Lead Scoring Case Study

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.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead
- Company devote resources on this lead through email, sms and call and try to convert lead to success. We intend to make this conversion with as optimised way possible by making that lead first which has high chances of getting converted.
- Problem is also with very low efficiency of conversion of lead which is 30% from all the leads

BUSINESS OBJECTIVE

- Prioritizing leads that require less time and resources; and for doing this we need to be able to clearly differentiate between hot lead (potential conversion) and normal lead (very low chances of conversion)
- For this create a model for sales team to make this classification of leads efficient and faster.
- Providing the insights and model to improve the conversion rate from 30% to much higher with low resources used possible

SOLUTION METHODOLOGY

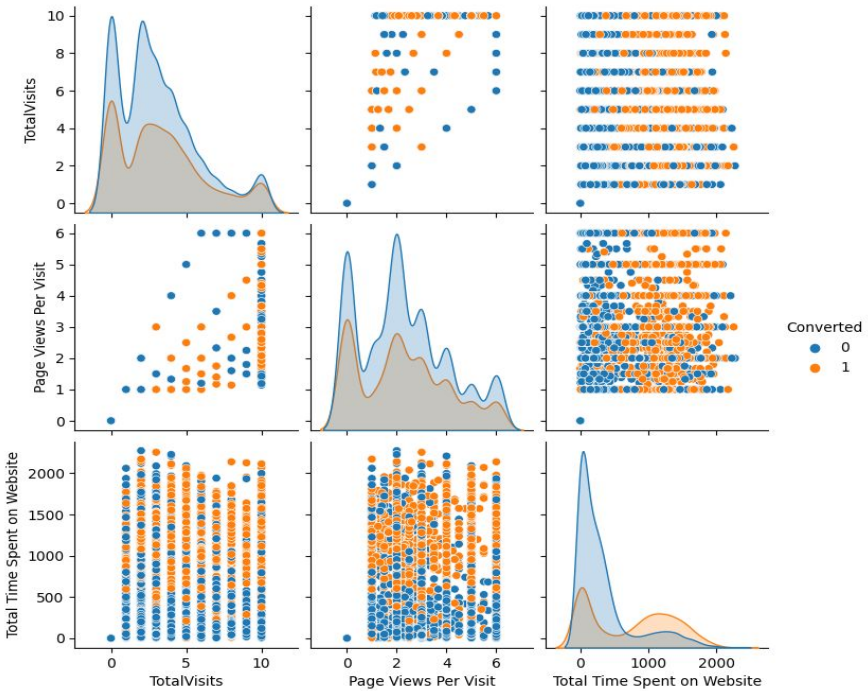
- Data cleaning and data manipulation
  - Check and handle duplicate data
  - Check and handle missing values
  - Drop columns, if it contains a large number of missing values and are not useful for the analysis
  - Handling skewed columns meanwhile imputation of the values, if necessary
  - Check and handle outliers in data

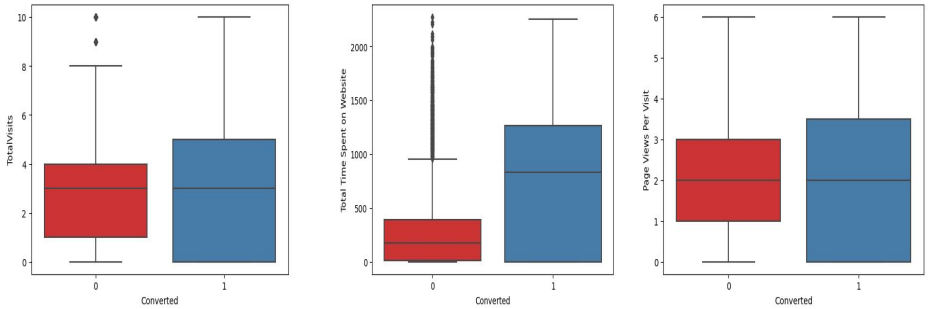
- Exploratory Data Analysis (EDA)
  - Univariate data analysis: value count, distribution of variables, etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
  - Feature Scaling & Dummy variables and encoding of the data
  - Classification technique: logistic regression is used for model making and prediction.
  - Validation of the model
  - Model presentation.
  - Conclusions and recommendations.

DATA MANIPULATION

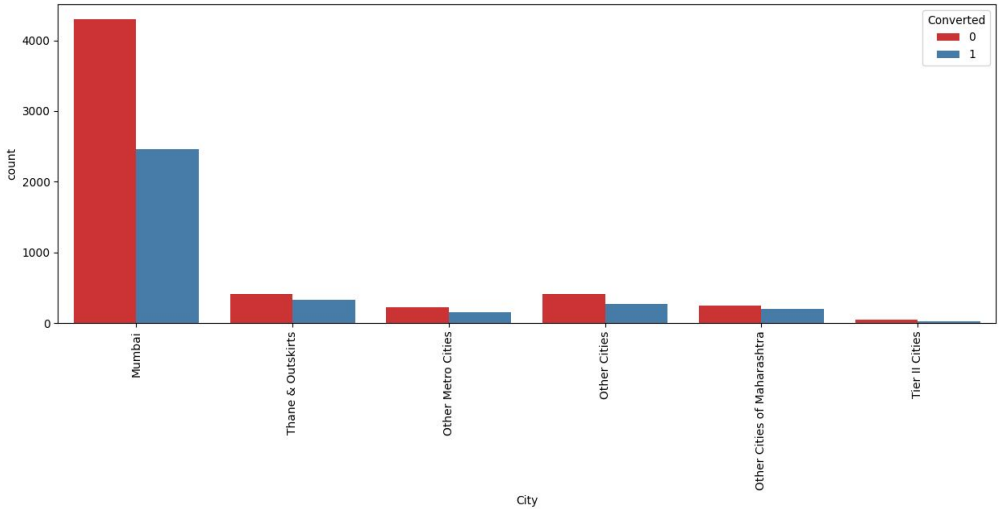
- Total Number of columns are 37 and total Number of rows =9240.
- Dropping single value features; columns that have only one value output
- Removing serial id columns - “ProspectID” and “Lead Number”
- Replacing the ‘Select’ with nan as given in problem description
- Imputing missing values and nan values with mean and mode as per dtype of the column
- Dropping the columns having more than 40% as missing values.
- There was high skewness in some columns so they were also removed from dataset.

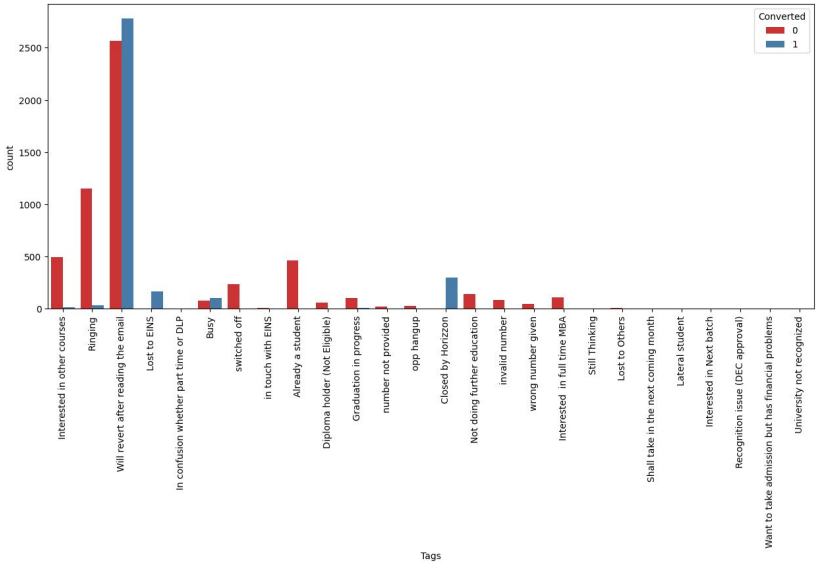
EDA PLOTS

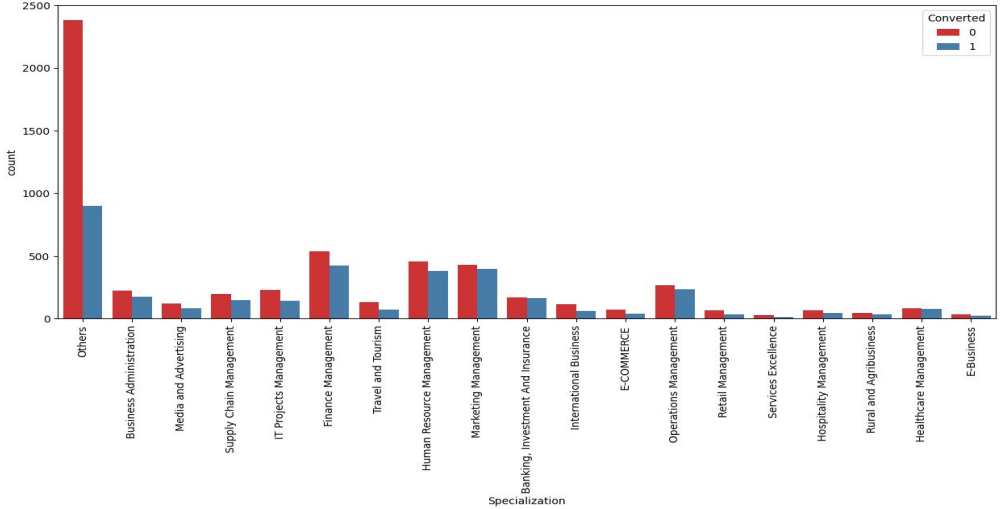


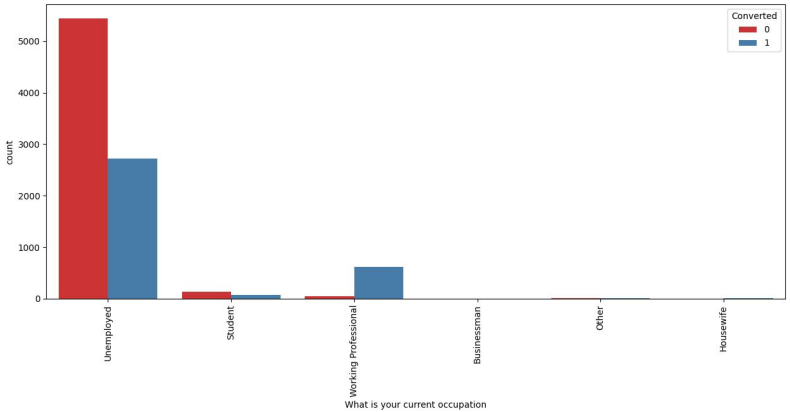


EDA performed using box plots & pair plots for continuous columns and countplot and bar graphs for categorical variables. Heatmap are also used to analyze correlation between column to remove multicollinearity.


Converted
0
1
count
4000
3000
2000
1000
0
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
City


Converted
0
1
count
2500
2000
1500
1000
500
0
Interested in other courses
Ringing
Will revert after reading the email
Lost to EINS
In confusion whether part time or DLP
Busy
switched off
in touch with EINS
Already a student
Diploma holder (Not Eligible)
Graduation in progress
number not provided
opp hangup
Closed by Horizzon
Not doing further education
invalid number
wrong number given
Interested in full time MBA
Still Thinking
Lost to Others
Shall take in the next coming month
Lateral student
Interested in Next batch
Recognition issue (DEC approval)
Want to take admission but has financial problems
University not recognized
Tags


Converted
0
1
count
2500
2000
1500
1000
500
0
Others
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


Converted
0
1
count
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

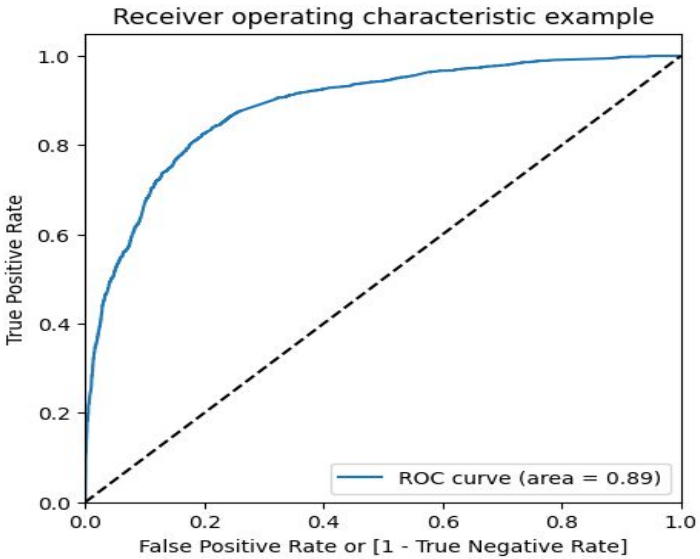
DATA CONVERSION

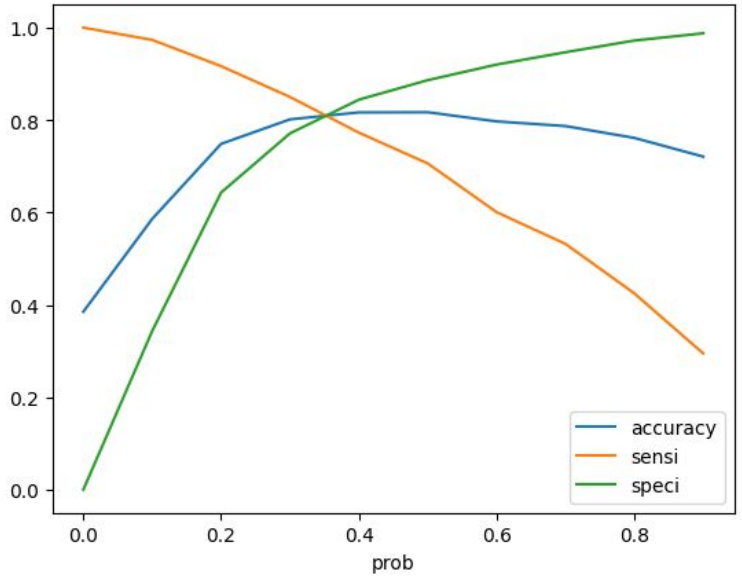
- Numerical Variables are normalized
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9240
- Total Columns for Analysis: 37

MODEL BUILDING

- Splitting the Data into Training and Testing Sets
- train_size=0.7, test_size=0.3 and random_state=100 ; ratio of split is 70% Train and 30% Test
- Use RFE for Feature Selection with 20 variables as n_features_to_select
- Building Model by removing the variable whose p-value is greater than 0.05 and vi value is greater than 5
- Predictions on test data set
- All accuracy metrics are around ~80%

ROC Curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.89)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensi
speci
prob
0.0
0.2
0.4
0.6
0.8
1.0

PREDICTION ON TEST SET

- Before predicting on the test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set, and the new prediction values were saved in a new data frame.
- After this we did model evaluation i.e. finding the accuracy, precision, and recall
- The accuracy score we found was Accuracy : 80.4 %, Sensitivity : 70.6 % , Specificity : 80.5 %
- This shows that our test prediction is having accuracy, precision, and recall scores in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of conversion, low the lead score lower the chance of getting converted

CONCLUSION & RECOMMENDATION

- It was found that the variables that mattered the most in the potential buyers are :
  - Lead Source_Welingak Website
  - Current_occupation_Working Professional
  - More time on website
- More budget/spend can be done on Welingkar Website in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage to provide more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.