X Education - Lead Scoring Case Study

Detection of Hot leads to improve conversion rate for X Eduacation

Team Members:
Manjunath Kannavalli,
Vaibhav Sahal,
Juber Nagani

Important sections:

1. Background of X Education co.
2. Problem Statement and Objective.
3. Suggested Ideas
4. Analysis
5. Data Cleaning
6. EDA
7. Data Preparation
8. Model Building
9. Evaluation
10. Recommendations

Background of X co.

- X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- People who are interested in the courses land on the website and browse for courses.
- People might fill up a form for the course or watch some videos.
- These people become leads when they provide their email address or phone number.
- The company also gets leads through past referrals.
- Employees from the sales team start making calls, writing emails, etc., to convert the leads.
- The typical lead conversion rate at X Education is around 30%.

Problem Statement and Objective

Problem Statement

- X Education gets a lot of leads, its lead conversion rate is at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

Objective of Case Study

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Suggested Ideas

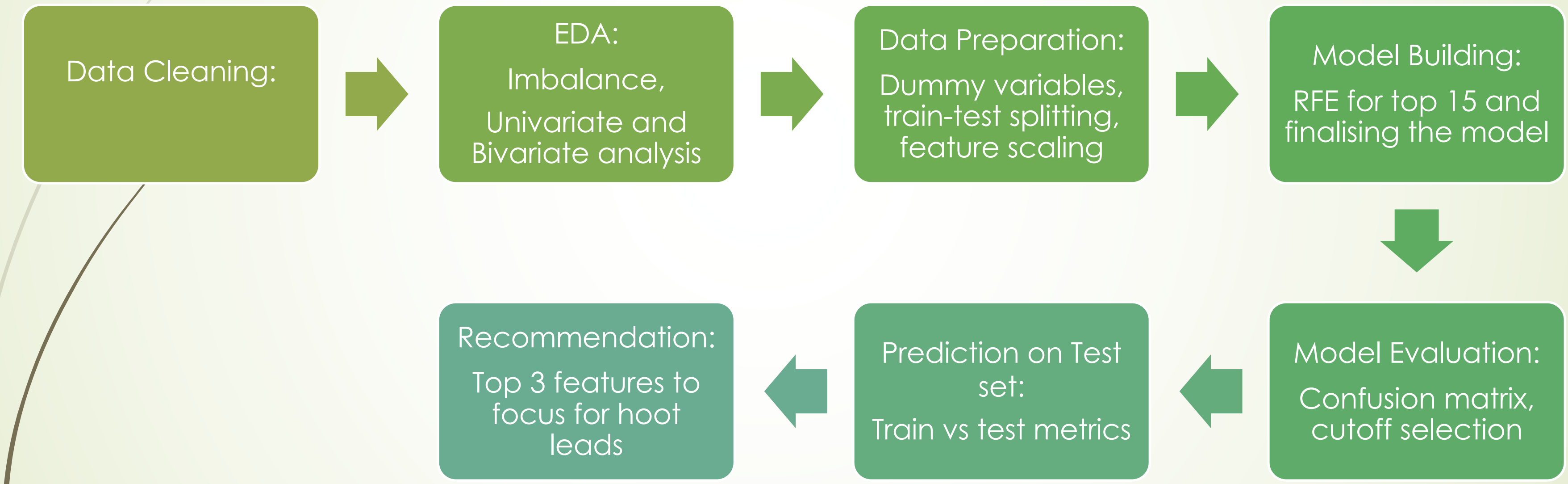
Grouping

- Groups are grouped or classified into probability of conversion
- Efforts are focused to hot leads

Boost Conversion

- We would have a greater conversion rate and be able to hit the 80% objective since we concentrated on hot leads that were more likely to convert.

Analysis Approach



Data Cleaning

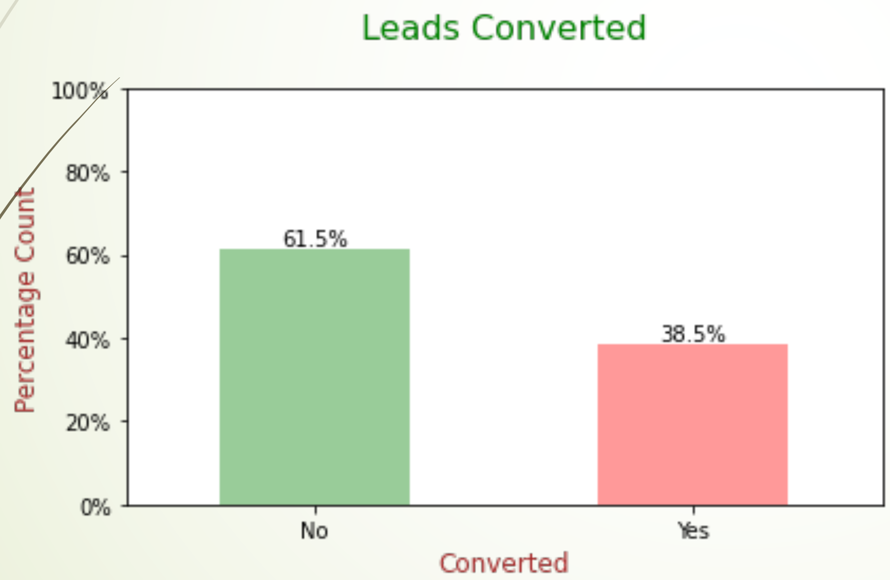
- **"Select"** level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective (tags, country)
- Imputation was used for some categorical variables.
- Additional categories were created for some variables.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution.

Data Cleaning

- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in **Total Visits** and **Page Views Per Visit** were treated and capped.
- Invalid values were fixed and data was standardized in some columns, such as lead source.
- Low frequency values were grouped together to “Others”.
- Binary categorical variables were mapped.
- Other cleaning activities were performed to ensure data quality and accuracy.
- Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc. (lead source has Google, google)

EDA

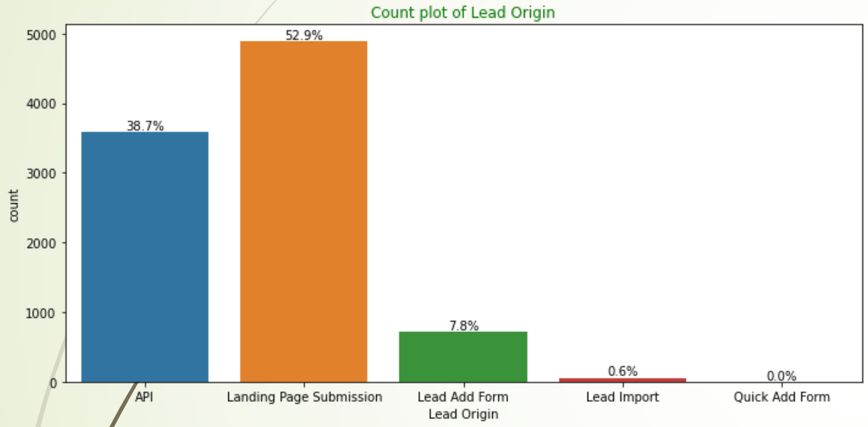
- Data has the imbalance with leads conversion criteria

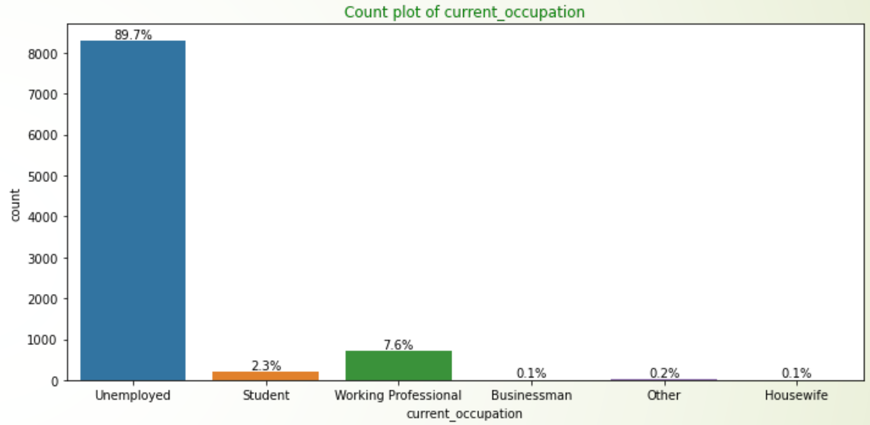


- Conversion rate is only 38.5%

EDA

- Univariate Analysis- categorical

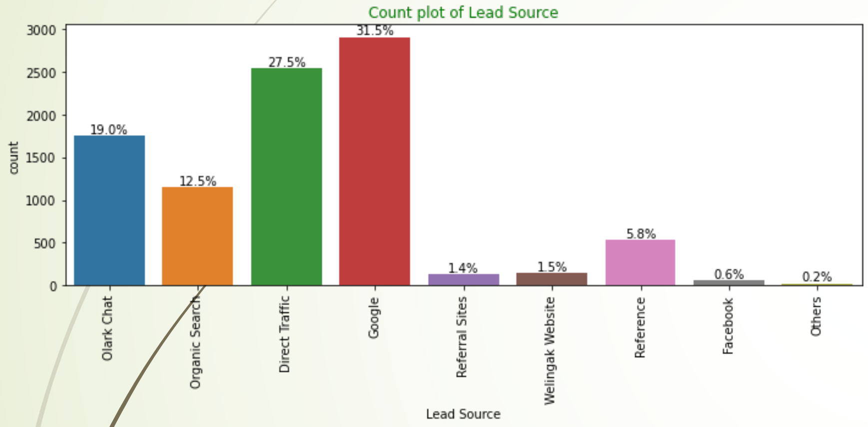


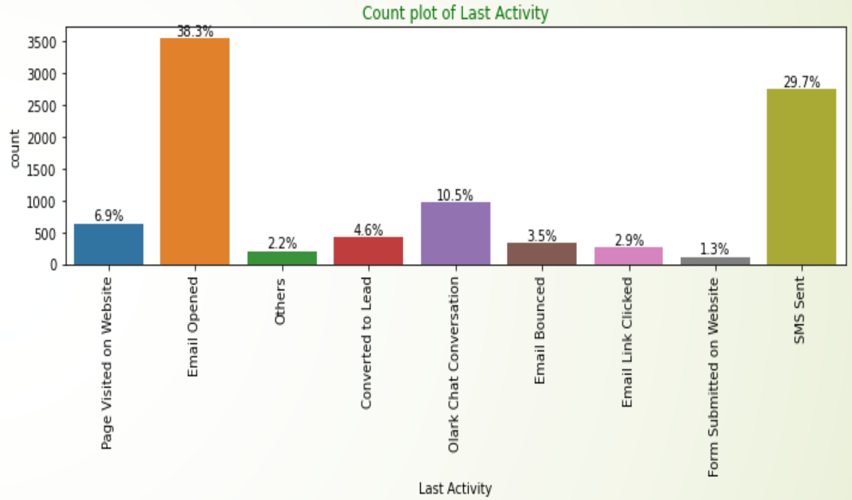


- **Lead Origin:** Landing page submission identified 53% of its customers, and API: 39%
- **Current_occupation:** It has 90% of the customers as Unemployed.

EDA

- Univariate Analysis- categorical

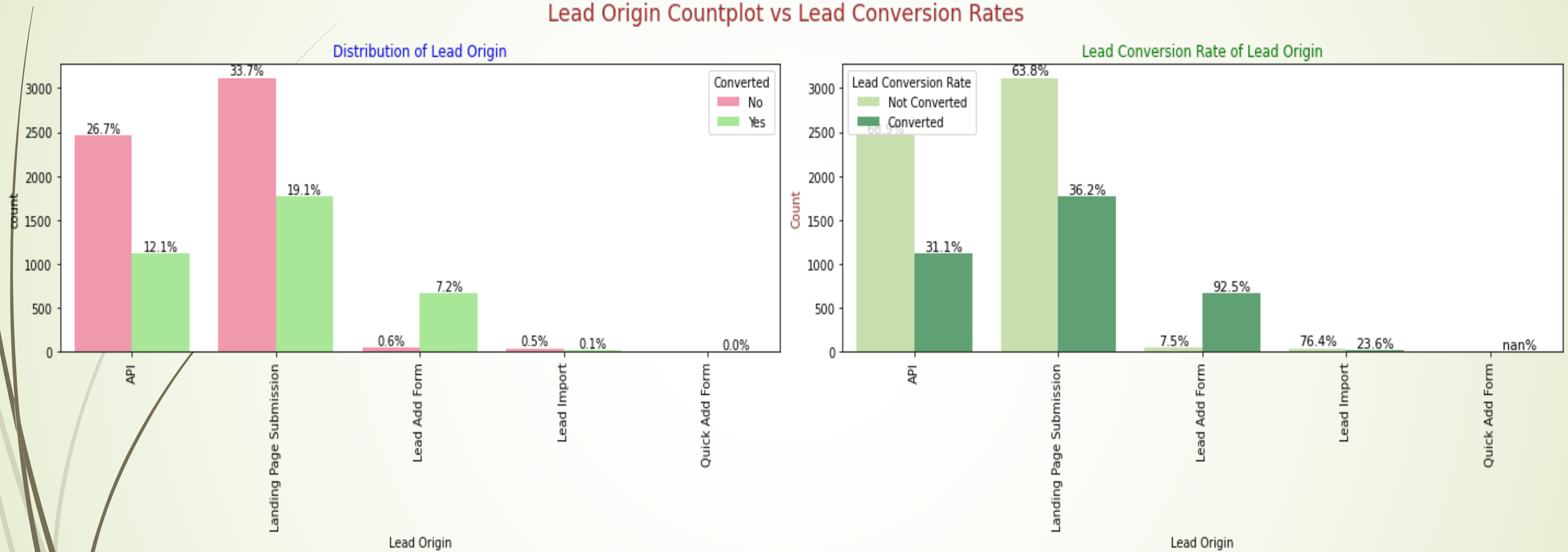




- **Lead Source:** Google and Direct traffic make up 58% of lead source

- **Last Activity:** SMS sent and Email Opened are two major categories for Leads

EDA- Bivariate for categorical

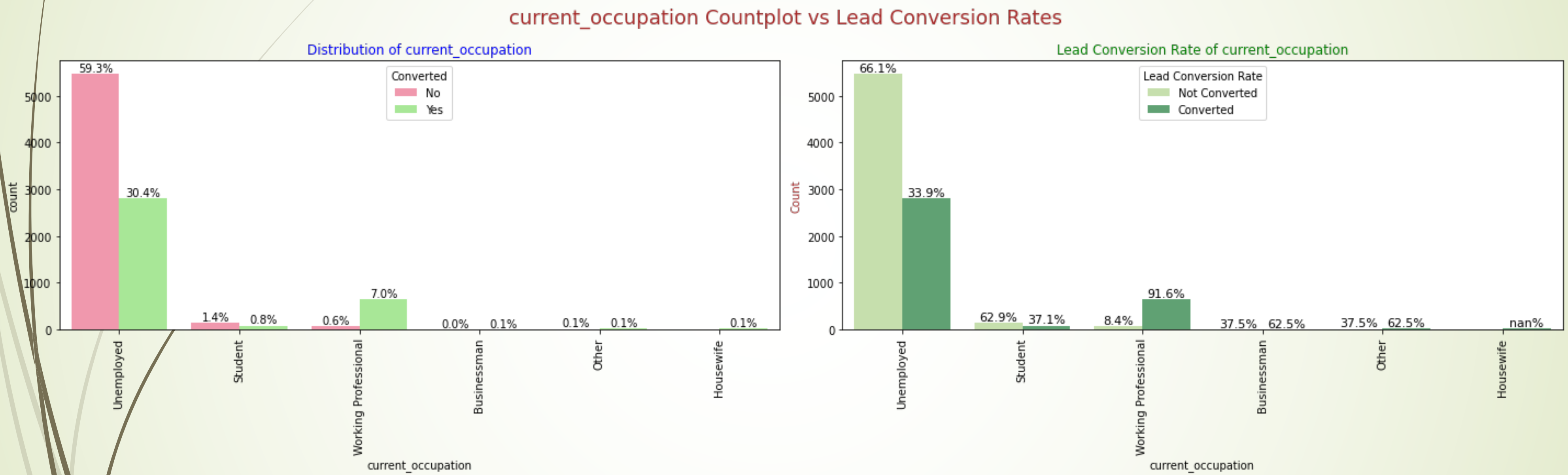


Lead Origin:

•Around 52% of all leads originated from "*Landing Page Submission*" with a **lead conversion rate (LCR) of 36%**.

• The "*API*" identified approximately 39% of customers with a **lead conversion rate (LCR) of 31%**.

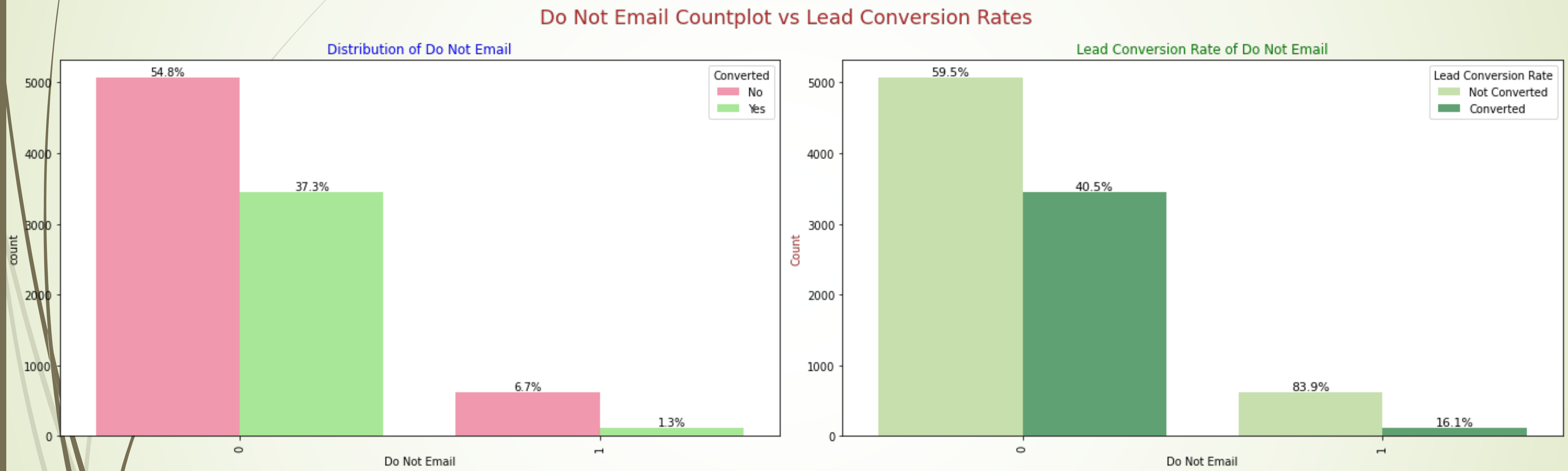
EDA- Bivariate for categorical



Current_occupation:

- Around 90% of the customers are *Unemployed,* with **lead conversion rate (LCR) of 34%.**
- While *Working Professional* contribute only 7.6% of total customers with almost **92% Lead conversion rate (LCR)**

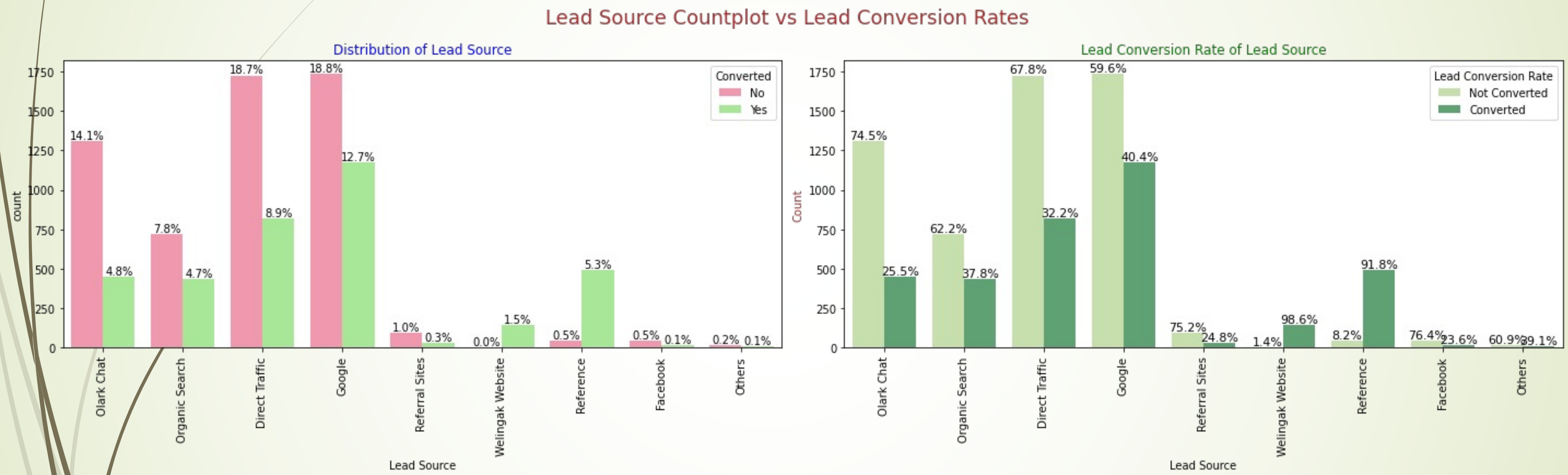
EDA- Bivariate for categorical



Do Not Email:

● 92% of the people has opted that they don't want to be emailed about the course & 40% of them are converted to leads.

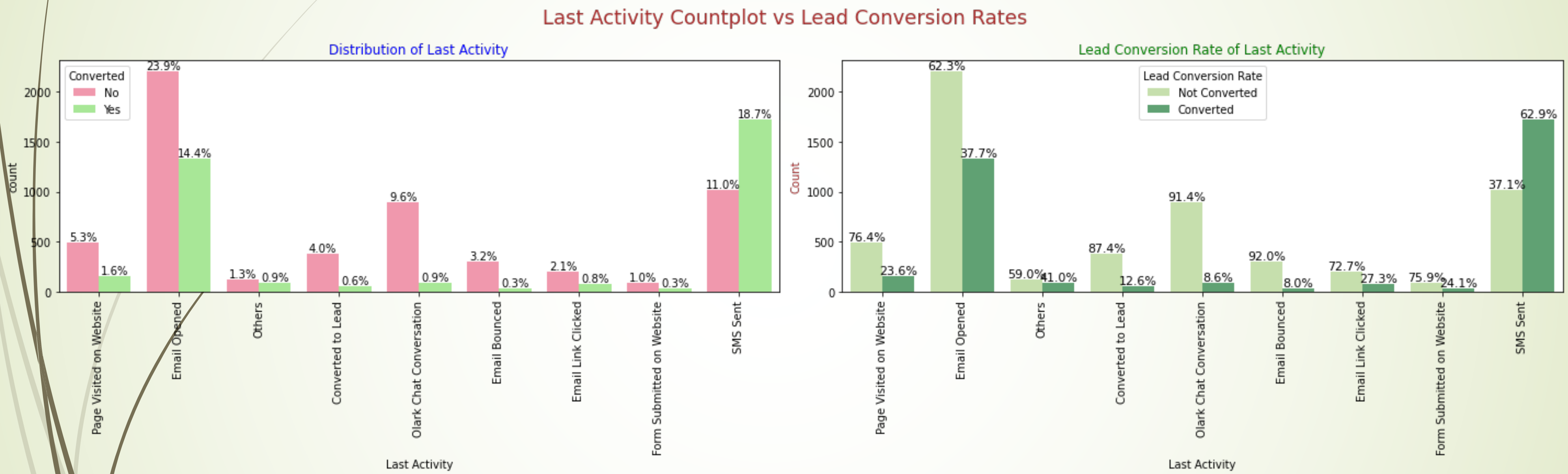
EDA- Bivariate for categorical



Lead Source:

- Google and Direct traffic are main sources for lead conversion
- Organic search has 37% of lead conversion
- Reference has 91% of success from 6% of the total customers

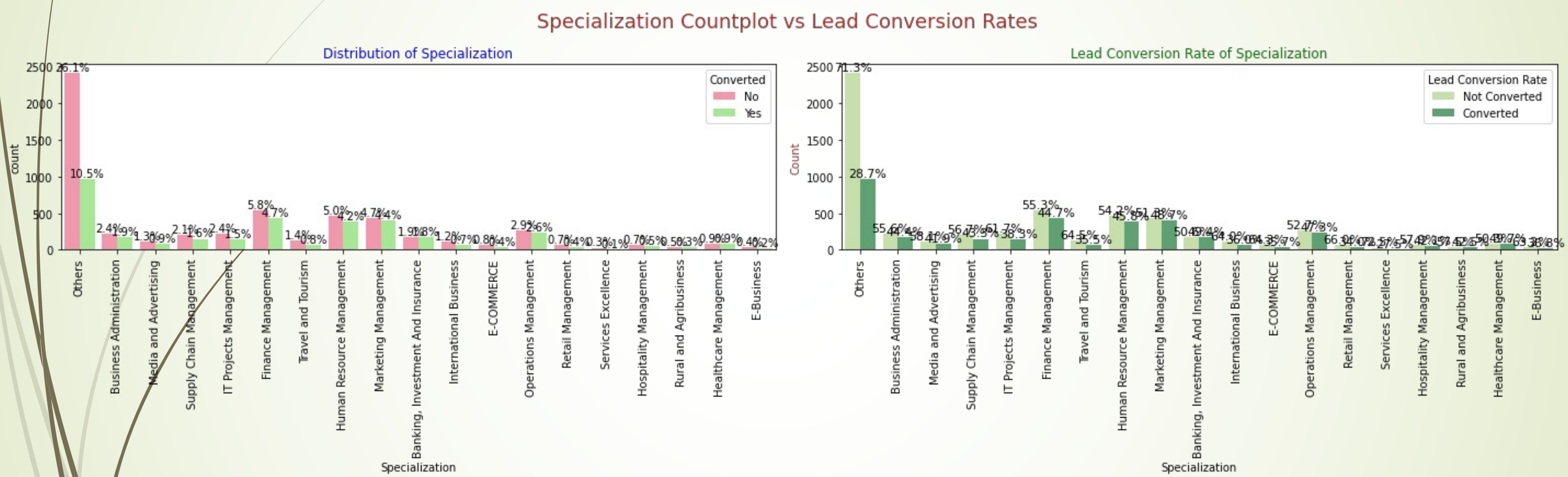
EDA- Bivariate for categorical



Last Activity:

- 'SMS sent' has high conversion rate of 63%.
- Email Opened has 38% of successful conversion rate

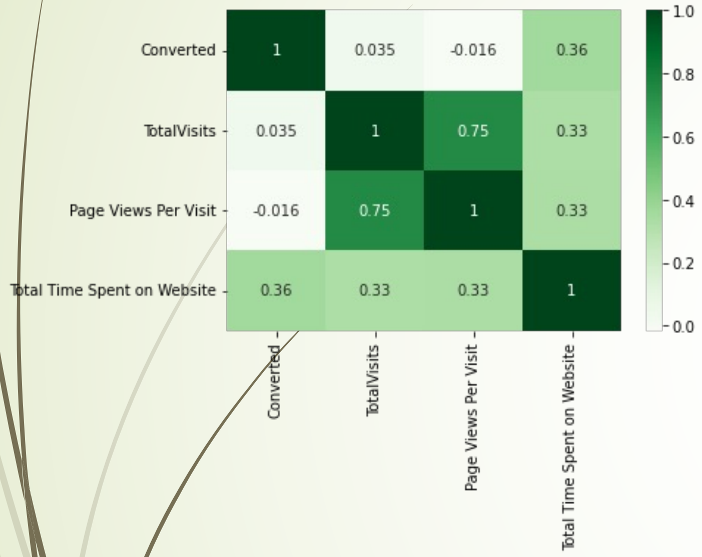
EDA- Bivariate for categorical

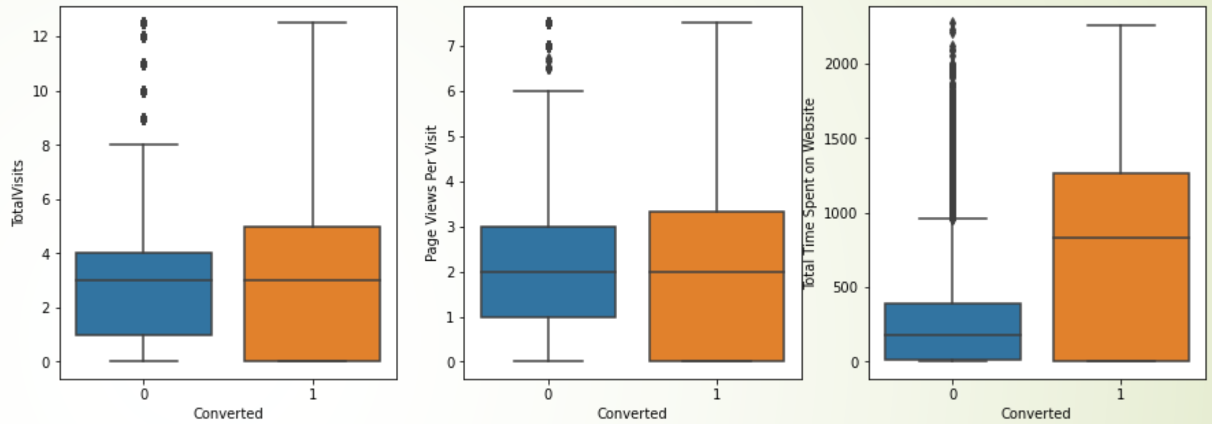


Specialization:

● Marketing Management, HR Management, Finance Management shows good contribution in Leads conversion than other specialization.

EDA- Bivariate for numerical





Past leads spending more time on the website have the chance for successful LCR

Data Preparation

- Binary level categorical columns were already mapped to 1 / 0 in previous steps
- Created dummy features (one-hot encoding) for categorical variables – Lead Origin, Lead Source,
- Last Activity, Specialization, Current_occupation
- Splitting Train & Test Sets
  - 70:30 % ratio was chosen for the split
- Feature scaling
  - Standardization method was used to scale the features
- Checking the correlations
- Predictor variables which were highly correlated with each other were dropped (Lead Origin_Lead Import and Lead Origin_Lead Add Form).

Model Building

- Feature selection:
- Using all the features in the dataset impacts accuracy and performance.
- With RFE(recursive feature elimination) to select only important features.

- Only 15 columns from 48 columns were left after RFE

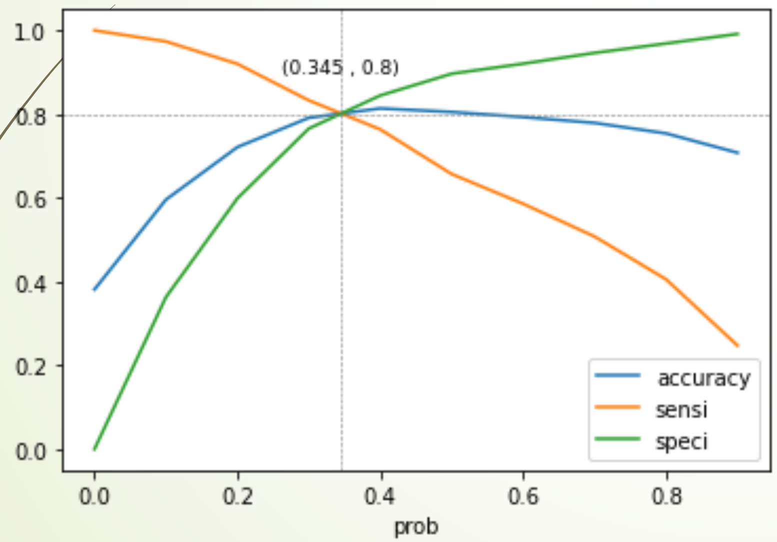
Model Building

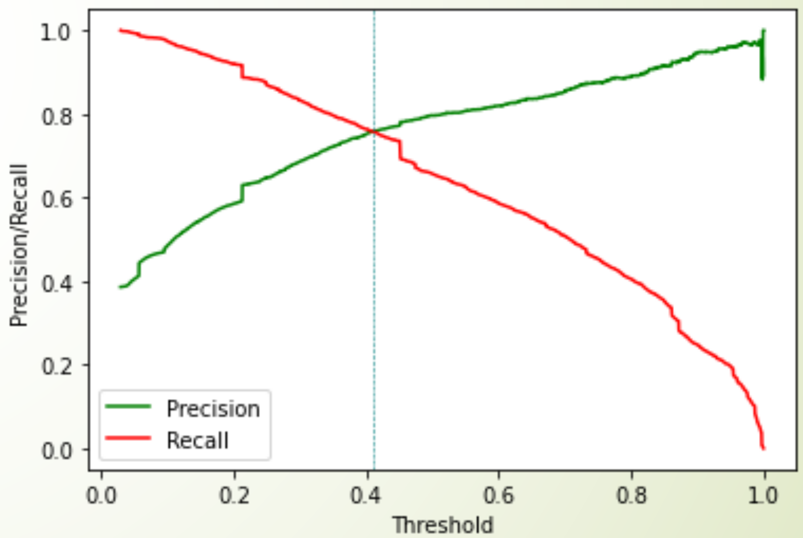
- Manual feature reduction is done from stats reports where we eliminate those columns whose p-value is greater than 0.05

- We built 4 models and the features have p-value less than 0.05 and no signs of multicollinearity

Model Evaluation

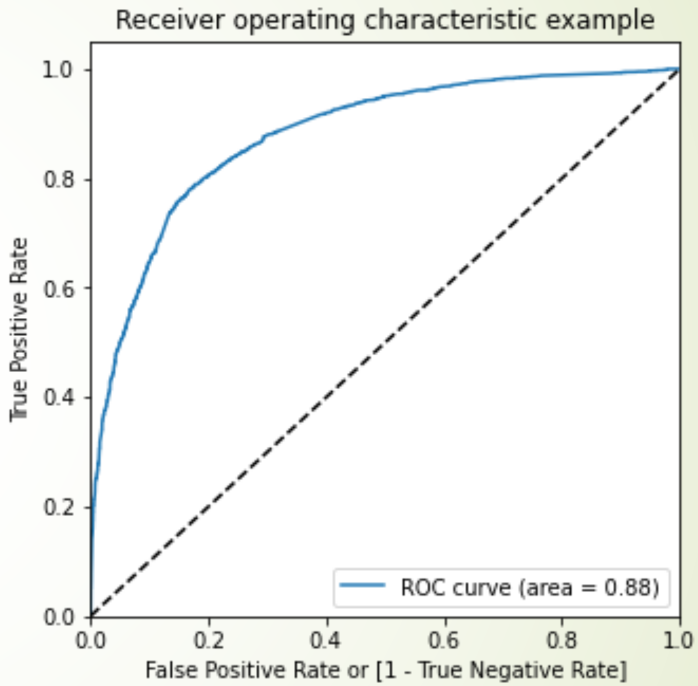
- Train test set
- Cutoff at 0.345 after checking the evaluation metrics is a good decision and 0.41 is the threshold





Model Evaluation

ROC Curve – Train Data Set
•Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.
•The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.

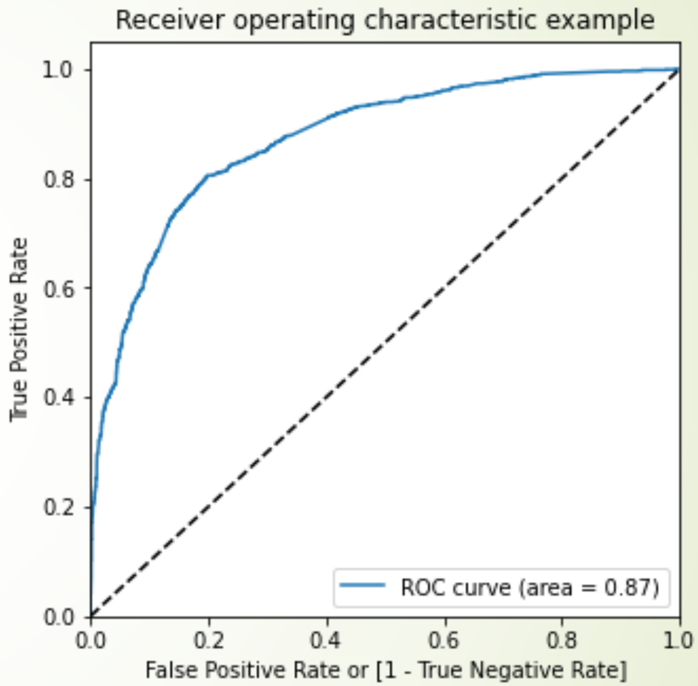


Model Evaluation

ROC Curve – Test Data Set

•Area under ROC curve is 0.87 out of 1 which indicates a good predictive model.

•The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



Model Evaluation

- Using a cut-off value of 0.345, the model achieved a sensitivity of 80.05% in the train set and 79.82% in the test set.
- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting
- The CEO of X Education had set a target sensitivity of around 80%.
- The model also achieved an accuracy of 80.46%, which is in line with the study's objectives.

Conclusion

- The final Logistic Regression Model has 12 features
- Top 3 features that contributing positively to predicting hot leads in the model are
  Lead Source_Welingak Website
  Lead Source_Reference
  Current_occupation_Working Professional
- The Optimal cutoff probability point is 0.345.Converted probability greater than 0.345 will be predicted as Converted lead (Hot lead) & probability smaller than 0.345 will be predicted as not Converted lead (Cold lead).

Recommendation

- **To increase our Lead Conversion Rates**:
- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Engage working professionals with tailored messaging.
- Optimize communication channels based on lead engagement impact.
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing 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.

Recommendation

To identify areas of improvement:
•Analyze negative coefficients in specialization offerings.
•Review landing page submission process for areas of improvement.

Thank you