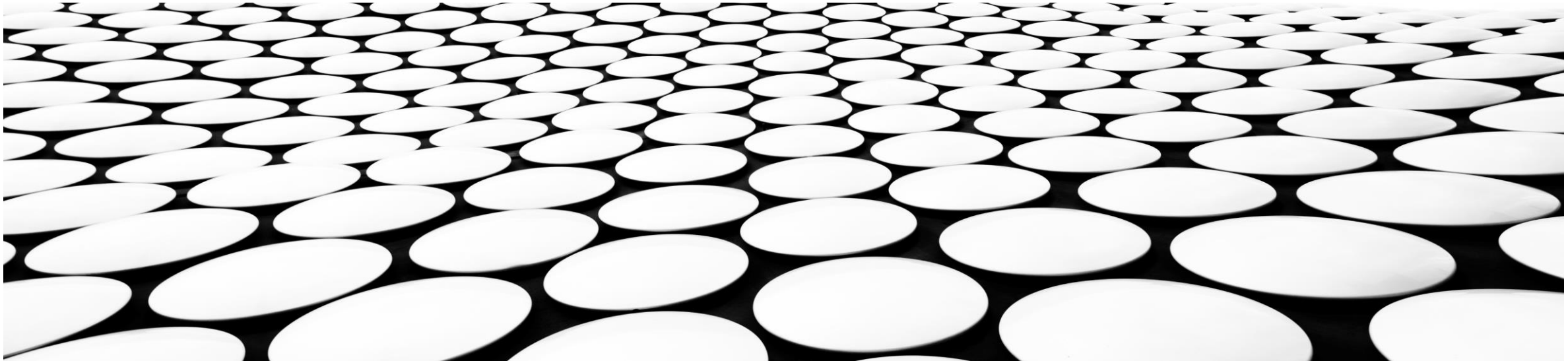
LEAD SCORE CASE STUDY

SUBMITTED BY

DHYEY RATANPARA , DHRUV AGARWAL, DEEP ROYCHAUDHRY



PROBLEM STATEMENT

- X Education sells online courses to industry professionals. 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. Moreover, the company also gets leads through past referrals.

- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

BUSINESS GOAL

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

STRATEGY

- Obtain data for analysis
- Clean and organize data
- Analyze data.
- Feature Scaling
- Combine data distribution with testing and training data.
- Create a logistic regression model and calculate lead score.
- Evaluate the model using different metrics (specificity and sensitivity or precision and recall).
- Use the best model for the test data based on measurement accuracy and specificity.

PROBLEM SOLVING METHODOLOGY

Data Sourcing, Cleaning and Preparation

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization.

Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data into train and test set.

Model Building

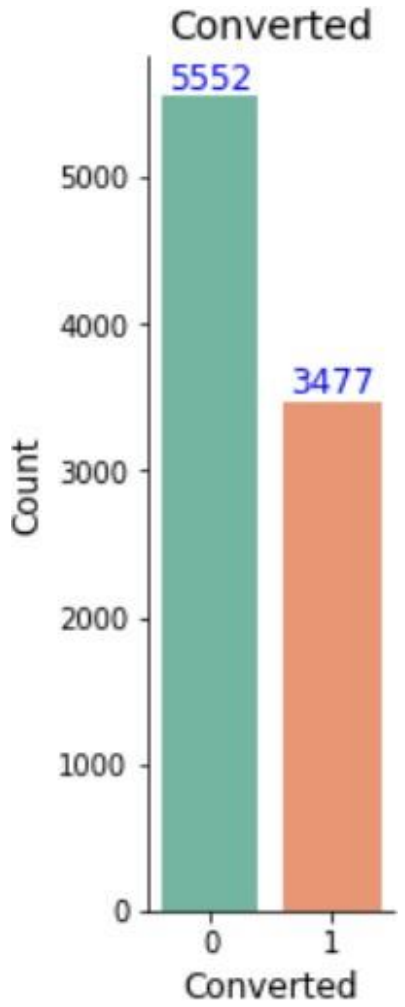
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

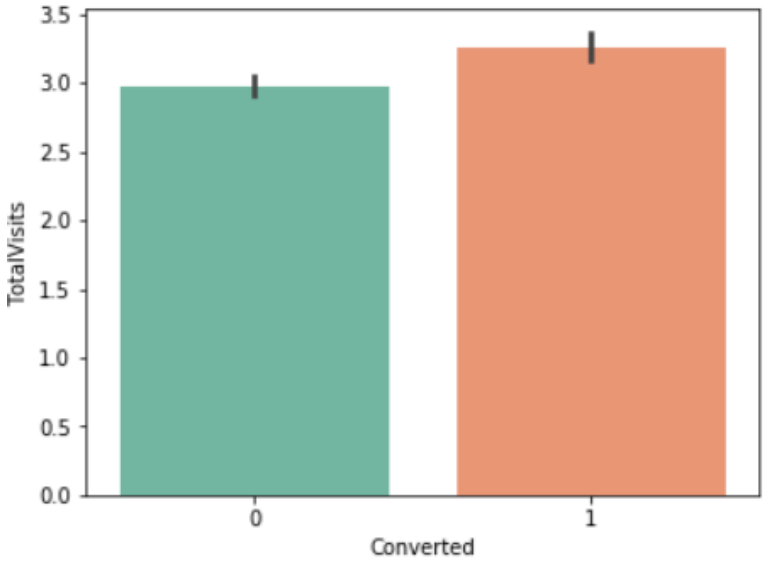
Result

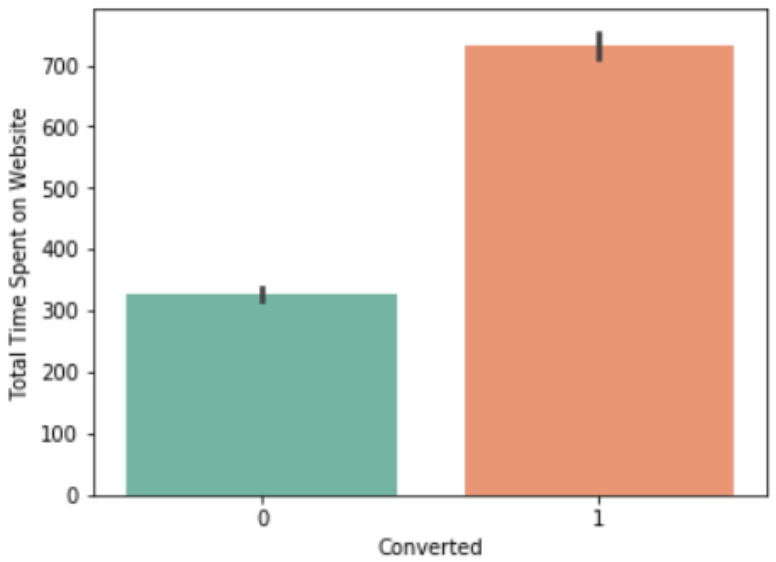
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

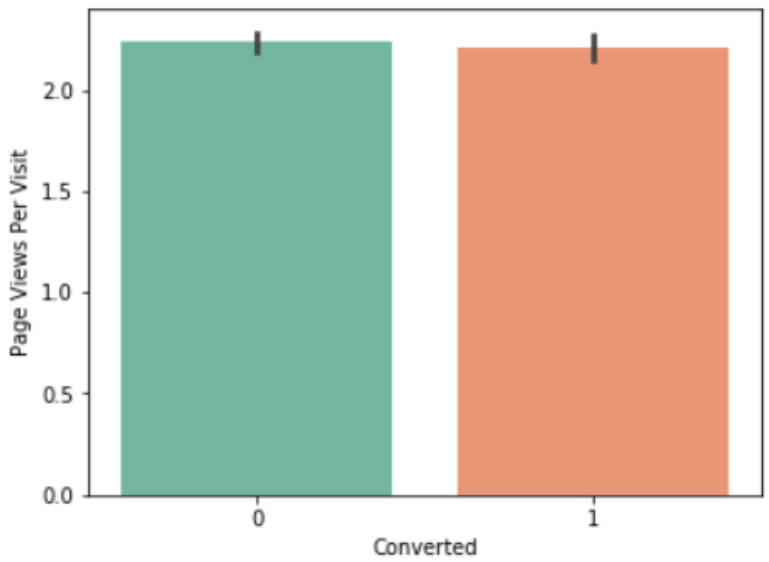
EXPLORATORY DATA ANALYSIS

- We have around 39% Conversion rate in Total



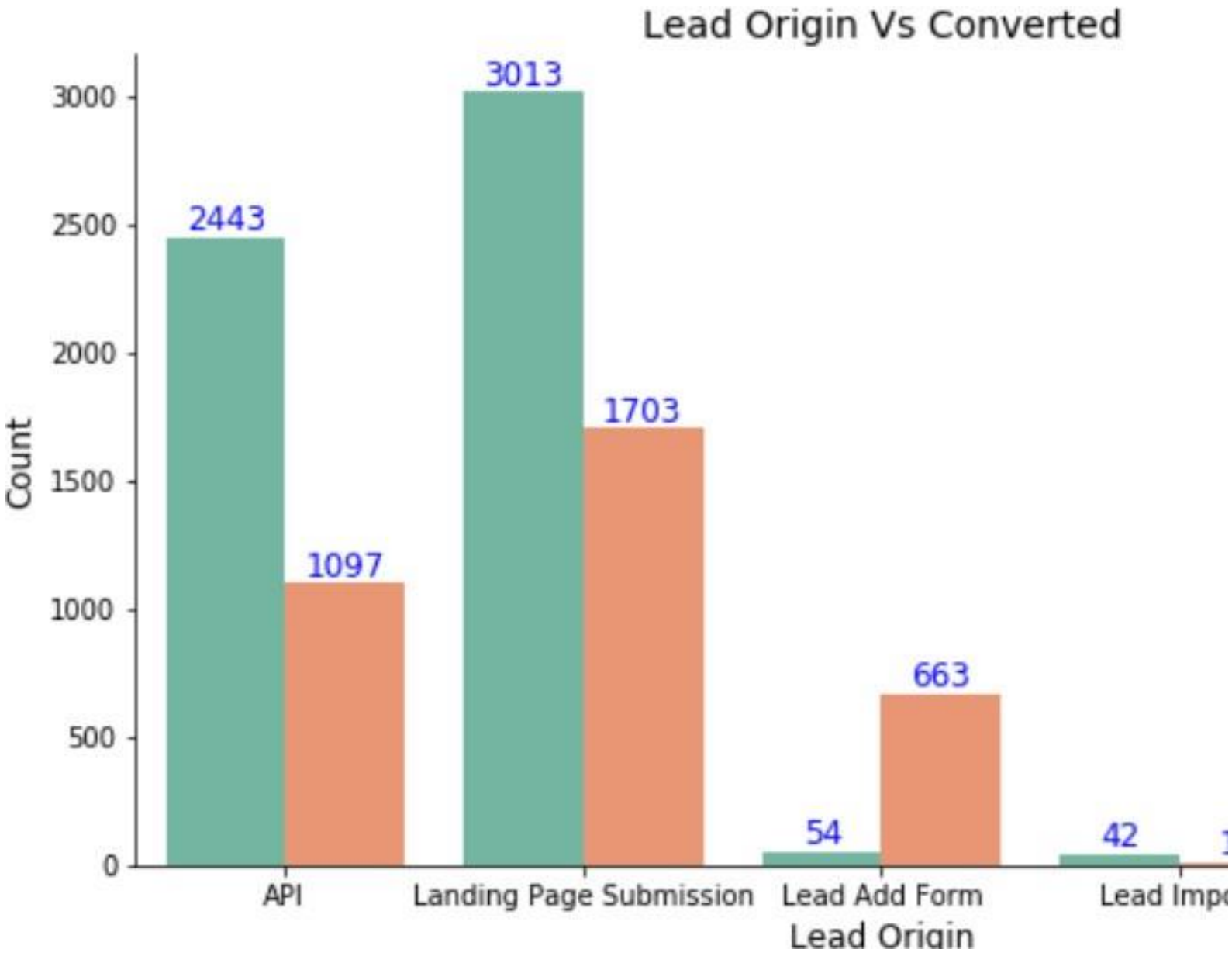




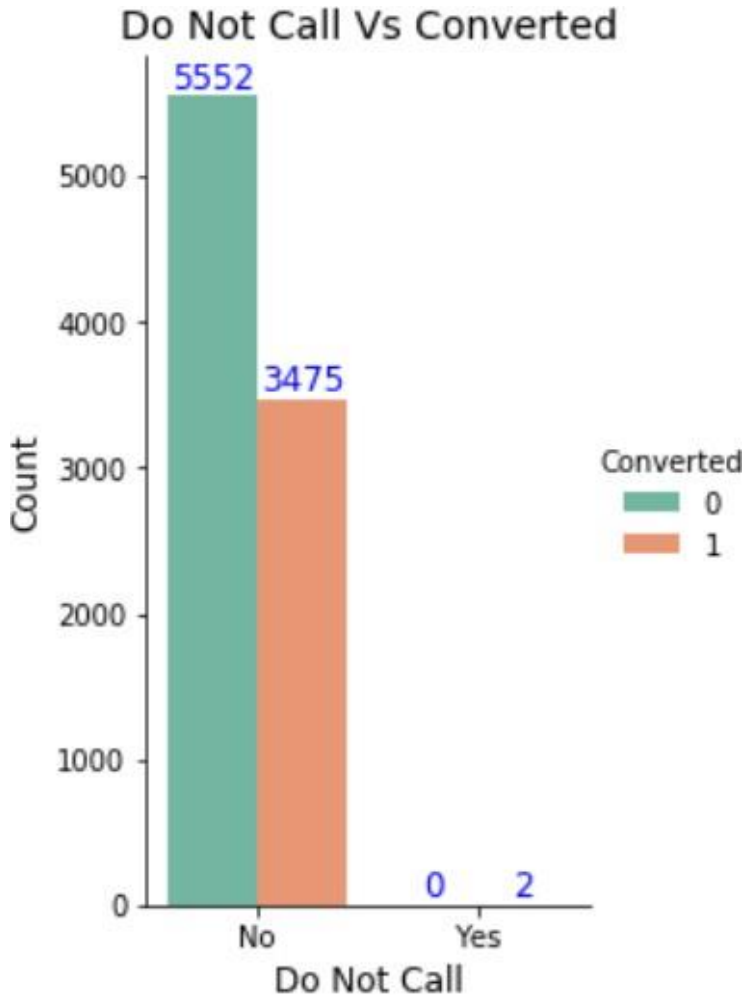


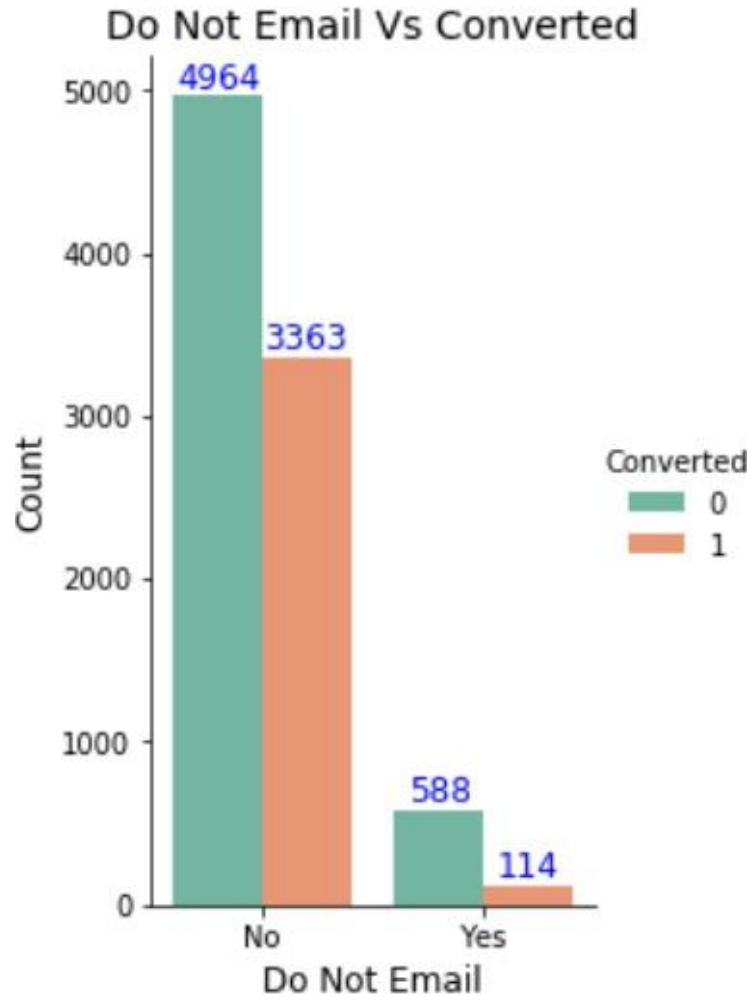
- The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit

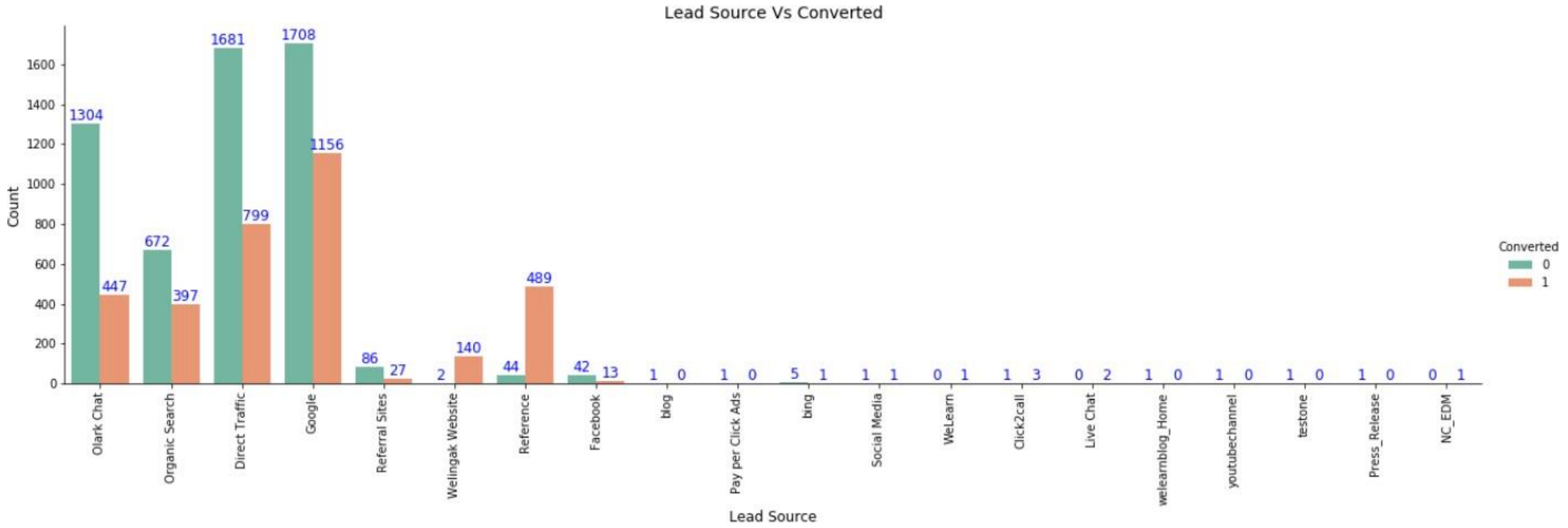
- In Lead Origin, maximum conversion happened from Landing Page Submission



- Major conversion has happened from Emails sent and Calls made

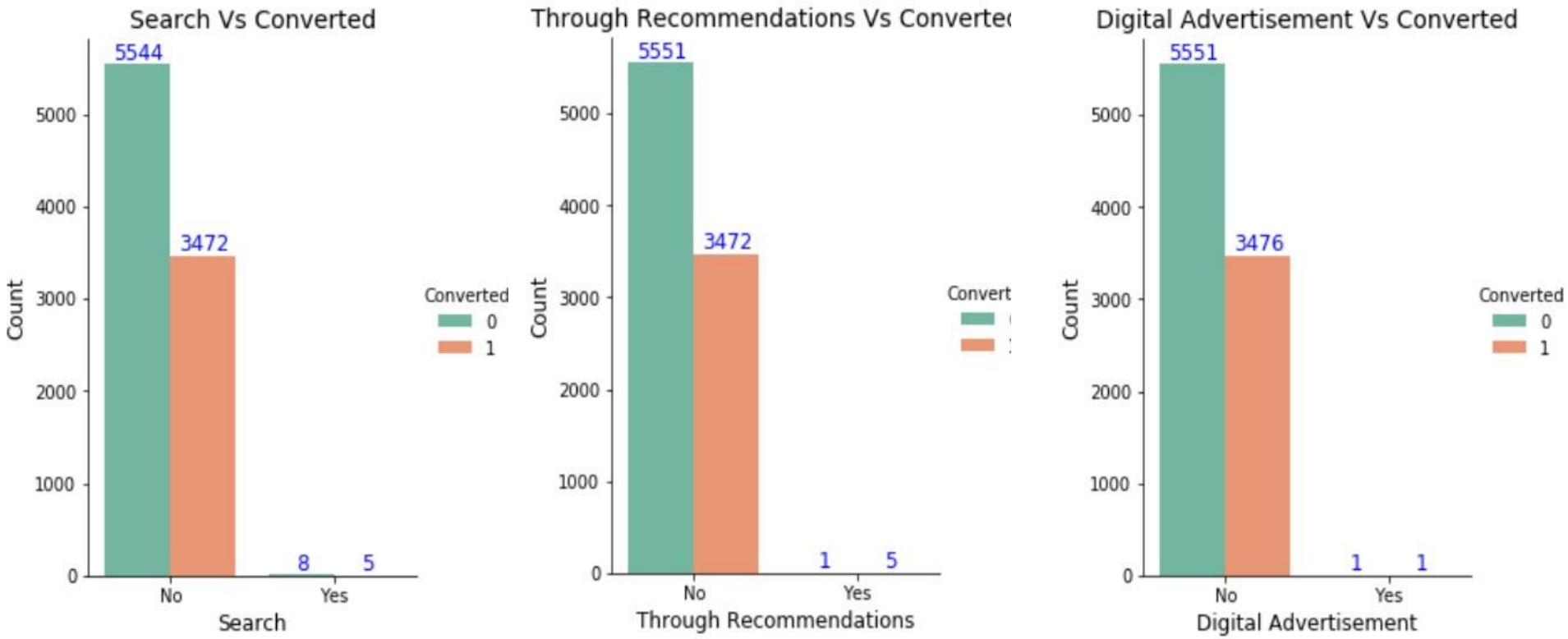




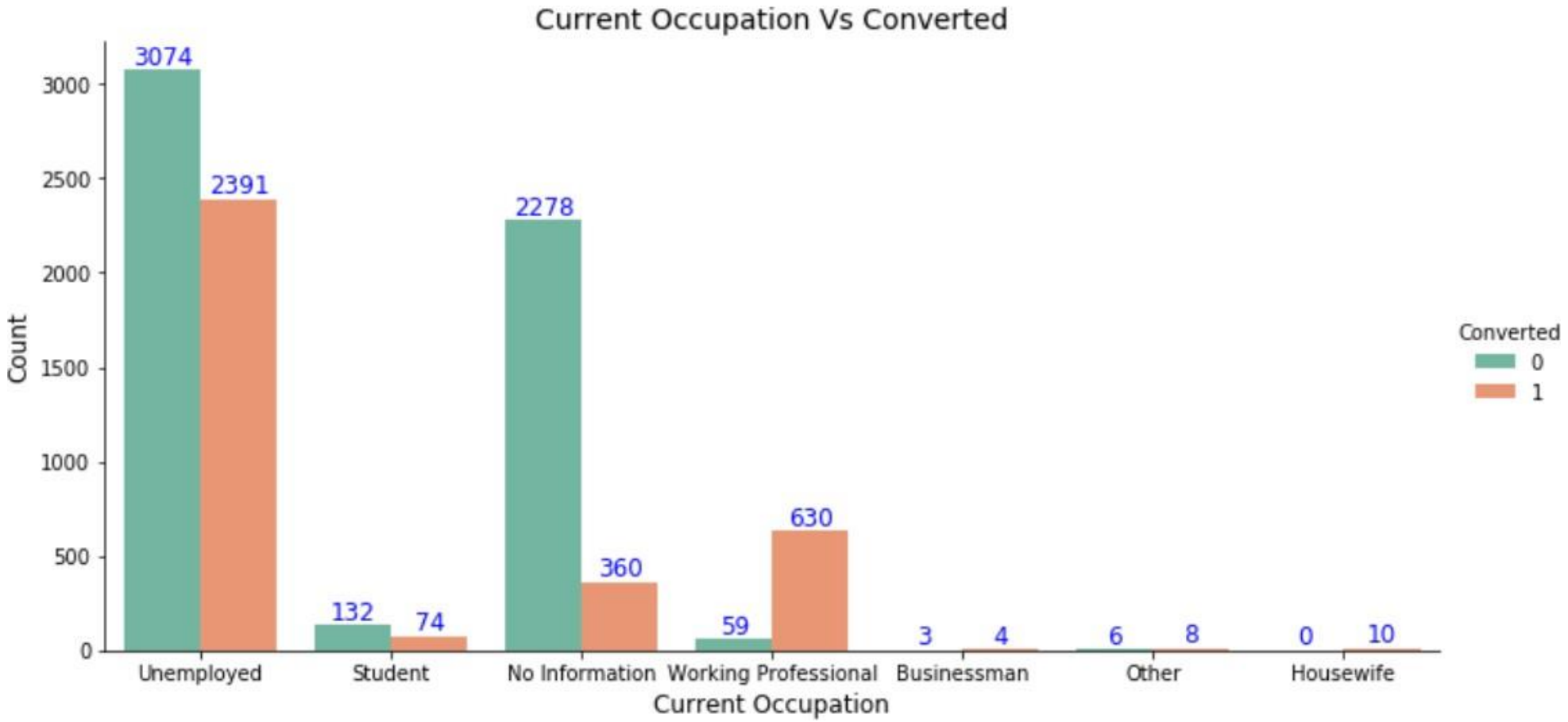


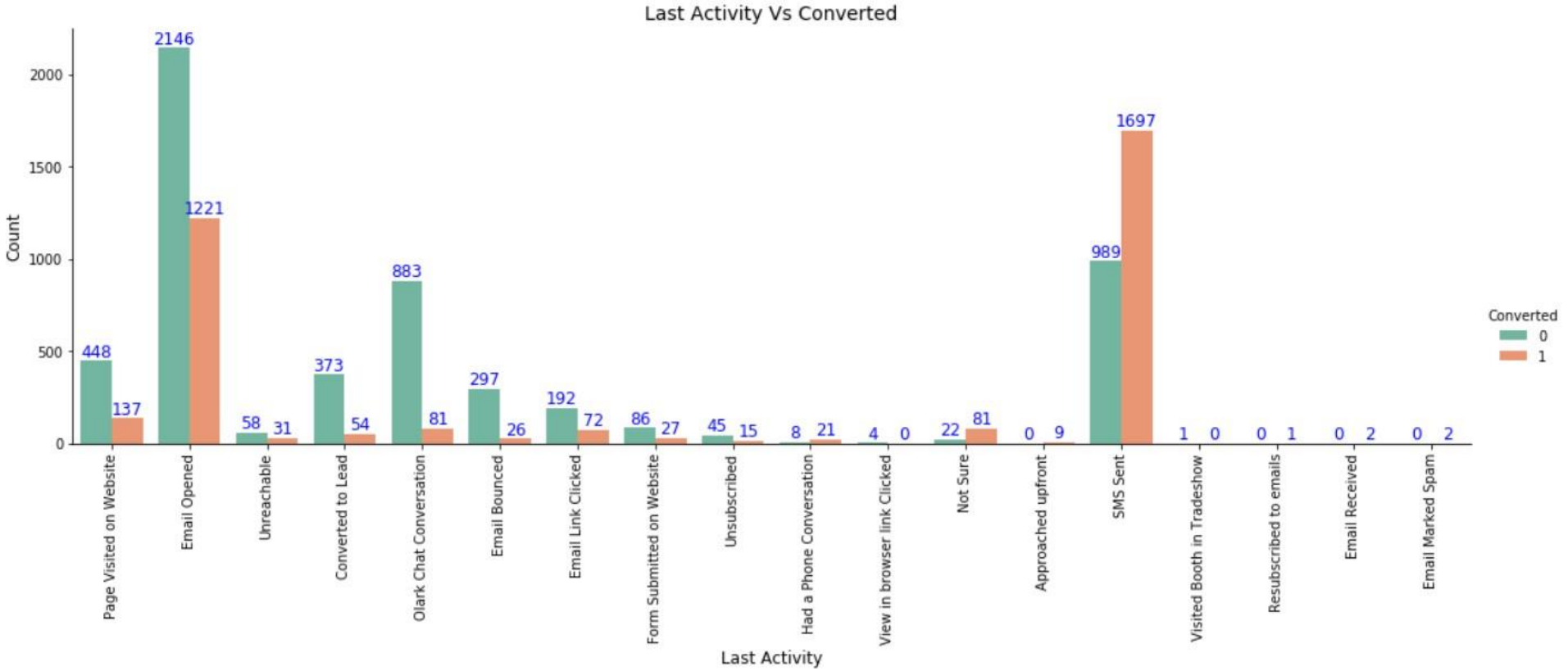
- Major conversion in the lead source is from Google

- Not much impact on conversion rates through Search, digital advertisements and through recommendations



- More conversion happened with people who are unemployed





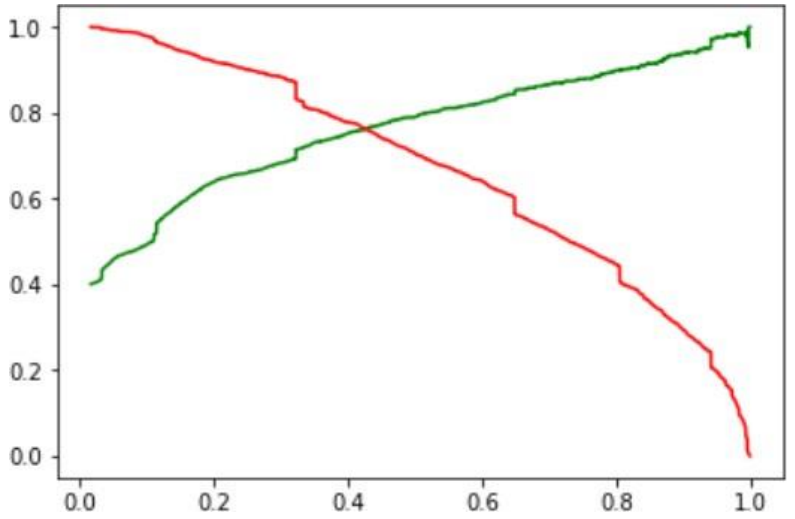
- Last Activity value of SMS Sent' had more conversion.

VARIABLE IMPACTING CONVERSIOIN RATE

- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin – Lead Page Submission
- Lead Origin – Lead Add Form
- Lead Source - Olark Chat
- Last Source – Welingak Website
- Last Activity – Email Bounced
- Last Activity – Not Sure
- Last Activity – Olark Chat Conversation
- Last Activity – SMS Sent
- Current Occupation – No Information
- Current Occupation – Working Professional
- Last Notable Activity – Had a Phone Conversation
- Last Notable Activity - Unreachable

MODEL EVALUATION

SENSITIVITY AND SPECIFICITY ON TRAIN DATA SET



- The graph depicts an optimal cut off of 0.42 based on Precision and Recall
- Confusion Matrix as shown
- Precision – 79%
- Recall – 71%

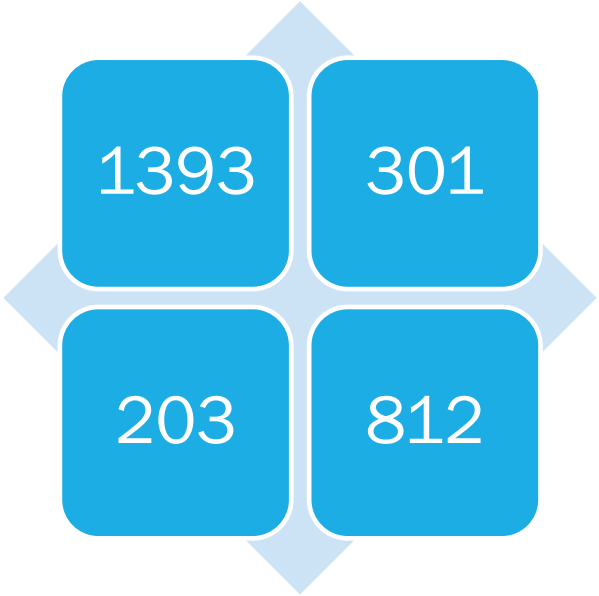
3412	446
712	1750

MODEL EVALUATION

SENSITIVITY AND SPECIFICITY ON TEST DATA SET

- Confusion Matrix

- Accuracy – 81%
- Sensitivity – 80%
- Specificity – 82%



CONCLUSION

- When we evaluated sensitivity-specificity based on accuracy and regression analysis, we determined the best cut-off point based on sensitivity and specificity to calculate the final prediction.
- The accuracy, sensitivity and specificity values of the test method are approximately 81%, 79% and 82% respectively , which are approximately close to the relevant values including training.
- Moreover, simulation scores show that the change of the final prediction model is about 80% (in the training process) and 79% in testing.
- Leaders in the model The 3 most important differences in the transformation are:
  Total time spent on website
  Lead onboarding form from lead source
  Had a phone conversation from last notable activity
- So, overall, this model looks good.