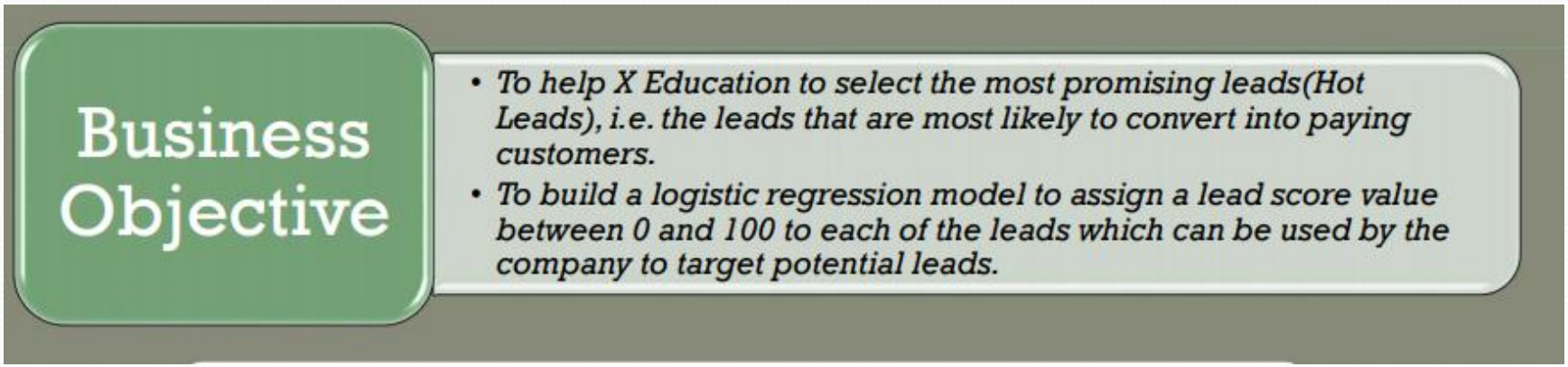
Lead Scoring Case Study

By

Lokesh Laddha

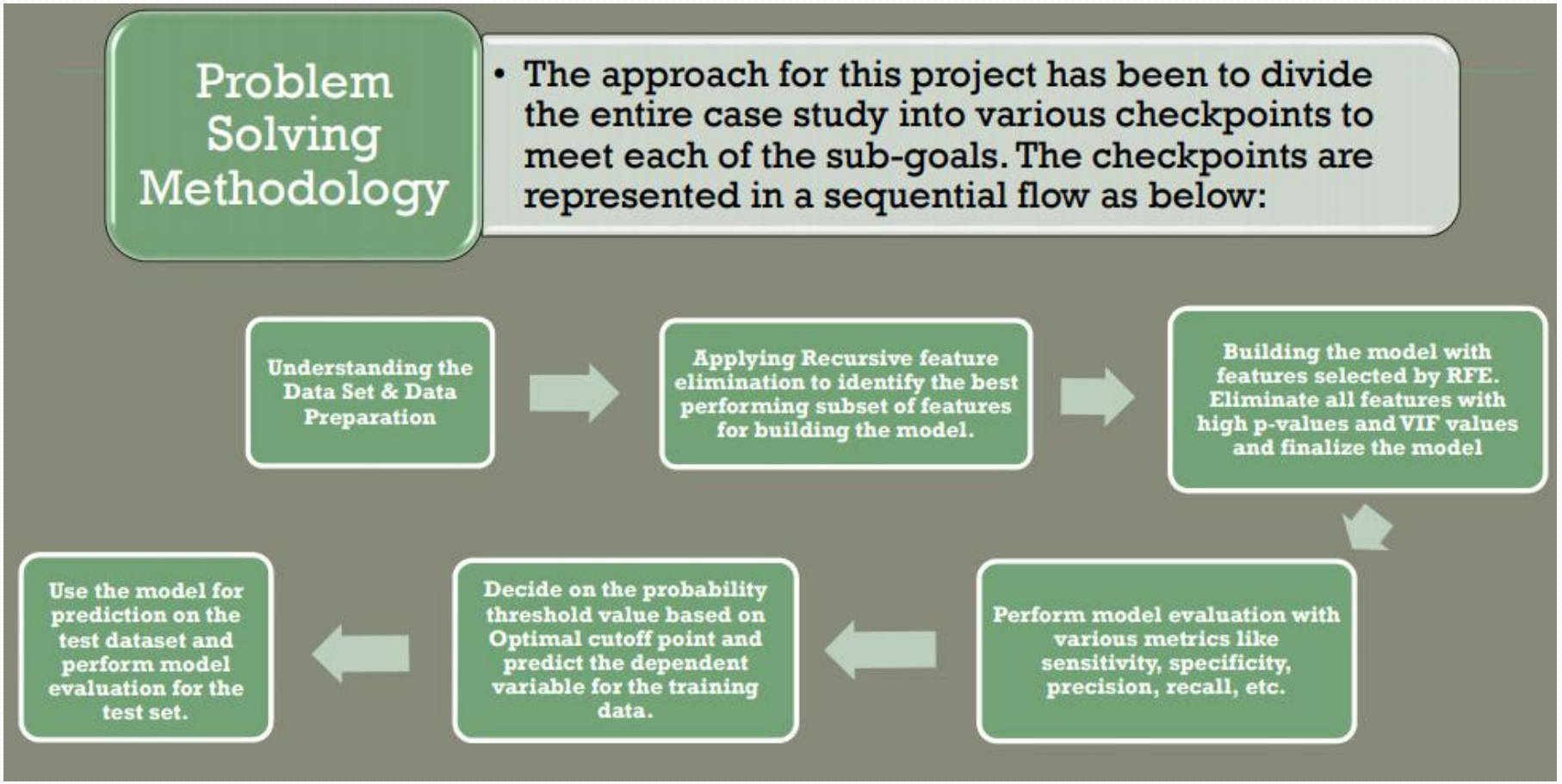
Karthika TV

Business Objective



Problem Solving Methodology

- The approach for this project has been to divide the entire case study into various checkpoints to meet each of the sub-goals. The checkpoints are represented in a sequential flow as below:



Solution Methodology

- Data cleaning and data manipulation.

  1.Check and handle duplicate data.

  2.Check and handle NA values and missing values.

  3.Drop columns, if it contains large amount of missing values and not useful for the analysis.

  4.Imputation of the values, if necessary.

  5.Check and handle outliers in data.

.

Continued..

- EDA

  1.Univariate data analysis: value count, distribution of variable etc.

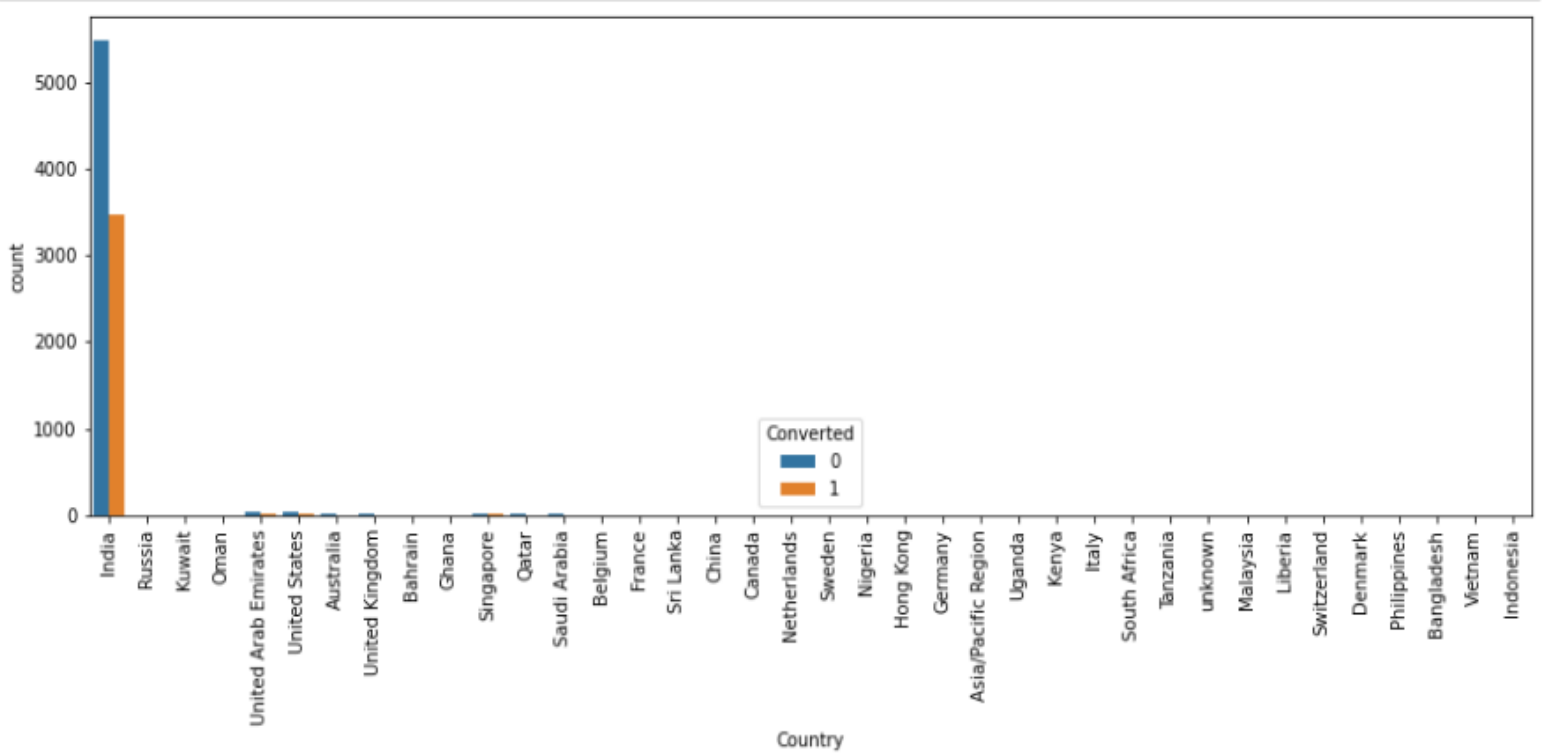
  2.Bivariate data analysis: correlation coefficients and pattern between the variables etc

Continued..

- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

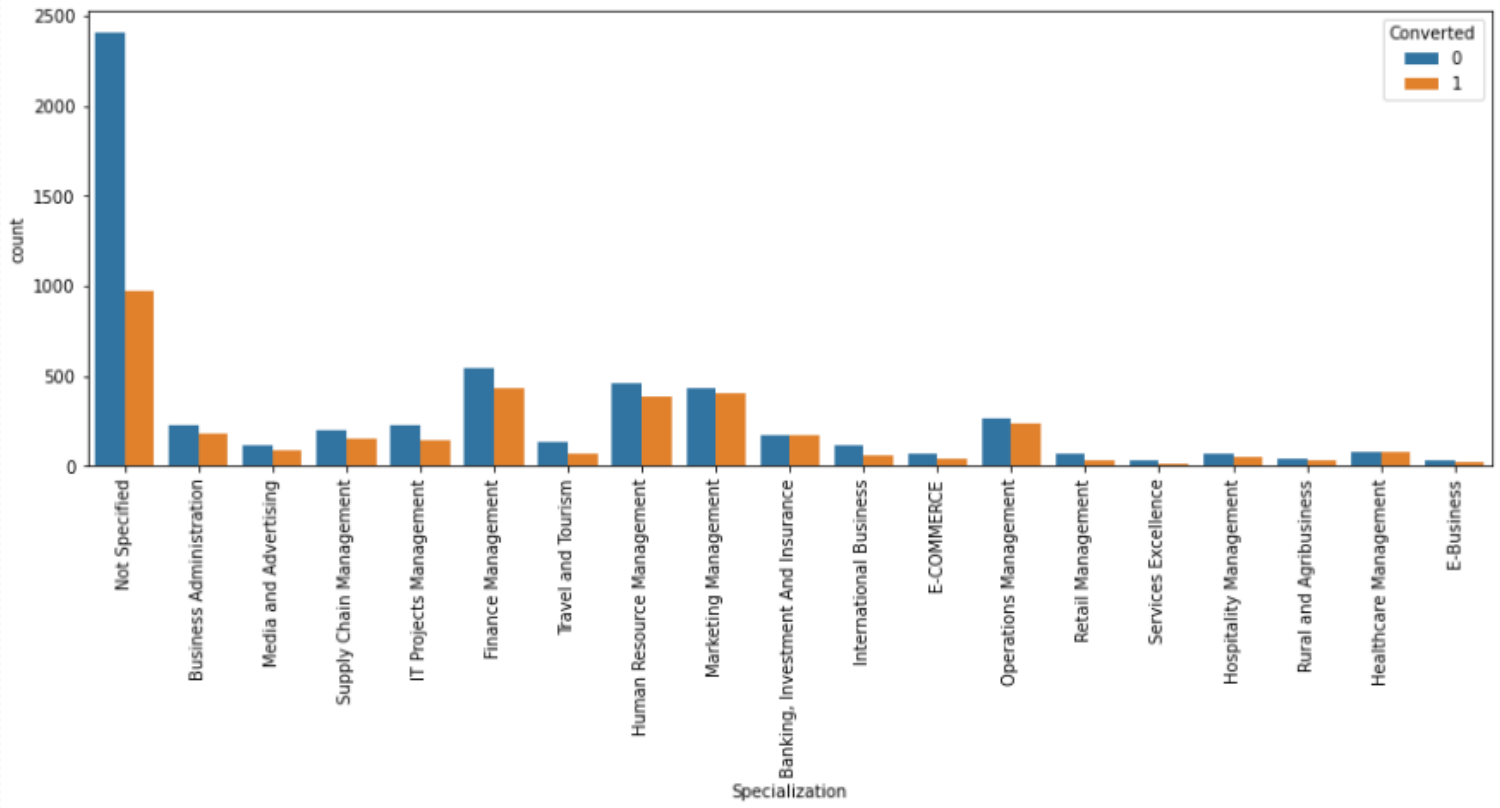
Categorical Attributes Analysis:

- Plotting spread of Country column



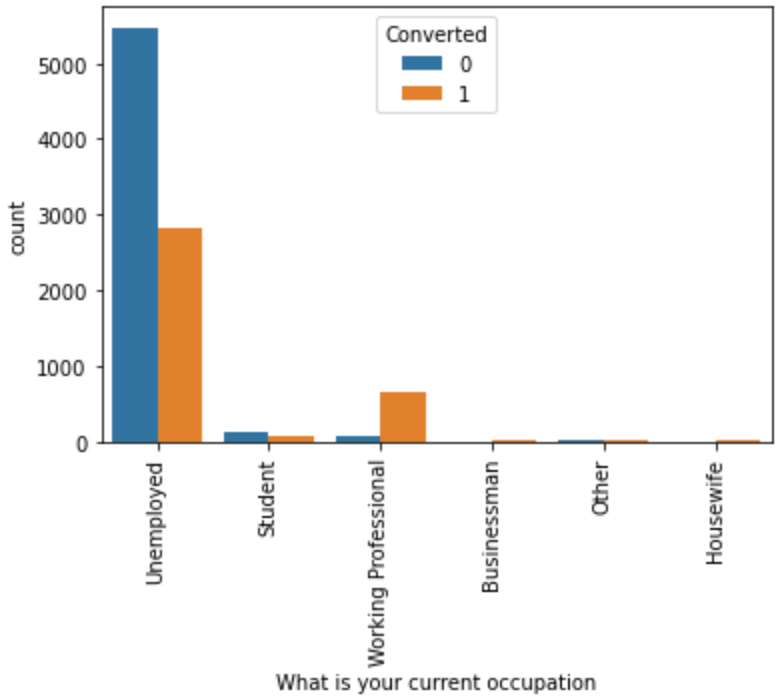
- As we can see the Number of Values for India are quite high (nearly 97% of the Data), this column can be dropped

Plotting spread of Specialization column



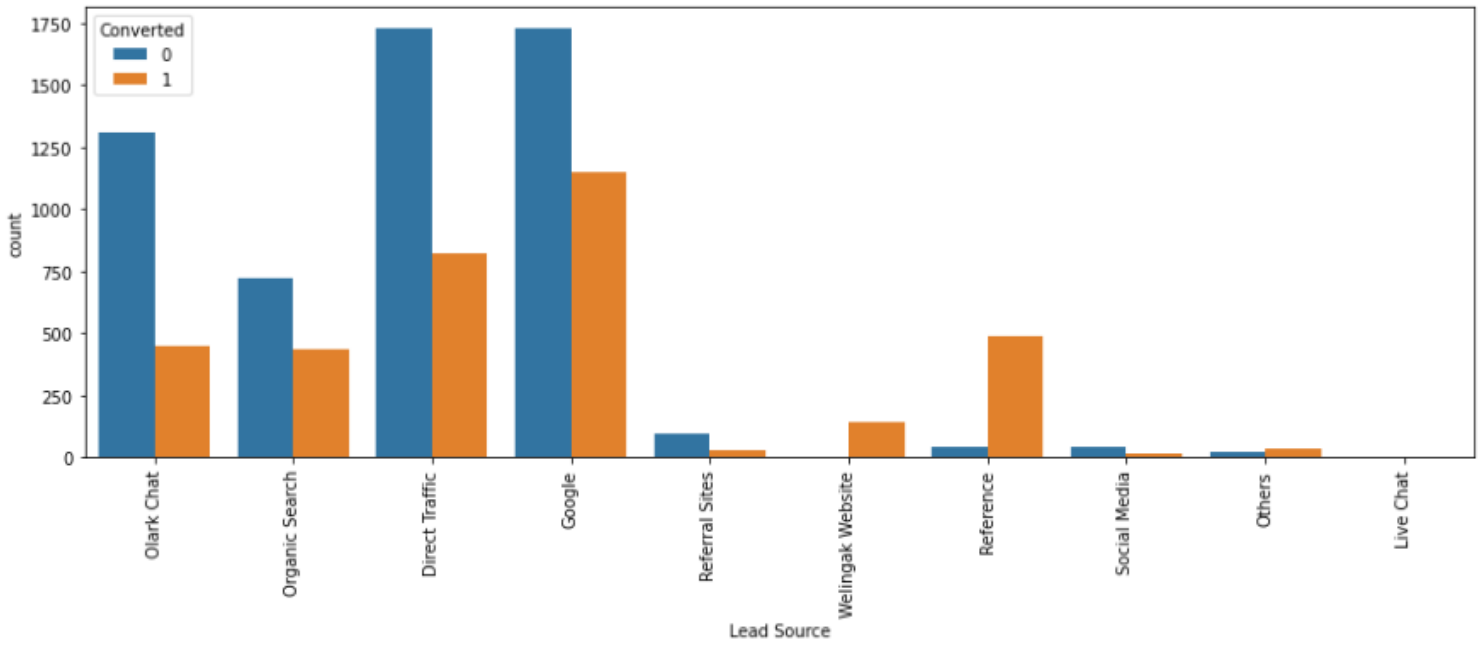
- We see that specialization with Management in them have higher number of leads as well as leads converted. So this is definitely a significant variable and should not be dropped.

Visualizing count of Variable based on Converted value



- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in terms of Absolute numbers.

Visualizing count of Variable based on Converted value

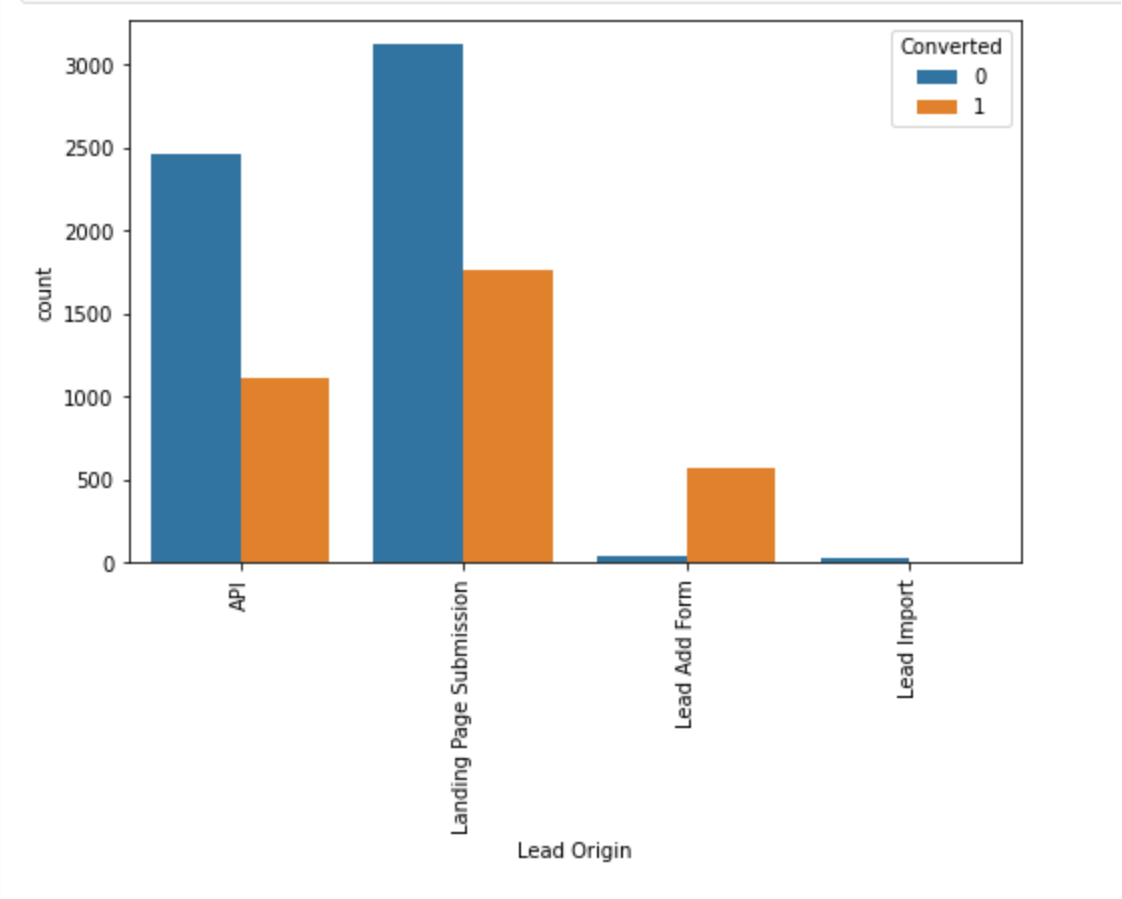


- Inference
- Maximum number of leads are generated by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct - traffic, and google leads and generate more leads from reference and welingak website.

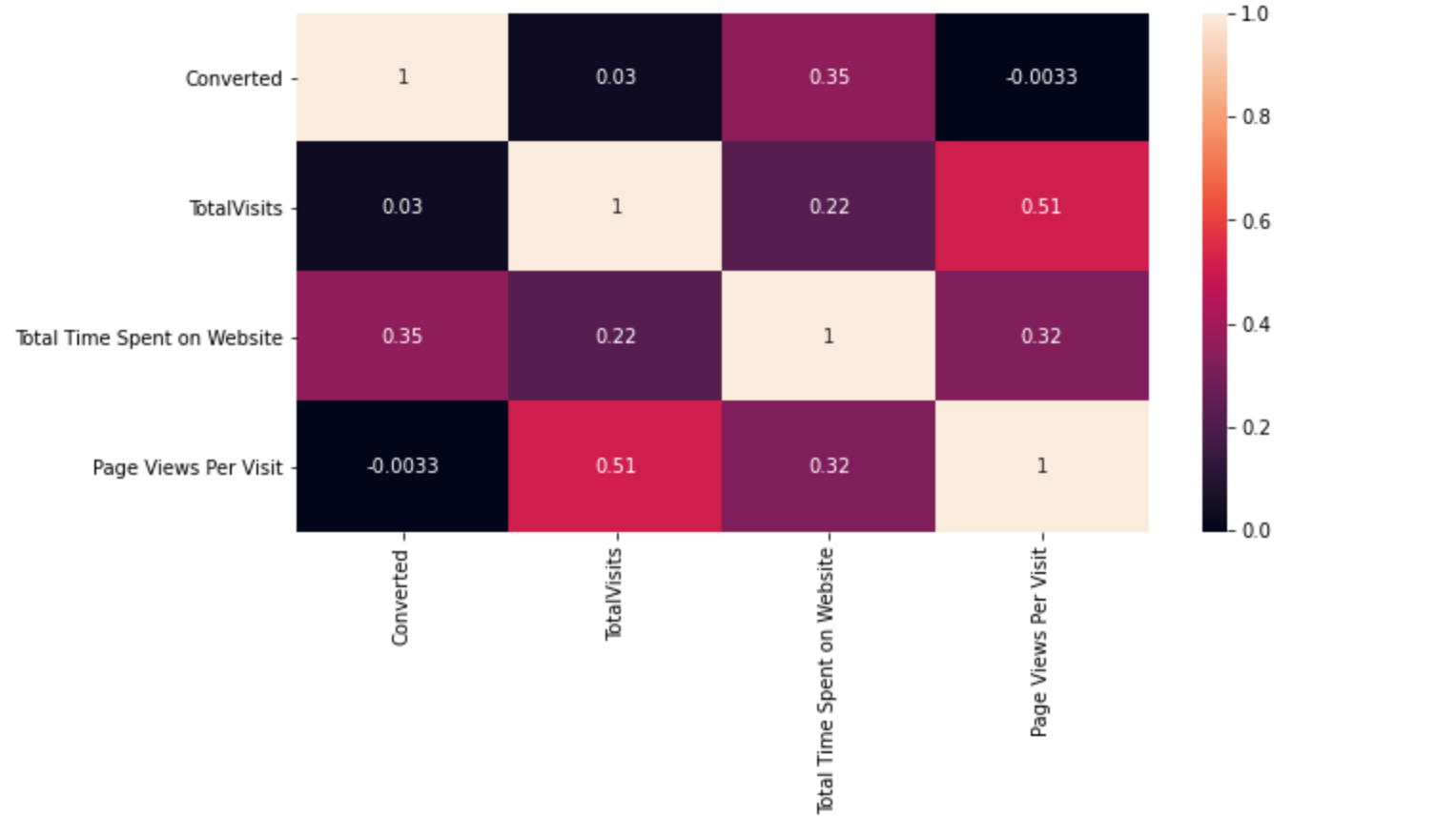
Visualizing count of Variable based on Converted value

- Inference
- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form get very few leads.
- In order to improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form

Continued..


Converted
0
1
3000
2500
2000
1500
1000
500
0
count
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin

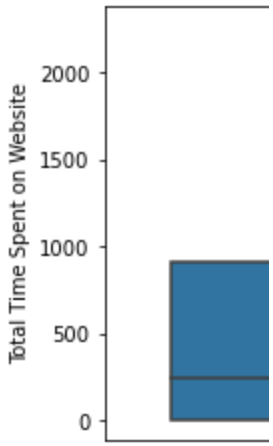
Numerical Attribute Analysis: Checking correlations of numeric values



Visualizing spread of numeric variable

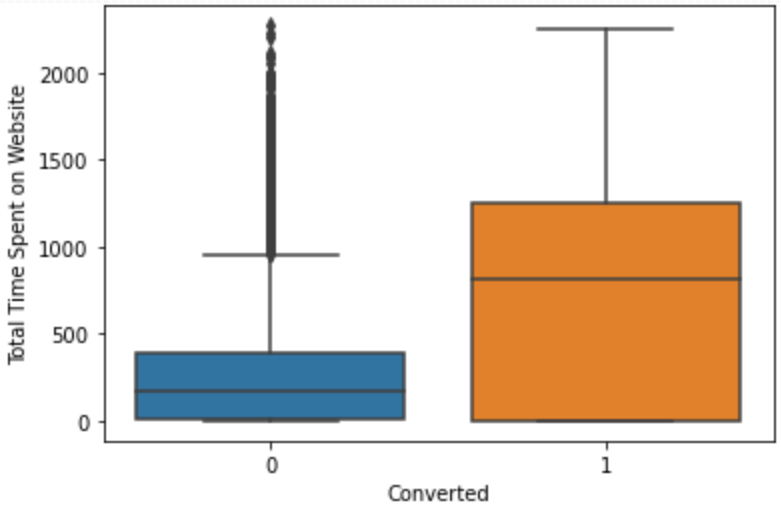
Lead Scoring Case Study

By Lokesh Laddha
Karthika TV



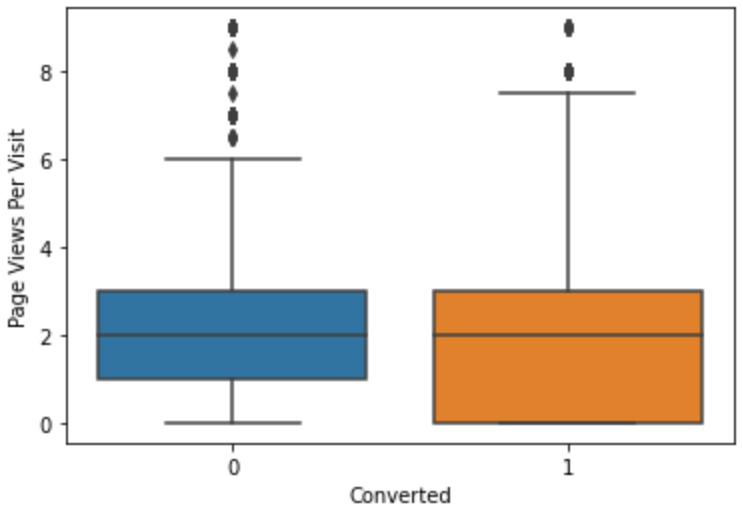
- Since there are nc ...is above Column
- Check for Page Views Per Visit:

Checking Spread of "Total Time Spent on Website" vs Converted variable



- Inference
- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.

checking Spread of "Page Views Per Visit" vs Converted variable

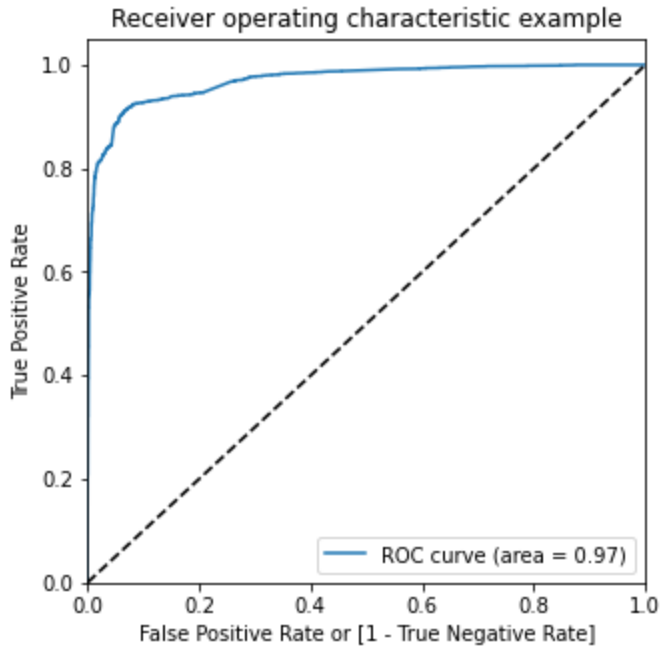


- Inference
- Median for converted and unconverted leads is the same.
- Nothing can be said specifically for lead conversion from Page Views Per Visit

Model Building

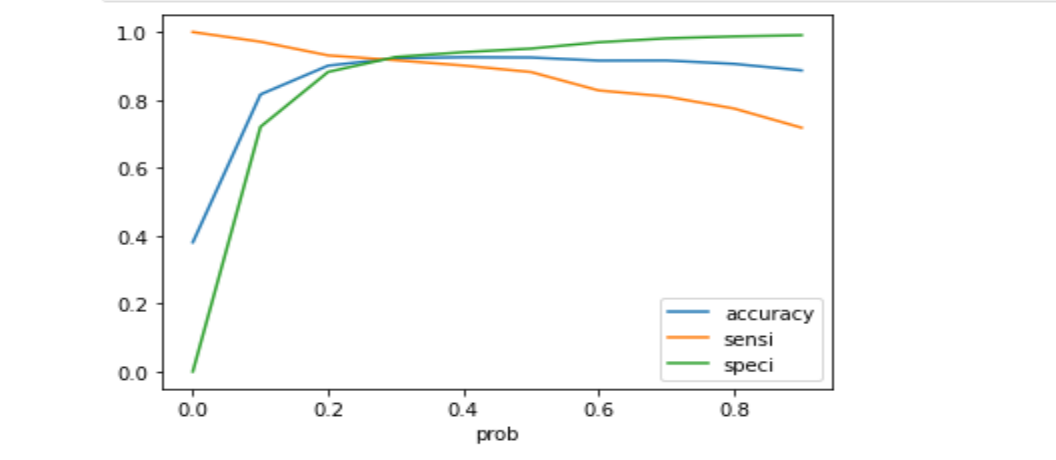
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and vifvalue is greater than 5
- Predictions on test data set

ROC Curve



- The ROC Curve should be a value close to 1. We are getting a good value of 0.97 indicating a good predictive model.

Continued…



Continued..

- Finding Optimal Cut off Point
- Optimal cut off probability is that
- probability where we get balanced sensitivity and specificity
- From the second graph it is visible that the optimal cut off is at 0.3

PREDICTIONS ON TEST SET

- The final prediction of conversions have a target rate of 79% (78.5%) (Around 1 % short of the predictions made on training data set)

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 92%, 84% and 97% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 85%
- Hence overall this model seems to be good.
- The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

Overall Metrics - Accuracy, Confusion Metrics, Sensitivity, Specificity on test set

- *Overall accuracy : 0.92*
- *Sensitivity of our logistic regression model :0.84*
- *Specificity of our logistic regression model :0.97*
- Precision_score : 0.94
- Recall_score: 0.84

Thank You