Lead Scoring Analysis for X Education

Logistic Regression for Hot Lead Identification

SUBMITTED BY:

ANKUR SARAN

ANKIT MATHUR

SAILAXMI ANUMOTHU

Contents

- Problem statement
- Business Objective
- Analysis approach
- Data Cleaning
- Visualization
- Business Recommendation
- Conclusion

Problem Statement

X Education struggles with a low lead conversion rate (~30%).

To make this process more efficient, the company wishes to identify the most potential leads, also known as Hot Leads. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

Business Objective

Lead X wants to build a model to assign every lead a lead score between 0 -100 , so that they can identify the hot leads and increase their conversion rate as well.

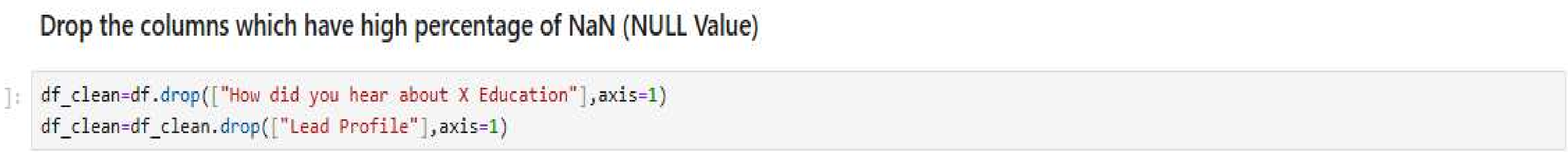
The CEO want to achieve a lead conversion rate of 80%.

They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approach.

Analysis Approach

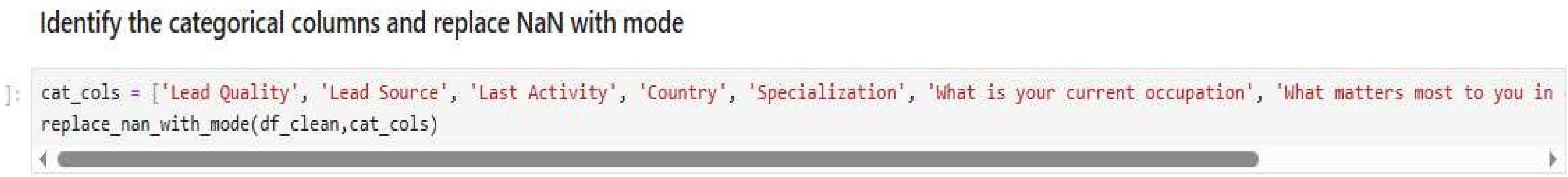
- Load Data
- EDA : Handle Missing Values, Visualize Data
  - Special focus on ‘Select’ values.
  - Identify the categorical columns and replace NaN with mode
  - Identify the Numeric columns and replace NaN with median
  - Univariate Analysis
- Data Preprocessing (one Hot Encoding)
- Test-Train Split
- Data Scaling (Min Max Scaler)
- Feature Selection (RFE for Top Predictors)
- Model Building (Logistic Regression)
  - Get the ‘LEAD SCORE’ based on Prediction Percentages
- Evaluation Metrics and Insights: Accuracy, Precision and Recall.

EDA



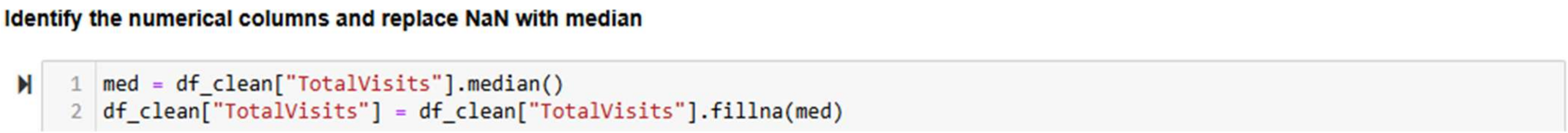
Drop the columns which have high percentage of NaN (NULL Value)

```
df_clean=df.drop(["How did you hear about X Education"],axis=1)
df_clean=df_clean.drop(["Lead Profile"],axis=1)
```



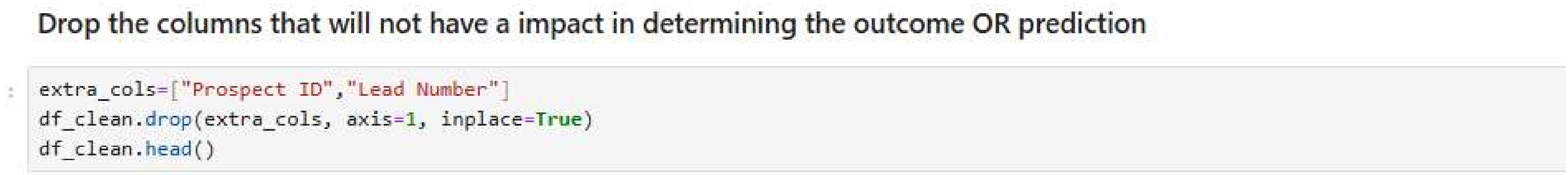
Identify the categorical columns and replace NaN with mode

```
cat_cols = ['Lead Quality', 'Lead Source', 'Last Activity', 'Country', 'Specialization', 'What is your current occupation', 'What matters most to you in
replace_nan_with_mode(df_clean,cat_cols)
```



Identify the numerical columns and replace NaN with median

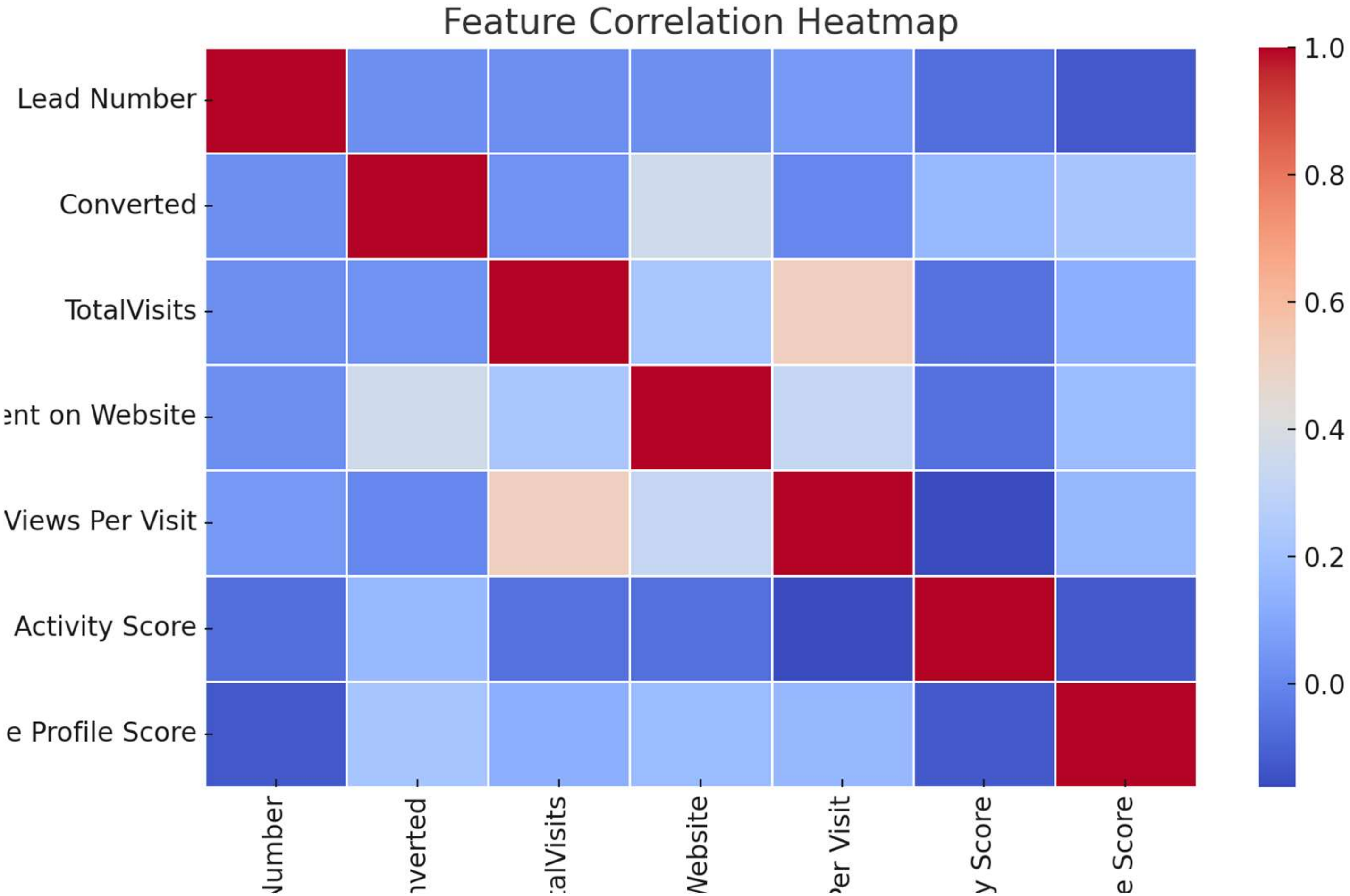
```
med = df_clean["TotalVisits"].median()
df_clean["TotalVisits"] = df_clean["TotalVisits"].fillna(med)
```



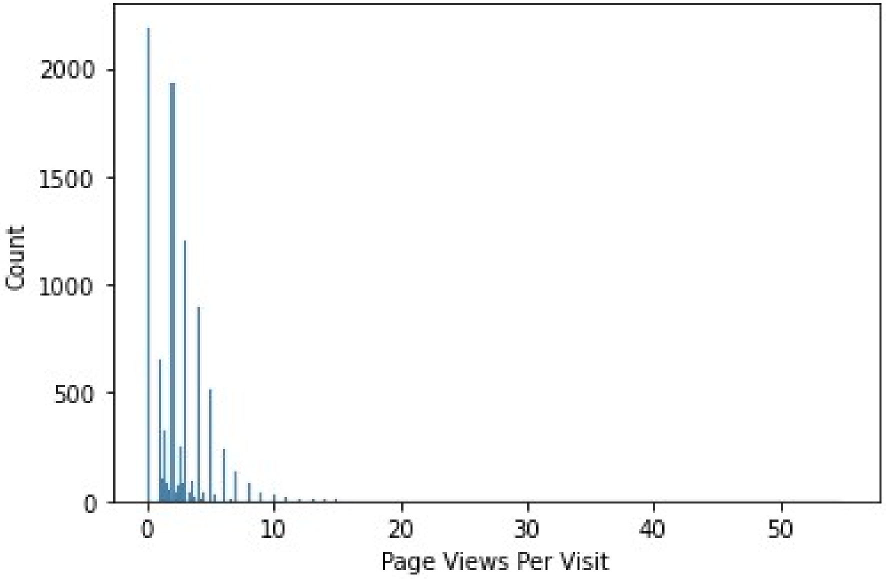
Drop the columns that will not have a impact in determining the outcome OR prediction

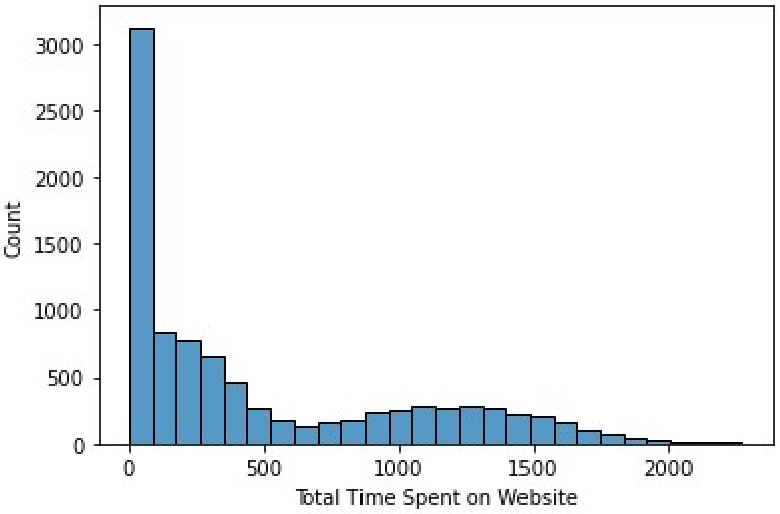
```
extra_cols=["Prospect ID","Lead Number"]
df_clean.drop(extra_cols, axis=1, inplace=True)
df_clean.head()
```

Visualization: Correlation Heatmap

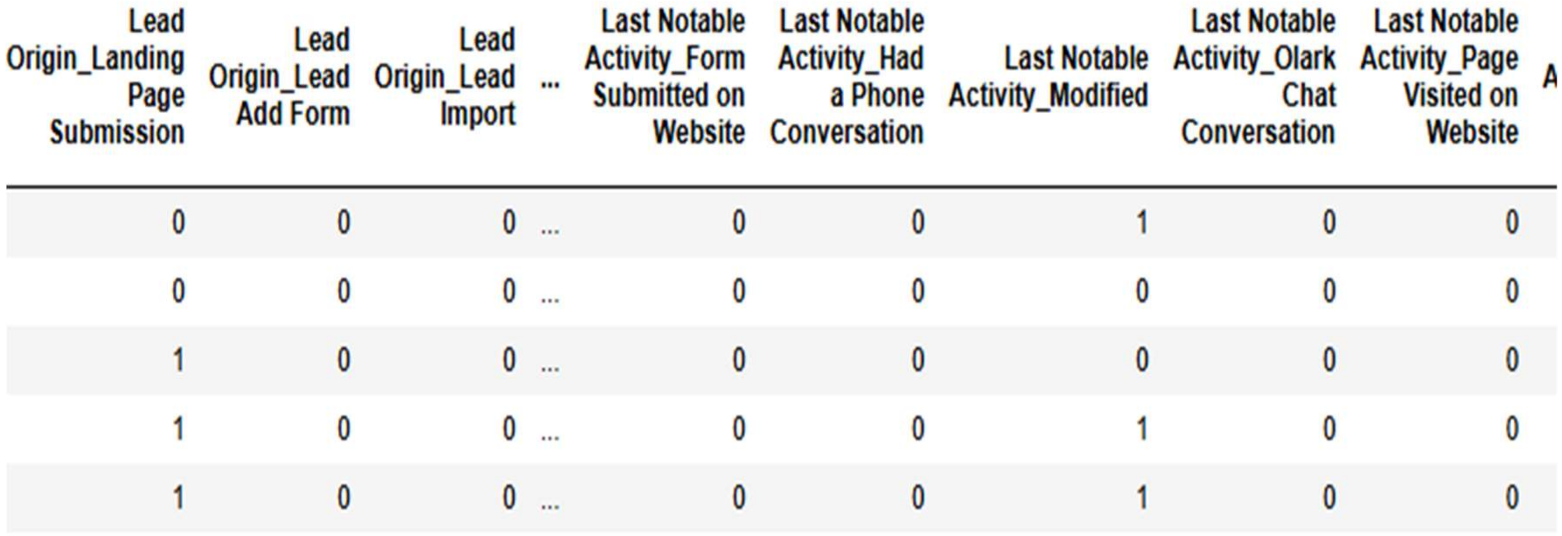


Visualization: Page view per visit and Time Spent Analysis





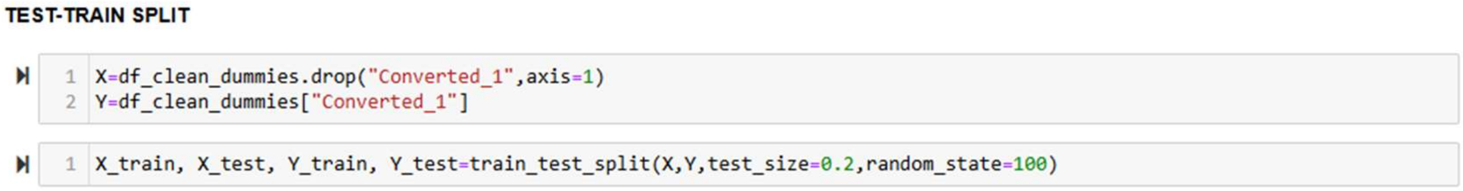
Pre-Processing: Encoding Categorical



Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	...	Last Notable Activity_Form Submitted on Website	Last Notable Activity_Had a Phone Conversation	Last Notable Activity_Modified	Last Notable Activity_Olark Chat Conversation	Last Notable Activity_Page Visited on Website	A
0	0	0	...	0	0	1	0	0	
0	0	0	...	0	0	0	0	0	
1	0	0	...	0	0	0	0	0	
1	0	0	...	0	0	1	0	0	
1	0	0	...	0	0	1	0	0	

Test-Train Split, Scaling, Feature Selection

- Test-Train Split
  - Randon State = 100



TEST-TRAIN SPLIT

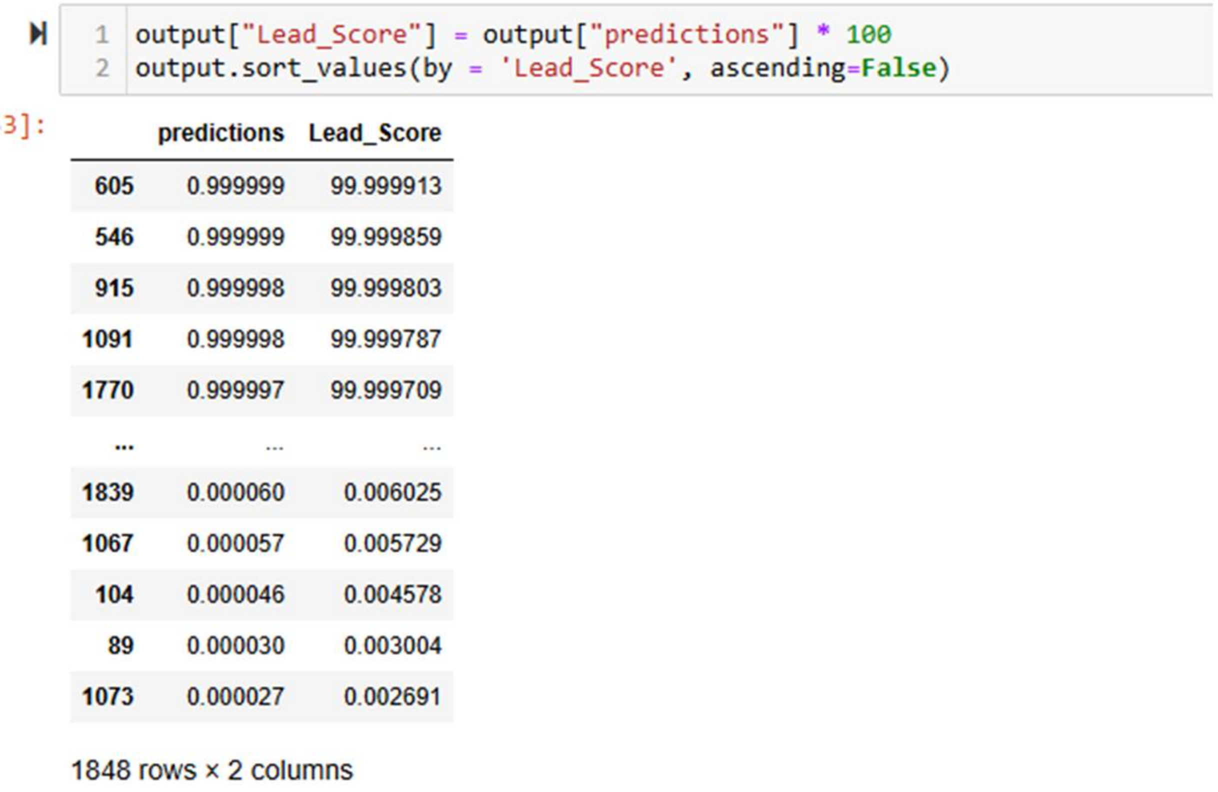
```
X=df_clean_dummies.drop("Converted_1",axis=1)
Y=df_clean_dummies["Converted_1"]
```

```
X_train, X_test, Y_train, Y_test=train_test_split(X,Y,test_size=0.2,random_state=100)
```

- Scale the data using Min-Max Scaler
- RFE
  - Use Recursive Feature Elimination (RFE) to select the most relevant features for a model, going with 20 features

Build and Train the Model

- Use GLM to study the model and observe VIF.
- Predict the Lead_Score using probabilities



```python
output["Lead_Score"] = output["predictions"] * 100
output.sort_values(by = 'Lead_Score', ascending=False)
```

3]:

	predictions	Lead_Score
605	0.999999	99.999913
546	0.999999	99.999859
915	0.999998	99.999803
1091	0.999998	99.999787
1770	0.999997	99.999709
...	...	...
1839	0.000060	0.006025
1067	0.000057	0.005729
104	0.000046	0.004578
89	0.000030	0.003004
1073	0.000027	0.002691

1848 rows × 2 columns

Evaluate the Model

- **Accuracy**
  - Overall Accuracy is 90.75% which is the ratio of correctly predicted instances to the total instances. Hence, overall performance measure of the model.
- **Precision**
  - Precision (also called Positive Predictive Value) measures the accuracy of the positive predictions. Seemingly 92% is high precision.
- **Recall**
  - Recall (also called Sensitivity or True Positive Rate) measures the model's ability to capture positive instances. Seemingly 93% is high recall.

Business Recommendations

1. Focus on high-probability leads from Welingak Website.

2. Prioritize leads engaging via SMS.

3. Reduce effort on invalid or switched-off numbers.

4. Optimize intern efforts for better conversion.

Conclusion

By implementing a lead scoring model, X Education can enhance sales efficiency, prioritize high-converting leads, and improve overall conversion rates.