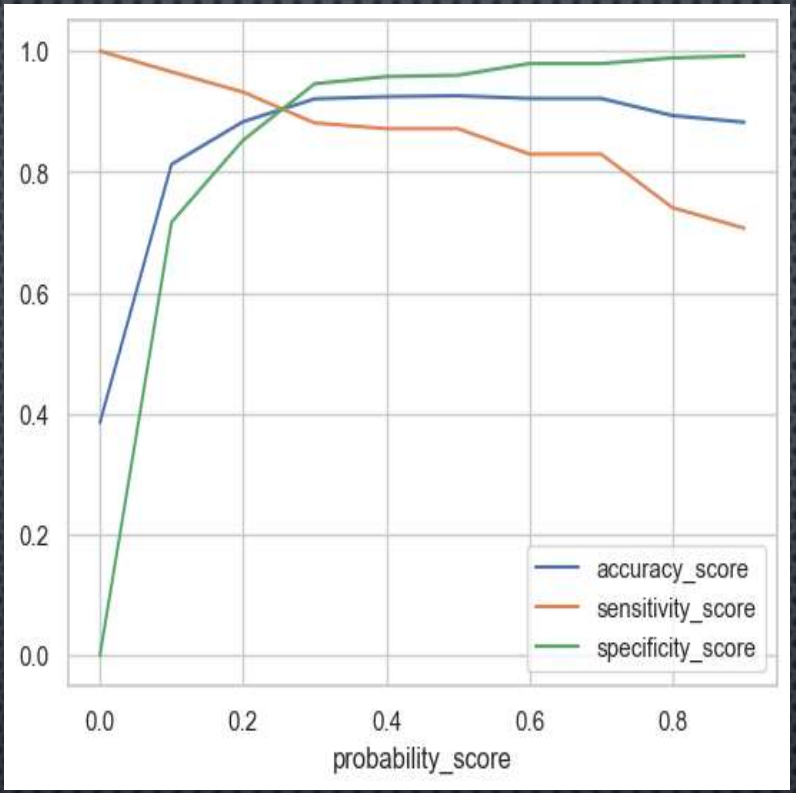
Sensitivity and Specificity Tradeoff: The ROC curve illustrates the tradeoff between sensitivity (true positive rate) and specificity (true negative rate). Improving one typically reduces the other.

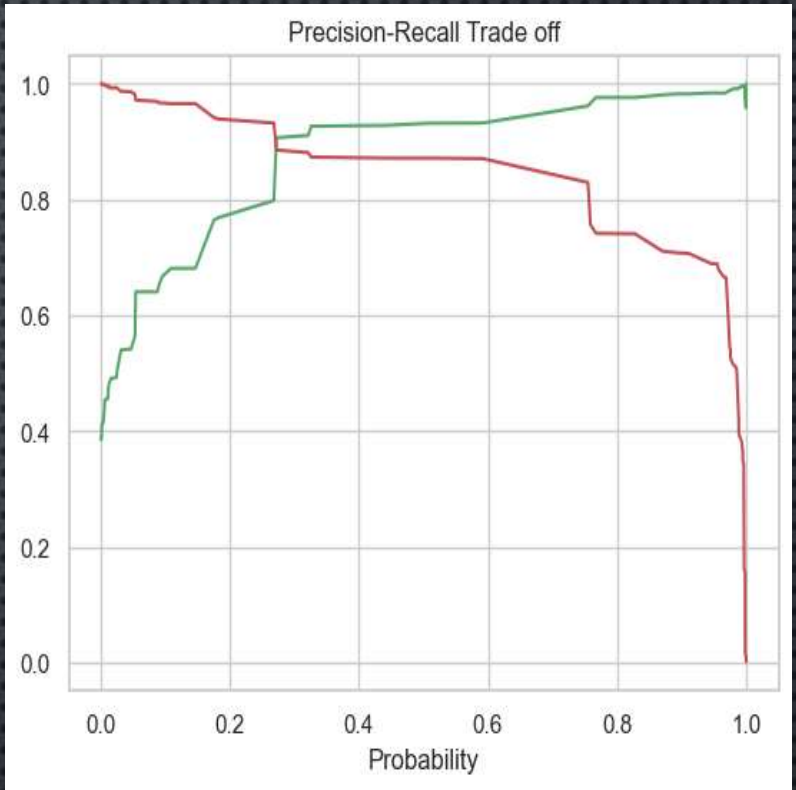
Area Under the Curve (AUC): A curve that closely follows the y-axis and the top border of the ROC space indicates a larger AUC, suggesting higher accuracy of the test or model.

Diagonal Reference Line: The closer the ROC curve is to the 45-degree diagonal line (the reference line), the smaller the AUC, indicating lower accuracy of the test or model.

Here, our goal is to have achieve good sensitivity score







Sensitivity-Specificity-Accuracy Plot: At a probability cutoff of 0.27, the Sensitivity-Specificity-Accuracy plot suggests an optimal point. This cutoff balances sensitivity (true positive rate), specificity (true negative rate), and overall accuracy effectively.

Precision-Recall Curve: The Precision-Recall curve indicates that a cutoff probability of 0.3 is optimal. This threshold maximizes precision (positive predictive value) and recall (sensitivity) for the model.

Cutoff Probability Selection: Based on these analyses, 0.27 is chosen as the optimal cutoff probability for assigning Lead Scores in the training data. This ensures a balance between correctly identifying leads and minimizing misclassifications.

Results of our final Model:

Accuracy: 0.91 (91% of the instances are correctly classified)

Confusion Matrix:

True Positives (TP): 856 True Negatives (TN): 1634 False Positives (FP): 100 False Negatives (FN): 133

Precision: 0.89 (89% of the predicted positives are correct)

Recall: 0.86 (86% of the actual positives are correctly predicted)

F1 Score: 0.88

Specificity: 0.94 (94% of the actual negatives are correctly predicted)

ROC-AUC Score: 0.90

Conclusion

Key Takeaways

So based on our final model we can see that we have achieved an Accuracy score of 91%.

We also have a Precision score of 89%.

We also have a Recall score of 86%

We also have a Specificity score of 94%

And finally we can see ROC-AUC score of 0.09.

Thank you