LOGISTIC REGRESSION – GRADED ASSIGNMENT

LEAD SCORING – LOGISTIC REGRESSION

Bhaveshkumar Thaker

UPGRAD – CU – M.SC. IN DATA SCIENCE

Problem Statement

- X Education sells online courses to industry professionals.
- Many professionals who are interested in the courses land on their website and browse for courses.
- The typical lead conversion rate at X education is around 30%.
- X Education gets a lot of leads, its lead conversion rate is very poor.
- 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:**
  - X education wants to know most promising leads.
  - For that they want to build a Model which identifies the hot leads.
  - Deployment of the model for the future use.

Analysis Approach

- Load and Clean the Data
  - Check for Null values and remove features with all unique values
- Perform Exploratory Data Analysis (EDA)
  - Perform target variable analysis, categorical & numerical variables analysis to understand the data
- Data Preparation
  - Remove outlier, normalize data and split data for train and test the model
- Model Building using Logistic Regression and Recursive Feature Elimination (RFE)
- Model Building using Statsmodel API and Variance Inflation Factor (VIF)
- Model Building using RandomForest Classifier
- Model Evaluation
- Conclusions and provide recommendations

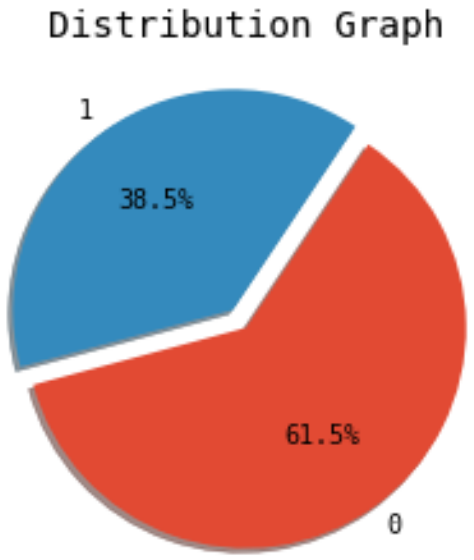
Data Preparation

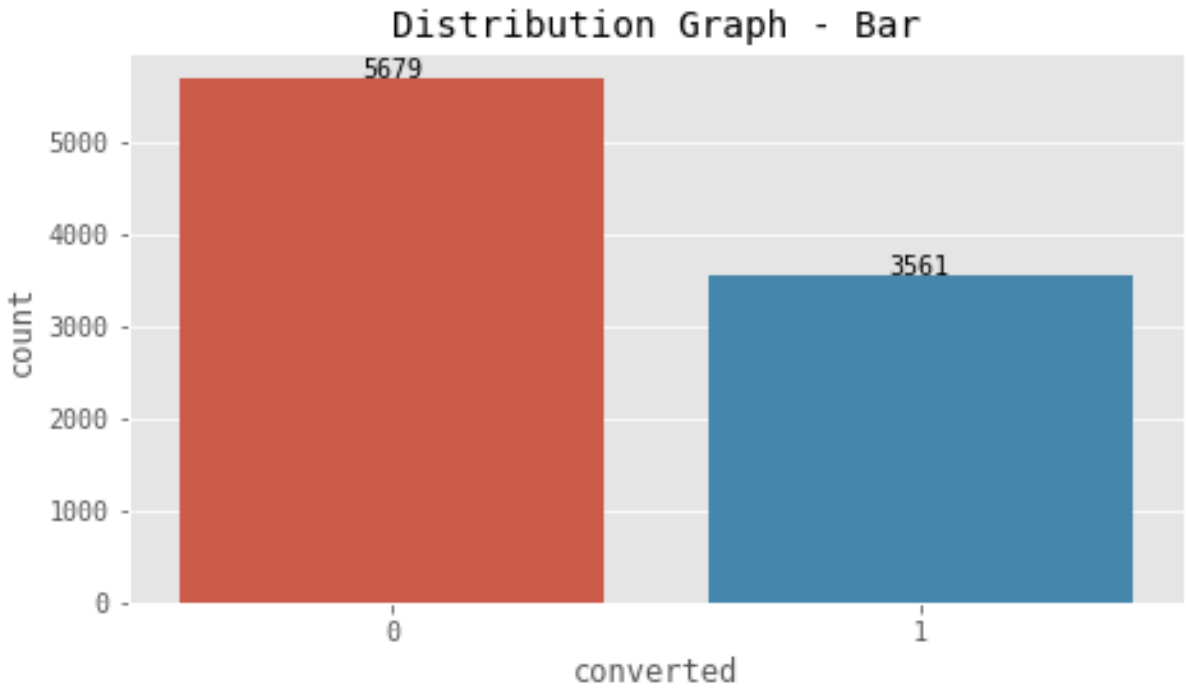
- Replace ”Select” value with NaN in various columns
- Drop features with only one value
  - “magazine”, “receive_more_updates_about_our_courses”, “update_me_on_supply_chain_content”, “get_updates_on_dm_content”, “i_agree_to_pay_the_amount_through_cheque”
- Drop features with all unique values
  - “Prospect Id”, “Lead Number”
- Drop columns with more than 40% null values
- Combine various values of features which are less than 2 percent.
- Drop highly correlated features
- Perform Outlier analysis and remove outliers
- Create dummy (one hot encode) variables
- Apply normalization on the dataset

VISUALIZATIONS

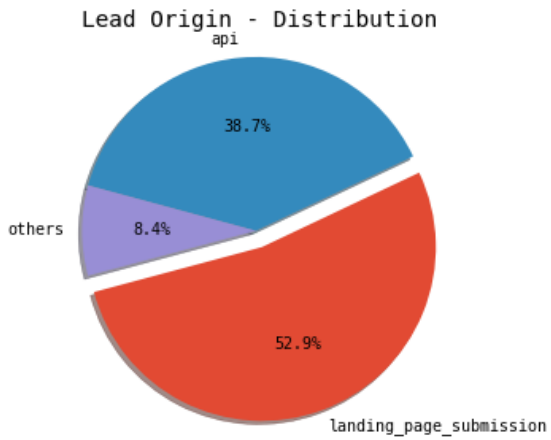
Summary

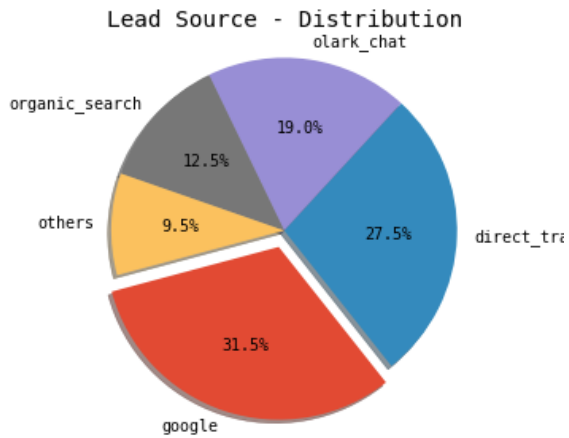
“converted” – Target Variable

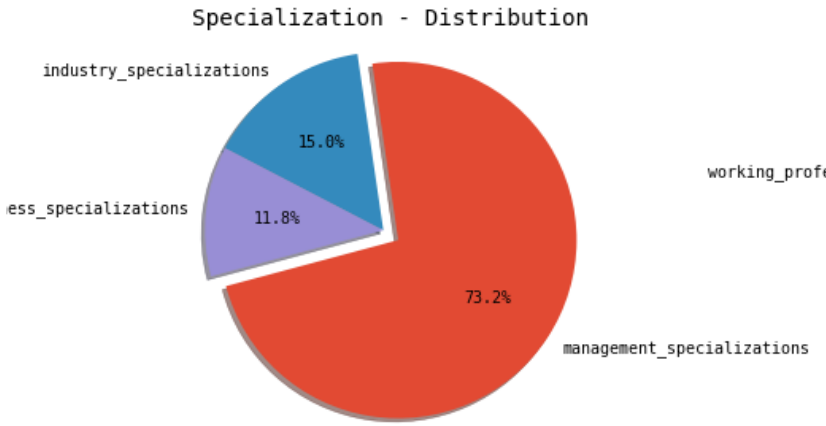


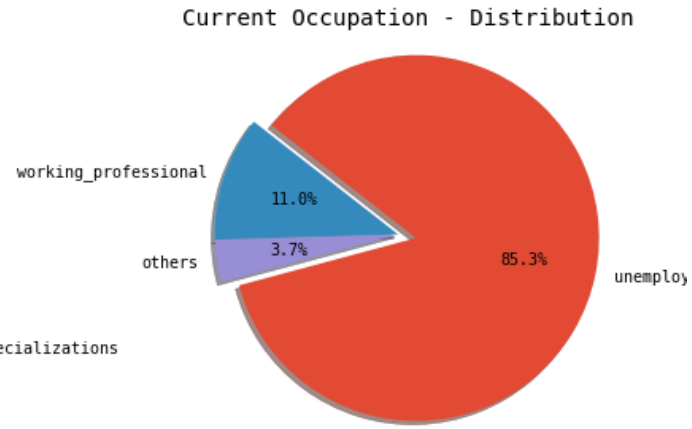


Categorical Distributions

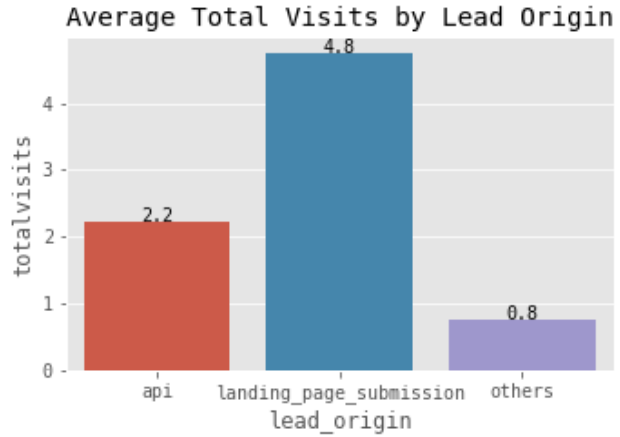


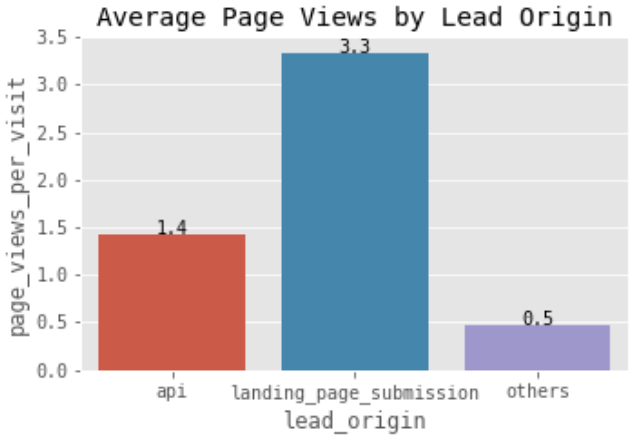


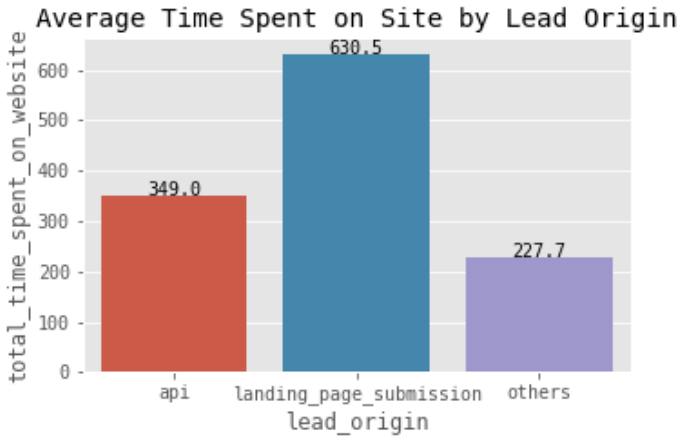




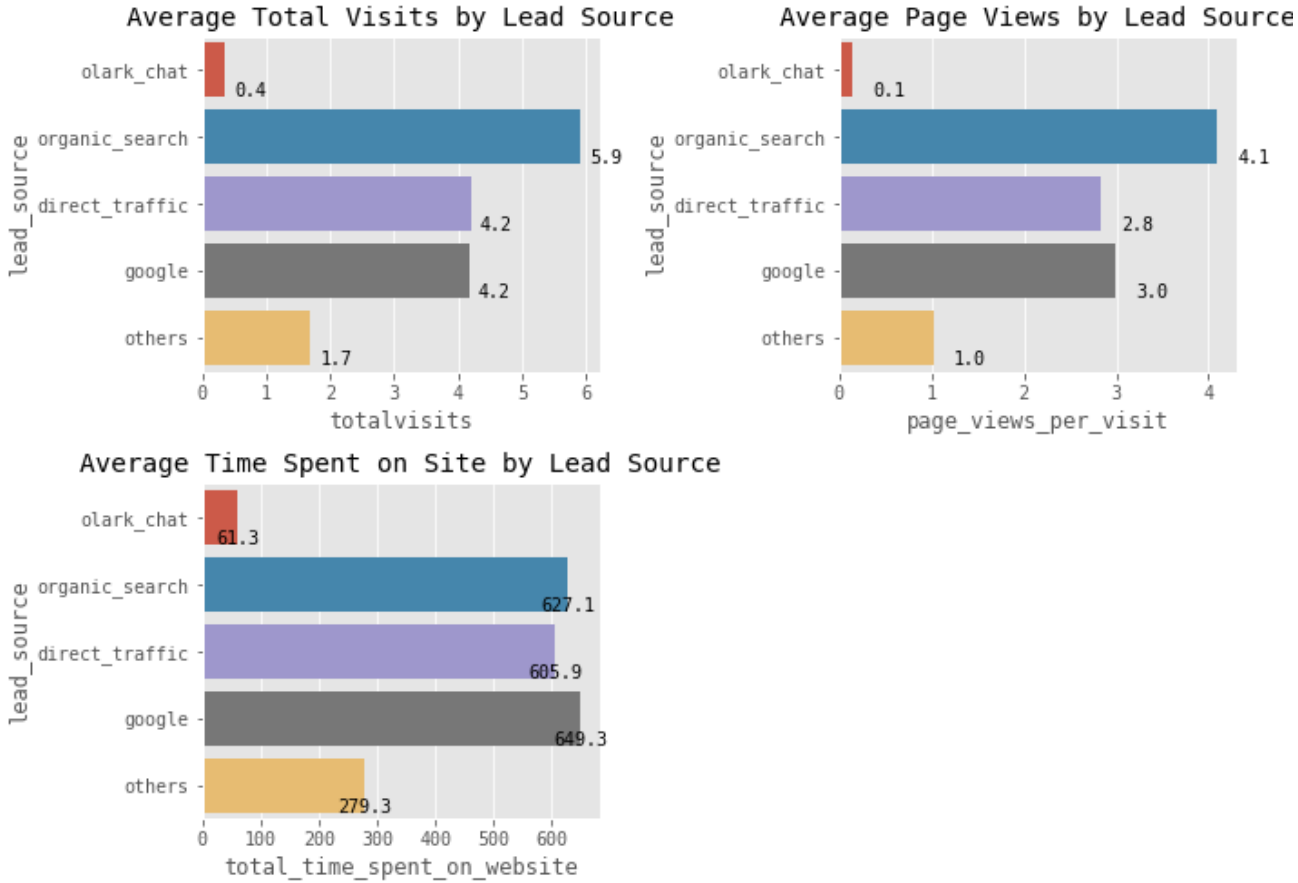
Lead Origin Analysis



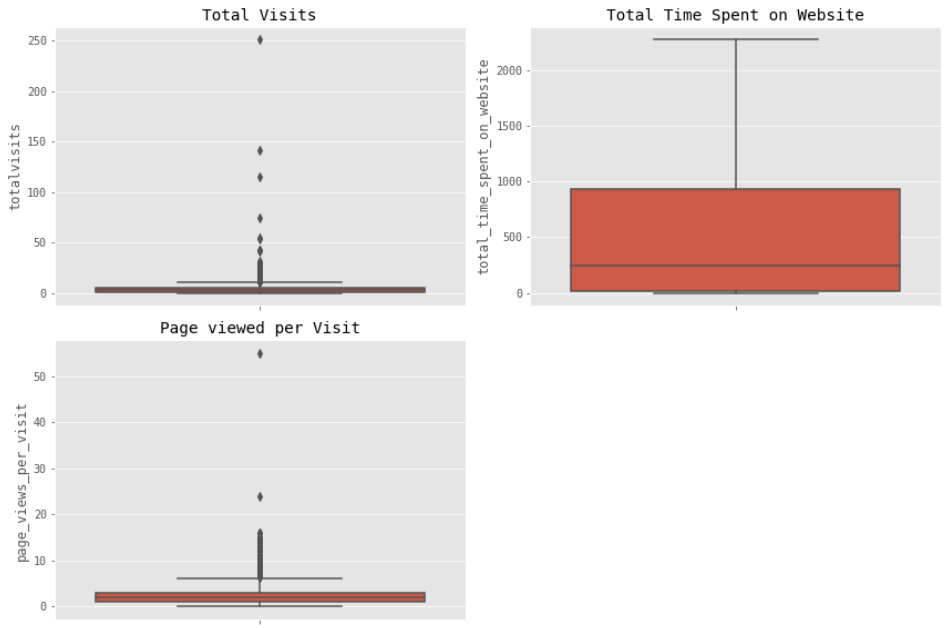


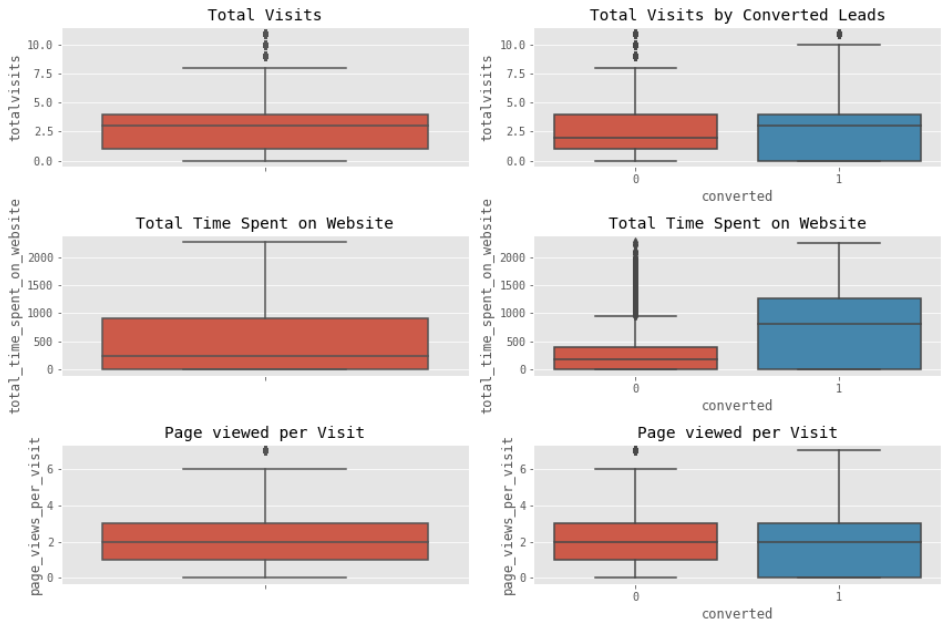


Lead Source Analysis

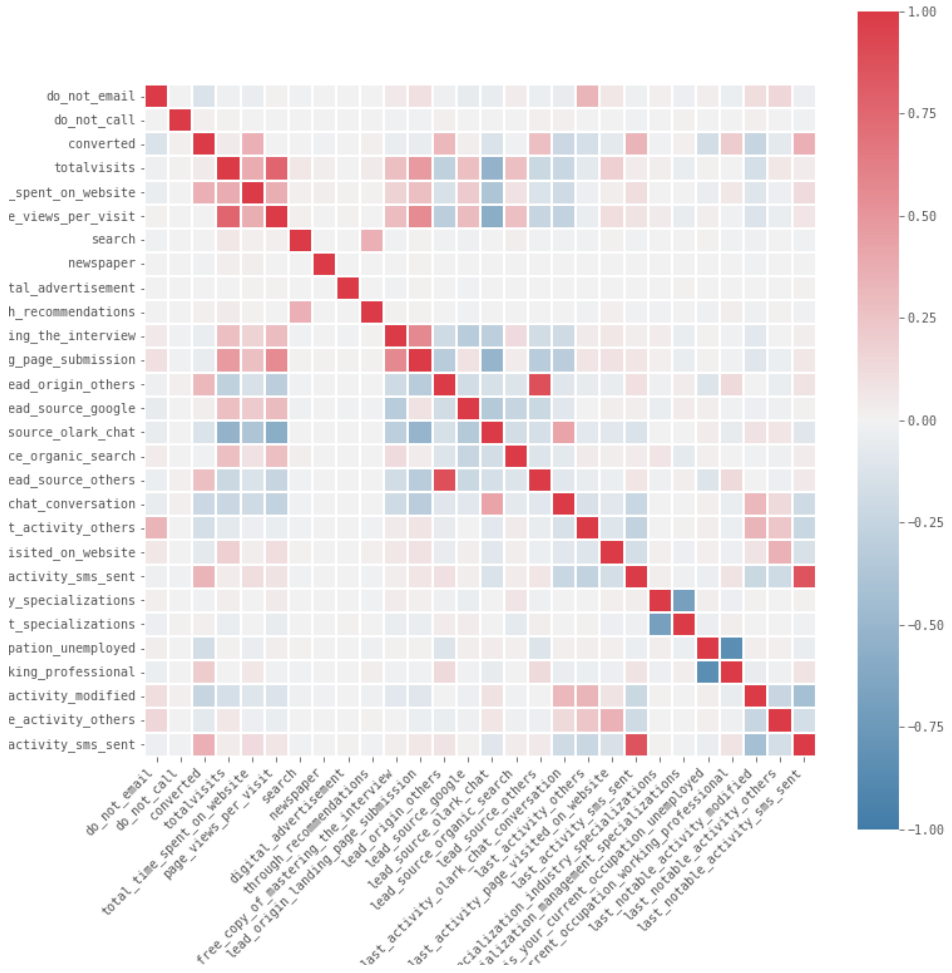


Outliers Treatment

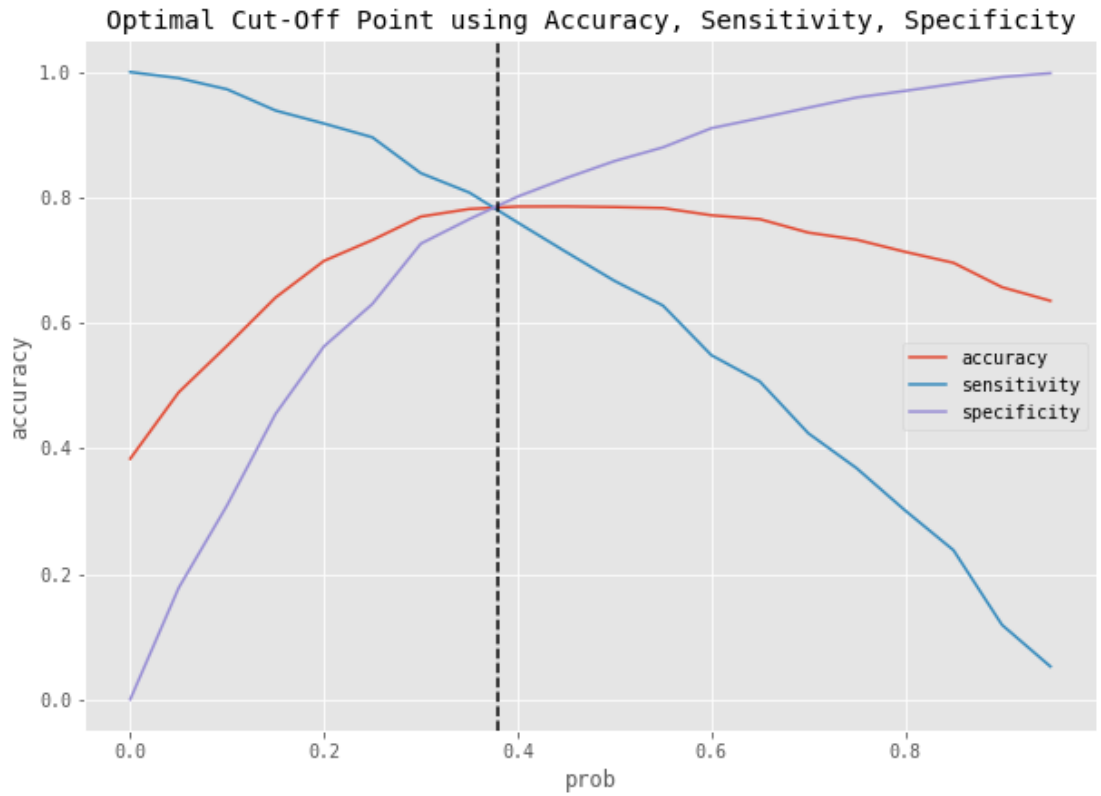




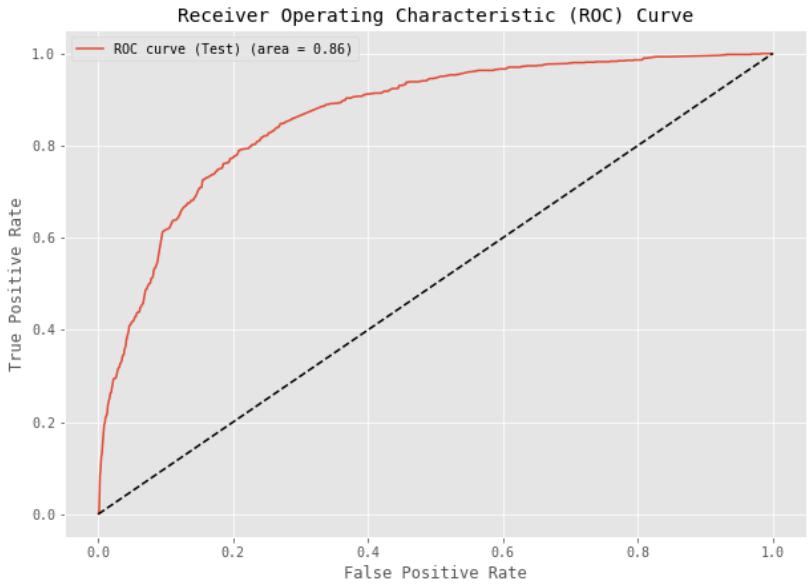
Dataset Correlations

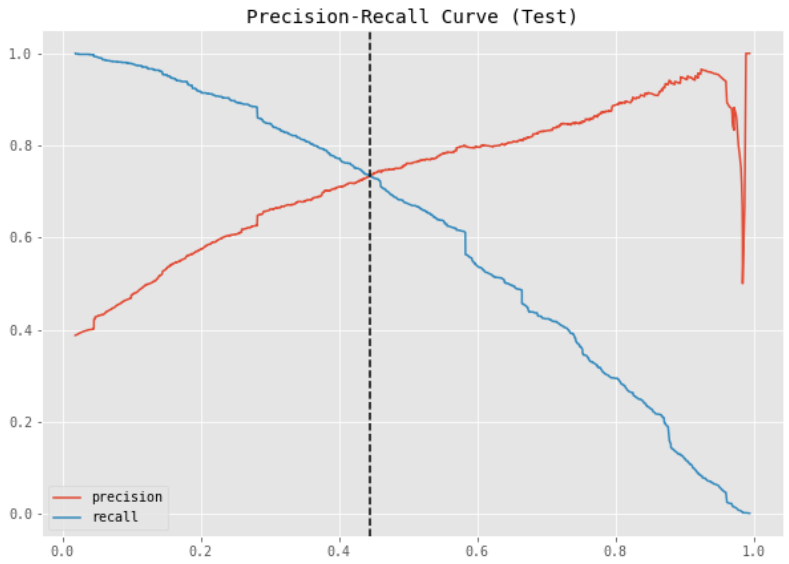


Optimal Cut-Off Point



ROC and Precision-Recall Curve (Test)





SUMMARY

Summary

- To increase the probability of the lead conversion, X Education's sales team should focus on following
  - Target leads which spends a lot of time on X Education website. Make the website more interactive, informative, and engaging.
  - Target leads which has come through recommendations. The X Education sales team should make phone calls to these leads and should explain competitive points why X Education is better than others.
  - Target leads which has come through various sources like Olark Chat, Organic Search, Google search.
  - Target leads which are Working Professionals and explain them specializations and industry readiness offered by the X Education courses.
  - Sales team should give digital advertisements to create potential leads and convert them.
  - Sales team should give advertisements and show banners on search engines like Google, Bing and other search engines to create potential leads and convert them.