LEAD SCORING CASE STUDY

PRESENTED BY: RISHAV KUMAR

MALVIKA CHAUHAN

Objective

- The objective is to help X Education an education company sells online courses to industry professionals, to select the most promising leads by building a model and assigning a lead score 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%

Data Cleaning and Preparation

- Below are the steps which are performed :-

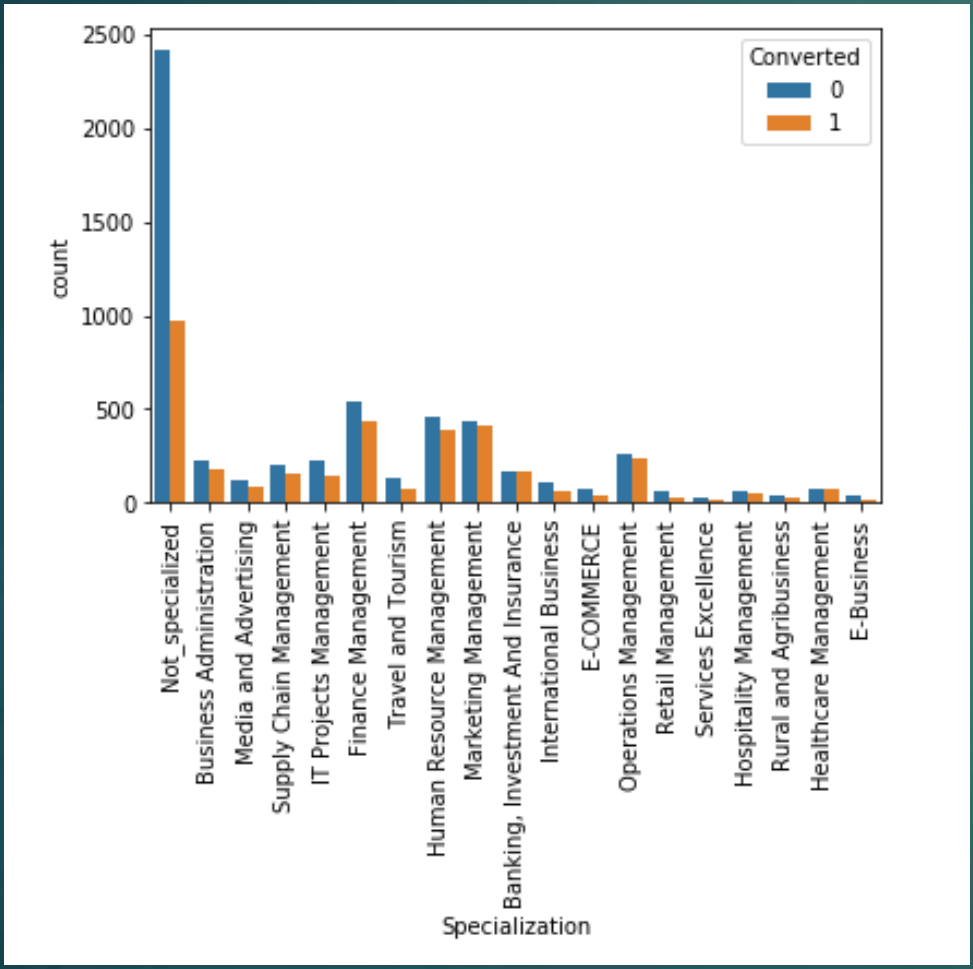
1. **Data Inspection**: It is the act of viewing data for verification and debugging purposes, before, during, or after a translation. With reference to dataset-app.info(), app.describe(), app.shape are some method in inspect the data.
2. **Identify percentage of null values in columns and rows** : We found there are multiple columns in the dataset containing null values in dataset.
3. **Drop columns with high null values:** We followed the strategy of dropping columns having null percentage greater than 45 .
4. **Replacing Select with nan.: columns like** Specialization , Lead Profile , City have Select as a value which needs to be imputed as null before model building.

- **Imputting missing values**

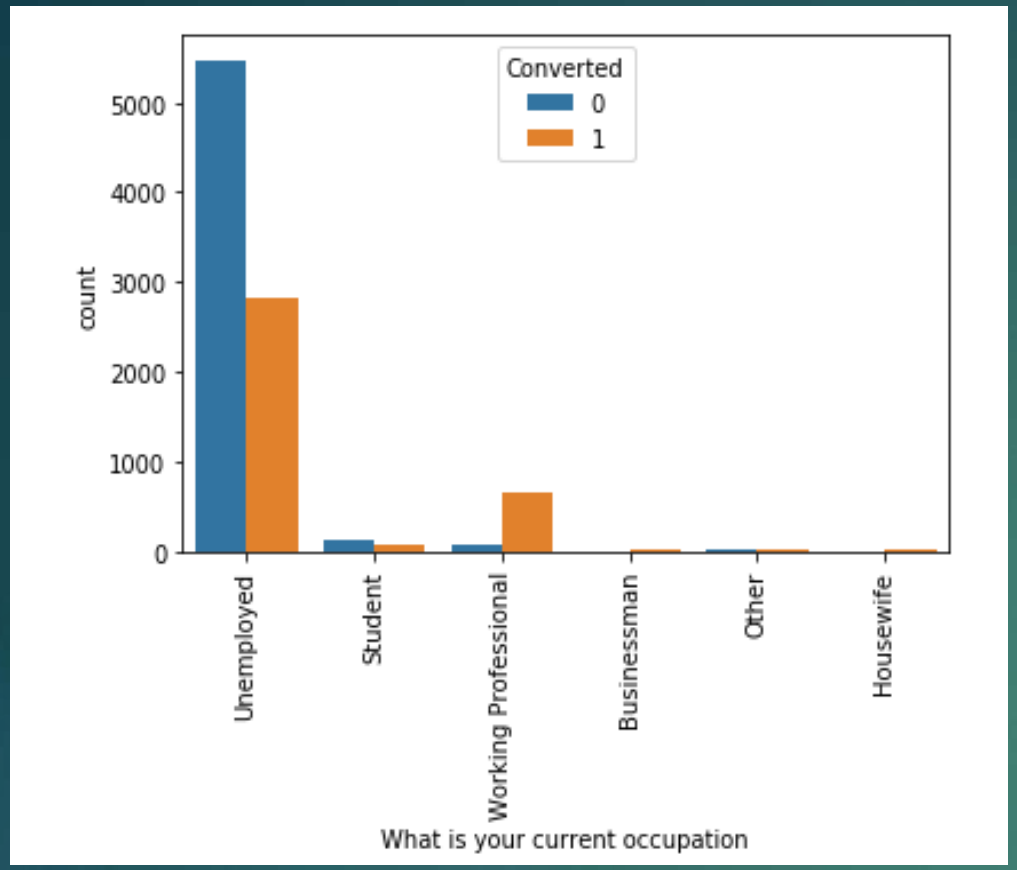
1. **Impute ‘YES’ & ‘NO’ columns with 0 and 1: Columns like** Search , Do Not Email , Do Not Call are some Boolean columns that need to imputed with 0 and 1.
2. **Country :** Column is quite disbalanced, so we dropped it other wise model will produce biased result.
3. **Specialization :** Specialization has wide range of values. It is bit unpredictable with whom we can replace missing values with. It is better to impute missing values with new value Others. Might be people do not wanted to took specialization in any of these field.
4. **Current occupation :** As majority people are unemployed so it is safer to impute missing values with unemployed.

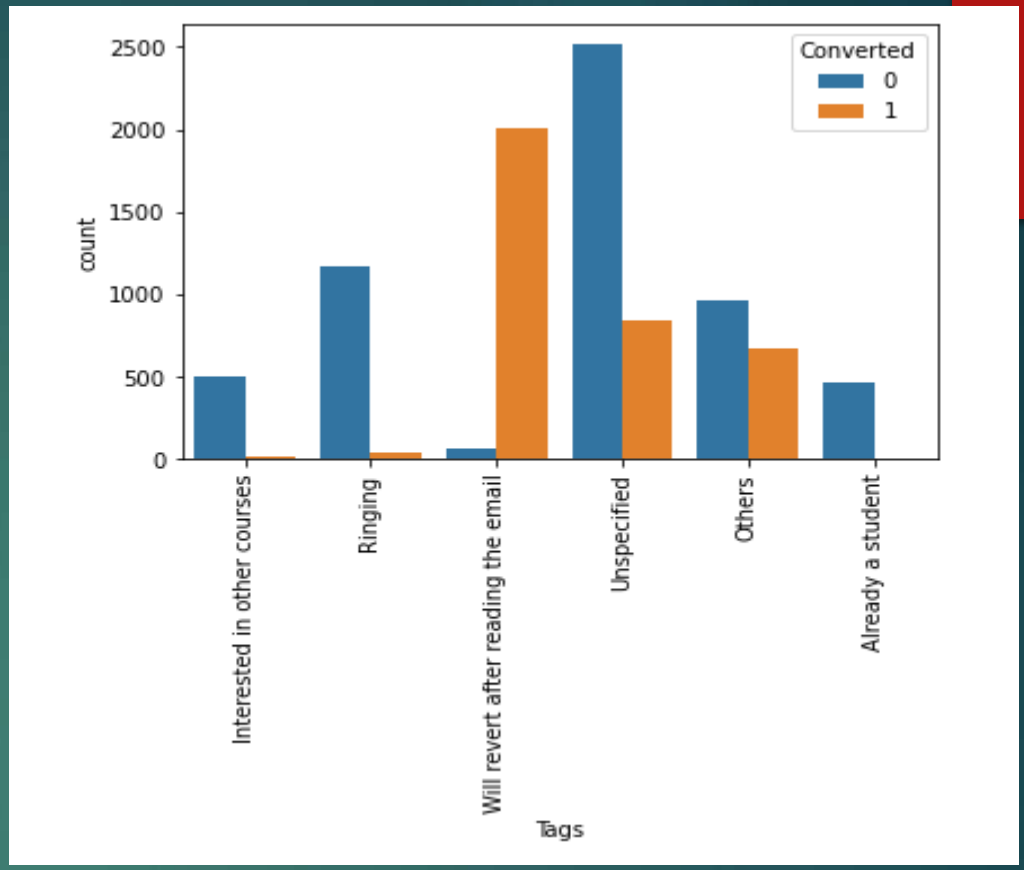
- **Handling Outliers: H**andling outlier in TotalVisits, Total Time Spent on Website and Page Views Per Visit by excluding values outside 99%ile.
- **Creating Dummies of categorical values**

EDA(Exploratory Data Analysis)

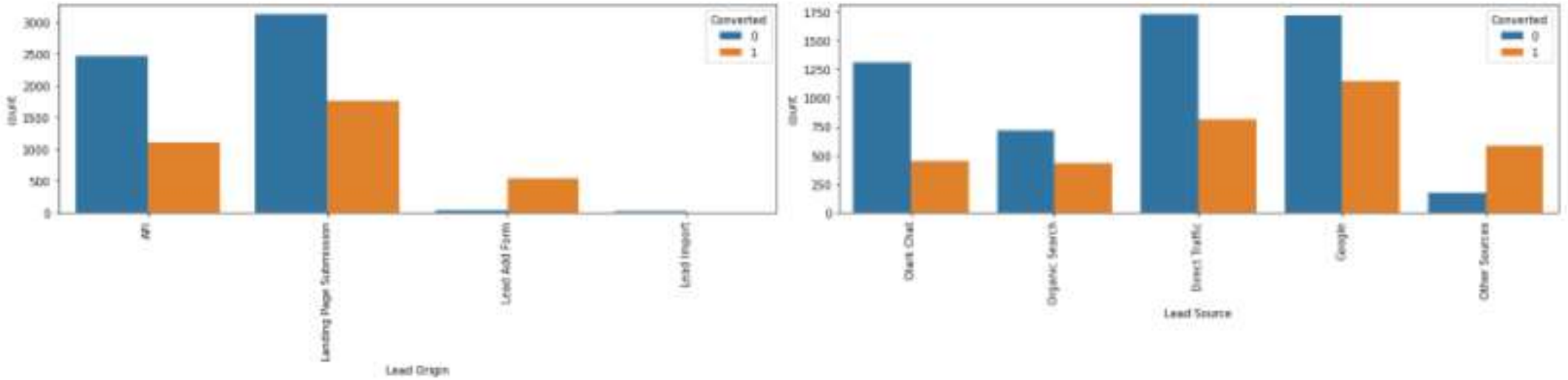


- Most of the leads are either unspecialized or from management specialization.
- Approaching the management specialized people will produce more conversion.





- Working Professionals have high chance of taking the course.
- Unemployed leads(can be students) are very high. If this portion handled properly, we can get more conversion rate.
- Almost all the leads given tag 'Will revert after reading the email' gets converted



- The conversion rate of 'Lead add form' is very high.
- API and Landing Page Submission brings higher number of leads and many are getting converted.
- Other sources though produce less leads but their conversion rate is high
- Direct Traffic and Google are the good source of leads as well their conversion rate is also quite decent.

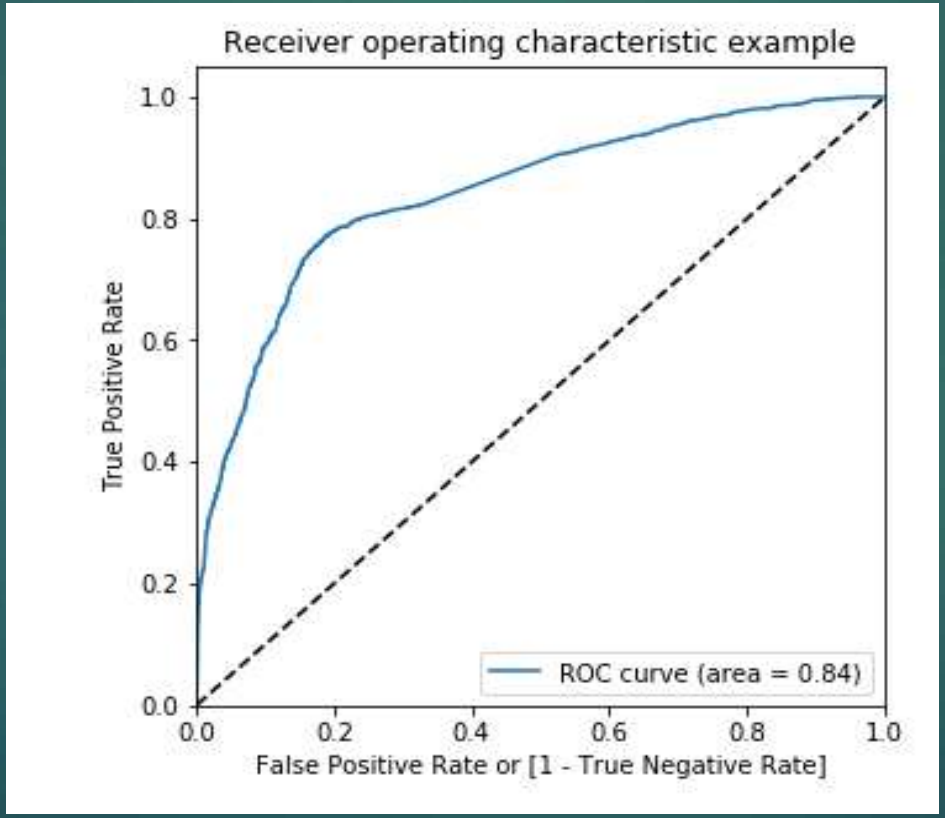
Building Logistic Model

- Top features left after an effective logistic modelling with low VIF and p value.

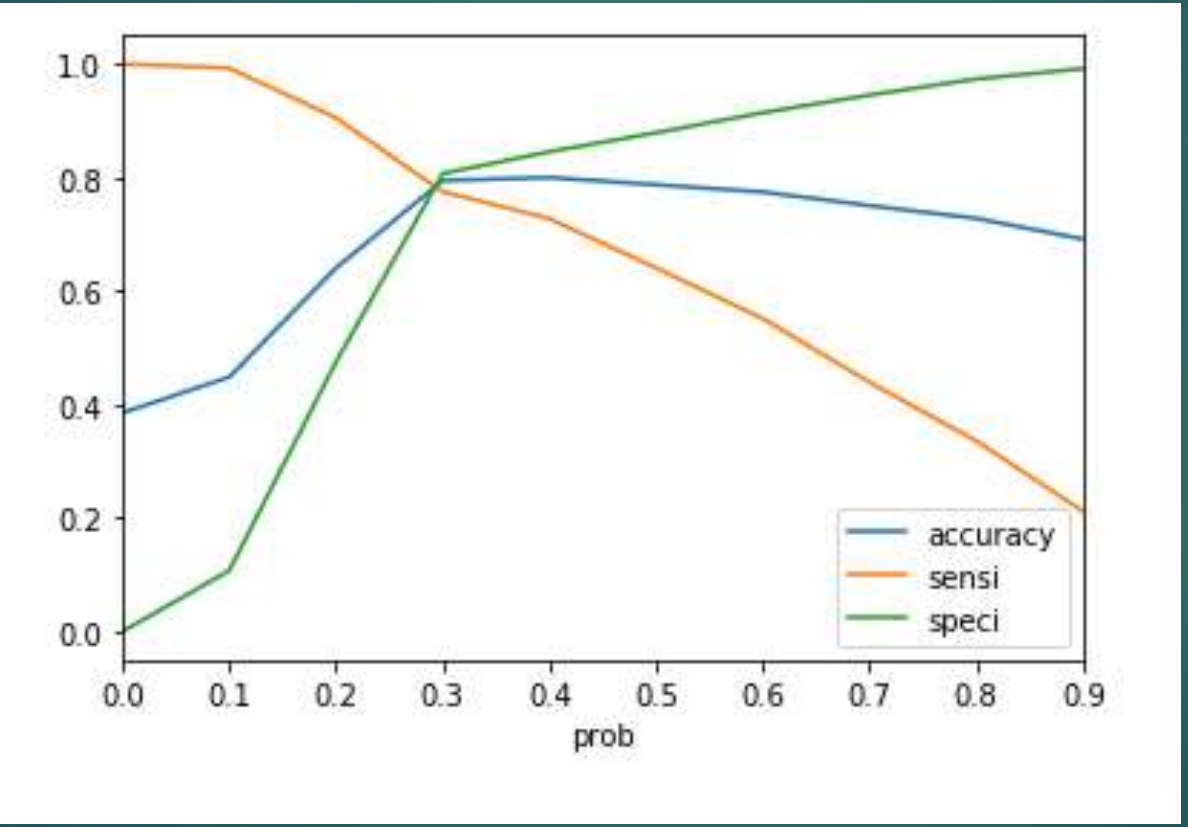
	Features	VIF
8	City_Mumbai	3.84
5	Specialization_Not_specialized	2.99
2	Lead Origin_Landing Page Submission	1.96
4	Lead Source_Olark Chat	1.90
3	Lead Origin_Lead Add Form	1.28
1	Total Time Spent on Website	1.27
7	What is your current occupation_Working Profes...	1.17
6	What is your current occupation_Student	1.03
0	Do Not Email	1.02

ROC Curve

- The ROC Curve should be a value close to 1. We are getting a good value of 0.84 indicating a good predictive model.



- From the curve below, 0.3 is the optimum point to take it as a cutoff probability.



- After testing test data on trained model we have final observations as below which states our train and test data prediction is almost similar.

Observation:

After running the model on the Test Data these are the figures we obtain:

- Accuracy : 80.02%
- Sensitivity : 77.76%
- Specificity : 81.31%

Final Observation:

Train Data:

- Accuracy : 79.45%
- Sensitivity : 77.47%
- Specificity : 80.69%

Test Data:

- Accuracy : 80.02%
- Sensitivity : 77.76%
- Specificity : 81.31%

Conclusion & Recommendation

- The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly. Then a cutoff of the probability is used to obtain the predicted value of the target variable.
- Optimum cut off is chosen to be 0.3 i.e. any lead with greater than 0.3 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.3 or less probability of converting is predicted as Cold Lead (customer will not convert)
- Our final Logistic Regression Model is built with 9 features.
- The top three categorical/dummy variables in the final model are '**Lead Add Form**' ,' **Working Professional**' and '**Lead Source_Olark Chat**' with respect to the absolute value of their coefficient factors.

X-Education has a better chance of converting a potential lead when:

- **The total time spent on the Website is high:** Leads who have spent more time on the website have converted.
- **Current Occupation is specified:** Leads who are working professionals have high chances of getting converted. People who were looking for better prospects like Unemployed, students and Business professionals were also good prospects to focus on.
- **When the Lead origin was Lead Add form** Leads who have responded/ or engaged through Lead Add Forms have had a higher chances of getting converted
- **Number of Total Visits were high** Leads who have made a greater number of visits have higher chances of getting converted.
- **When the last activity was SMS sent or Email opened** Members who have sent an SMS for enquiry or who have opened the email have a higher chance of getting converted.
- **Approaching the management specialized leads, will produce more conversion.**
- **Almost all the leads given tag 'Will revert after reading the email' gets converted , they have a high chance of conversion.**

Thank You