LEAD SCORING CASE STUDY USING LOGISTIC REGRESSION

SUBMITTED BY AADITYA MAHAMUNI

AGRATA ARYA

AAYUSH SINGH

CONTENTS

- Problem Statement
- Problem Approach
- EDA
- Correlations
- Model Evaluation
- Observations
- Conclusion

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. They provide a form for the potential customers posted on various websites and search engines like google, etc.
- These people (leads) are then contacted by the company employees to be guide, informed about the different courses available on the platform through mails, calls, and other mediums and converted into course takers. The lead conversion rate for X Education is around 30% and they want to increase this using logistic regression to get “Hot Leads” (leads with high chance of conversion).
- If they are successful in doing so their sales would increase by huge margins and simultaneously save a lot of resources by minimizing on expenses as they’d know what to put their focus on exactly. They require a Lead Score assigned to each lead to between 0 and 100 to show their likelihood to become a permanent customer.

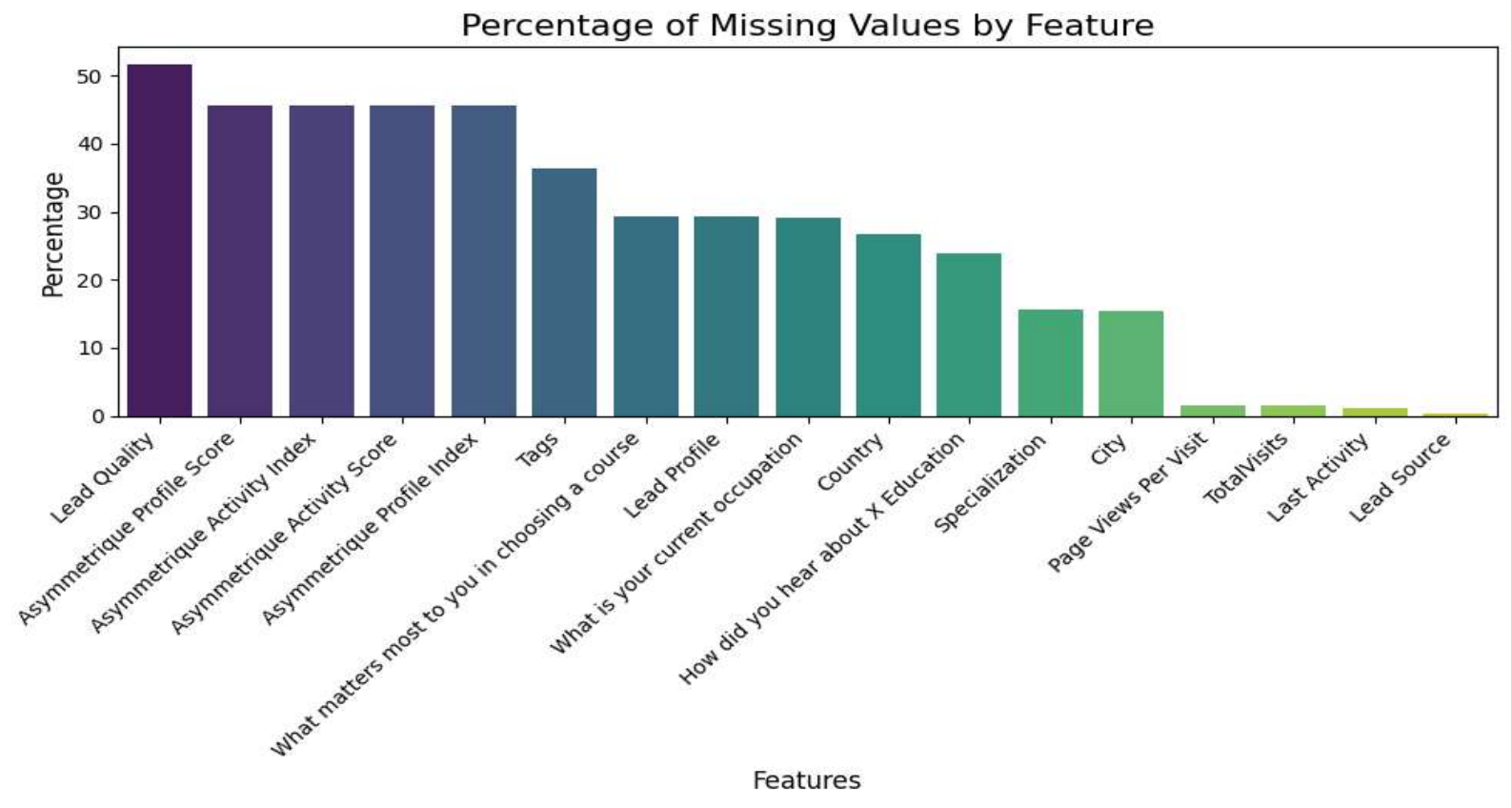
BUSINESS OBJECTIVE

- The company wants us to build a binary logistic regression which will predict potential ‘Hot Leads’ so they can convert them into customers.
- They aim to improve the conversion rate to around 80% after the model development and deployment.
- They would also like to know when they can reallocate their resources to save on employee working hours and other kinds of expenses.

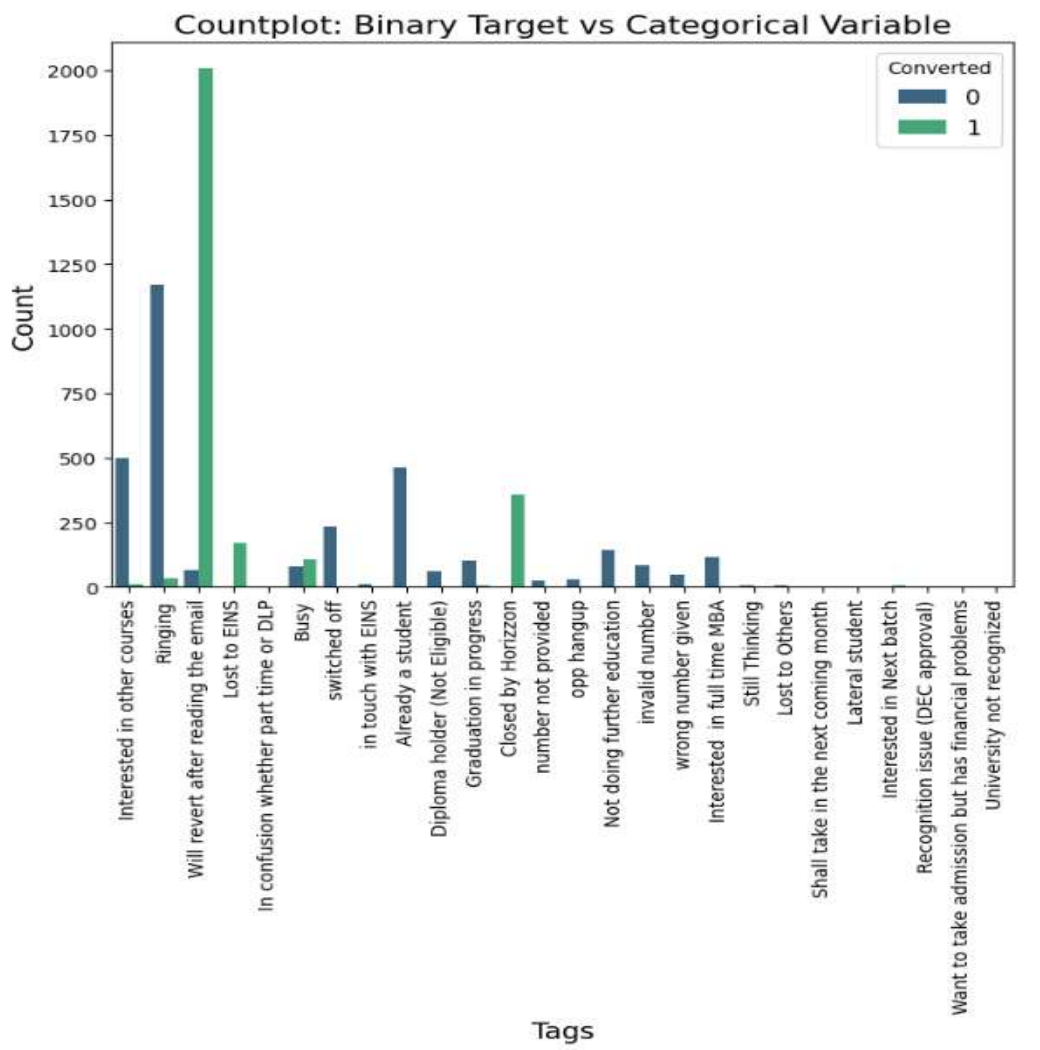
PROBLEM APPROACH

1. Importing the required libraries and packages for data analysis, visualization, model building and evaluation.
2. EDA
3. Train-Test Split
4. Feature Scaling
5. Model Building (VIF and P-value)
6. Evaluation Metrics
7. Predictions on Test Dataset.

EDA- DATA CLEANING



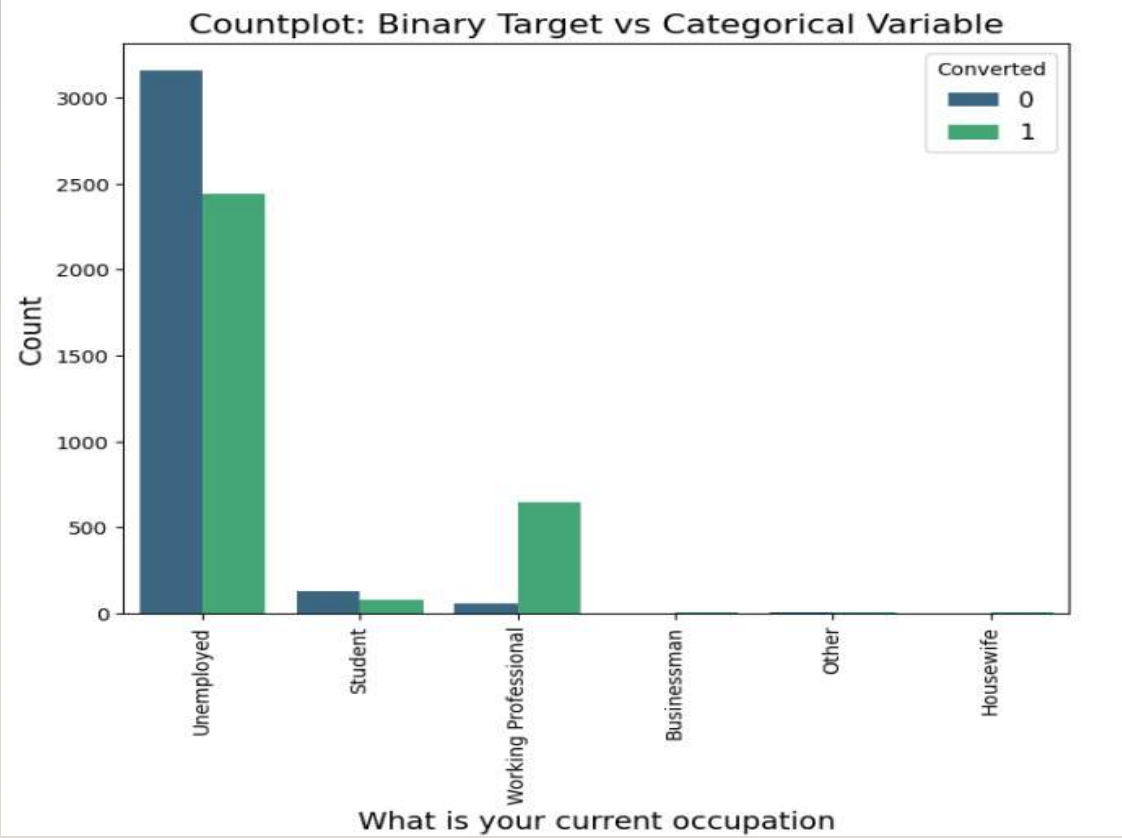
EDA- VISUALIZATION

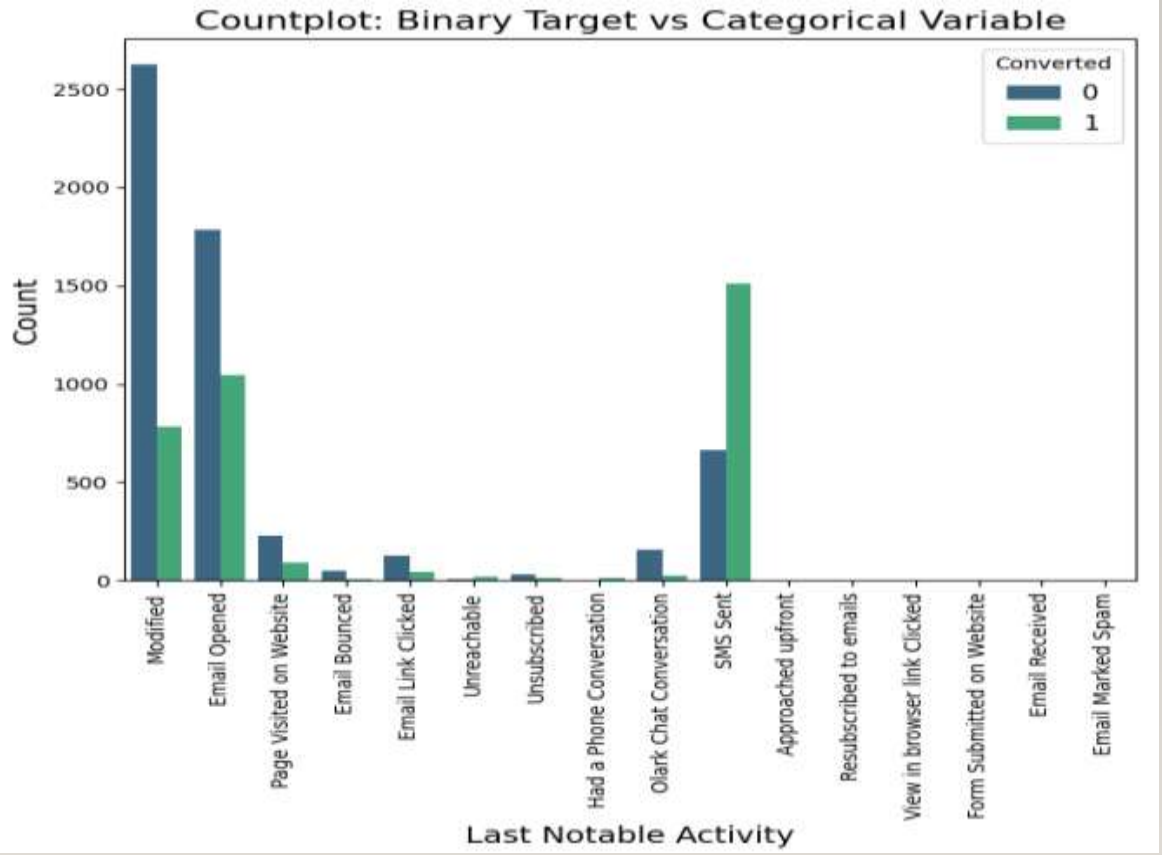


The Tags categorical variable shows highest conversion rate in "Will revert after reading the email" and second most in "Closed by Horizzon" category.

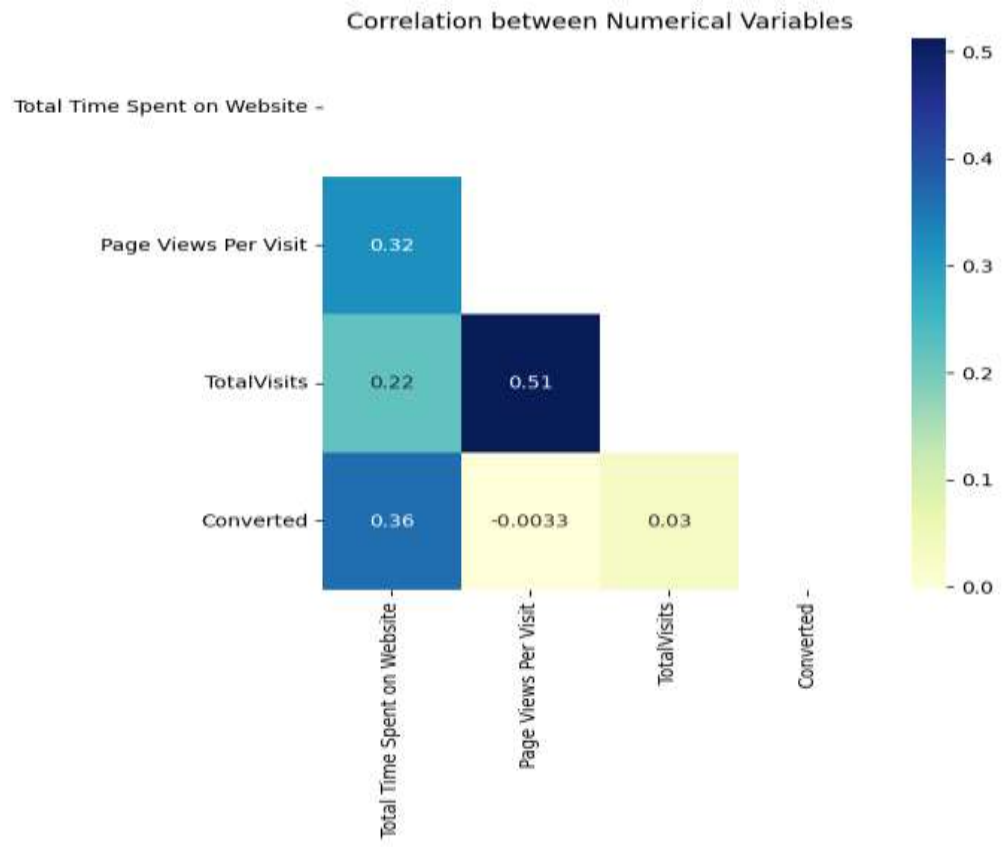
This is one of the top features in our logistic regression lead predictor model showing

EDA (OCCUPATION & LAST NOTABLE ACTIVITY)





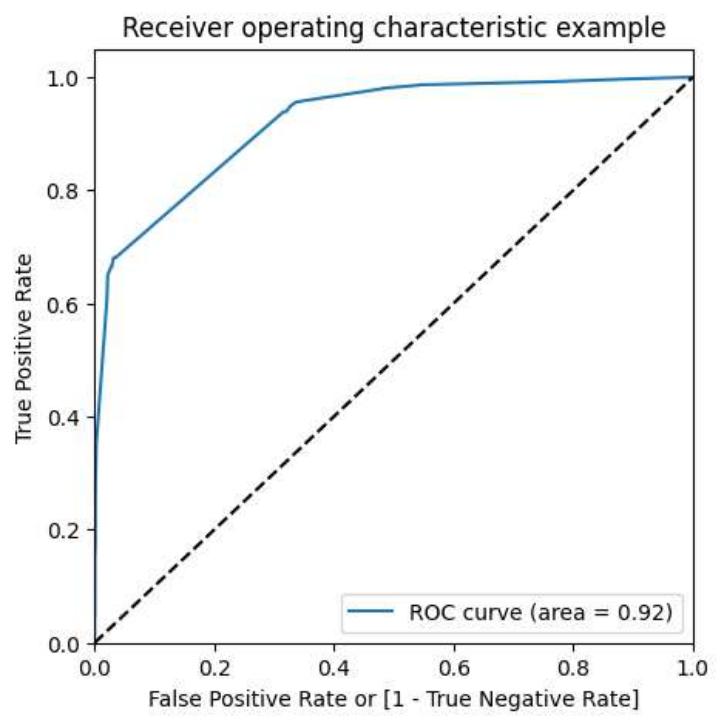
CORRELATION MATRIX FOR NUMERICAL VARIABLES AND TARGET

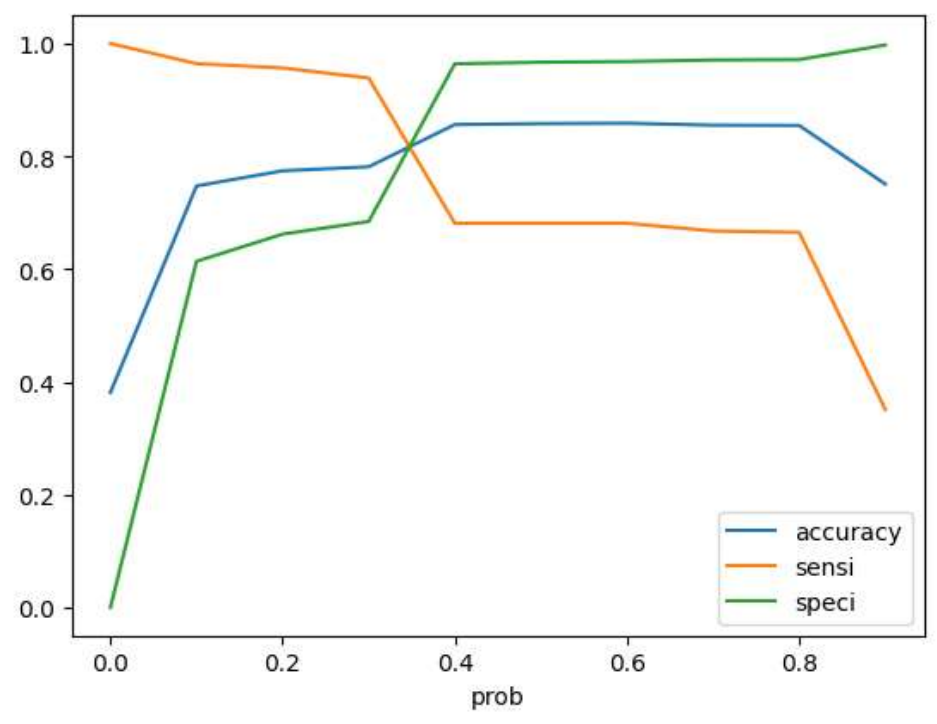


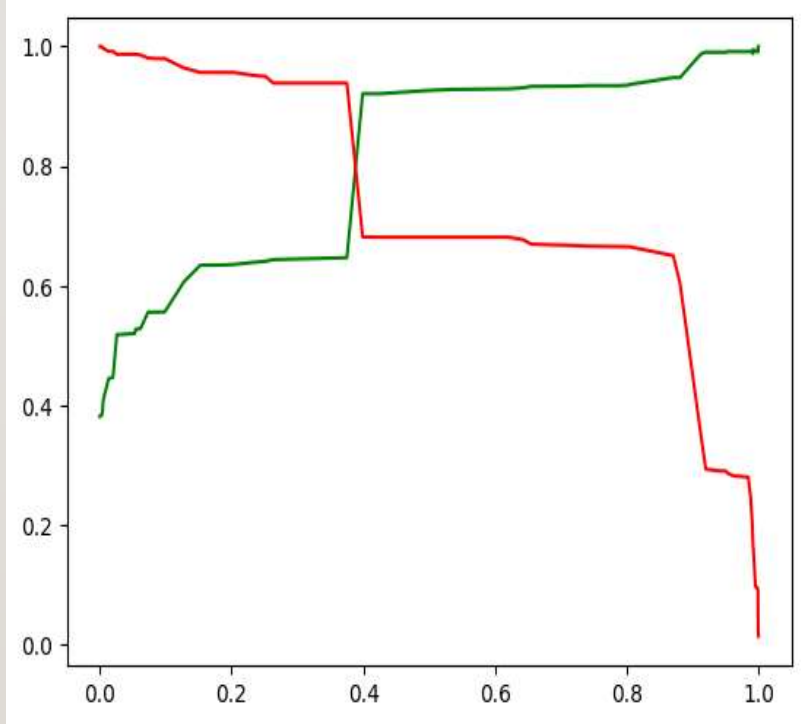
The correlation matrix shows the relation of the target variable ‘Converted’ with other numerical variables like “TotalVisits”, “Total Time Spent on Website”, etc.

We can observe that ‘TotalVisits’ and ‘Page Views Per Visit’ show a little higher correlation with each other at 0.51.

MODEL EVALUATION (ROC, PRECISION-RECALL TRADEOFF)







OBSERVATIONS

Train Data:

- Accuracy : 85%
- Sensitivity : 68%
- Specificity : 96%

Test Data:

- Accuracy : 84%
- Sensitivity : 67%
- Specificity : 96%

Final Columns list:

- Do Not Email
- Lead Origin_Lead Add Form
- Last Activity_Olark Chat Conversation
- Occupation_Working Professional
- Tags_Busy
- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Ringing
- Tags_Will revert after reading the email
- Tags_in touch with EINS
- Tags_switched off
- Last Notable Activity_Had a Phone Conversation
- Last Notable Activity_SMS Sent

CONCLUSION

- We found out that the most important factors which could tell whether a lead would be converted were as follows: -
  a. Tags
  b. Last Notable Activity
  c. When occupation was a working professional
  d. Lead Origin from Lead Add form

  Keeping in mind all this X Education can improve their profit margins significantly while not wasting a lot of resources and efforts on unnecessary, redundant practices which would not really cause a lead to convert in the future.