LEAD SCORE CASE STUDY

- Submitted by:

  ○ Nutan Patel
  ○ Shivam
  ○ Samantha

PROBLEM STATEMENT

X Education, an online course selling education company, needs assistance in recognising and selecting promising leads that will most likely convert to paying customers.

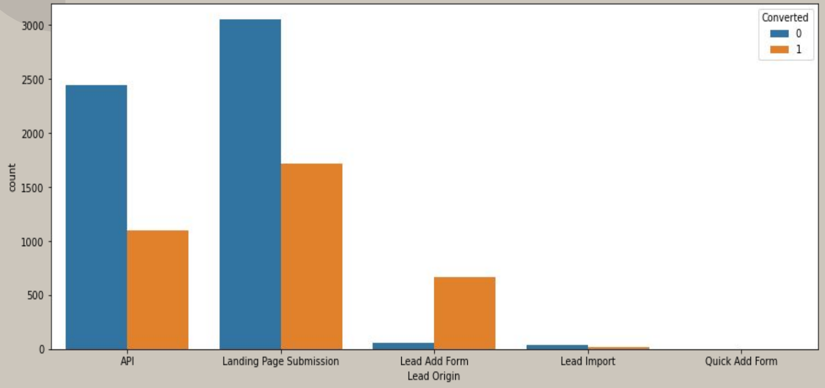
X Education gets a lot of leads but lead conversion rate is poor

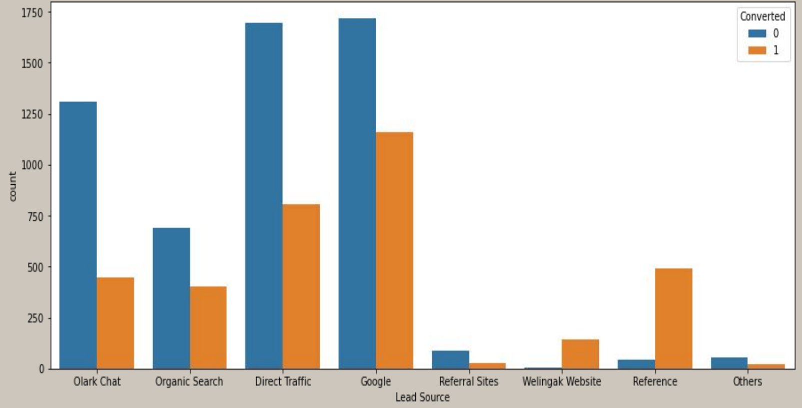
- So our aim is to analyse and identify variables which leads to higher customer conversion rate. This is done by building a logistic regression model that can assign a lead score to each of the leads such that customers with high lead score have a higher chance of conversion.

ASSUMPTIONS AND MODIFICATIONS ON THE DATA SET

- We dropped columns in which 45 or more percent of the data were missing.
- Since nearly half of the data were missing in few columns, we couldn't consider them for model building. We assumed that these columns will not have much influence on conversion rate.
- We imputed the missing values in continuous variables with median and those in categorical variables with mode. However, few categorical variables had missing value count more than that of mode, in such cases we replaced the missing values with 'Not mentioned'.
- In few of the categorical columns we had highly varying counts for levels ,so we grouped less frequent levels together to avoid model picking on the highly underrepresented dummy levels.
- We handled outliers in continuous variables by capping them at 99th percentile.

EDA RESULTS





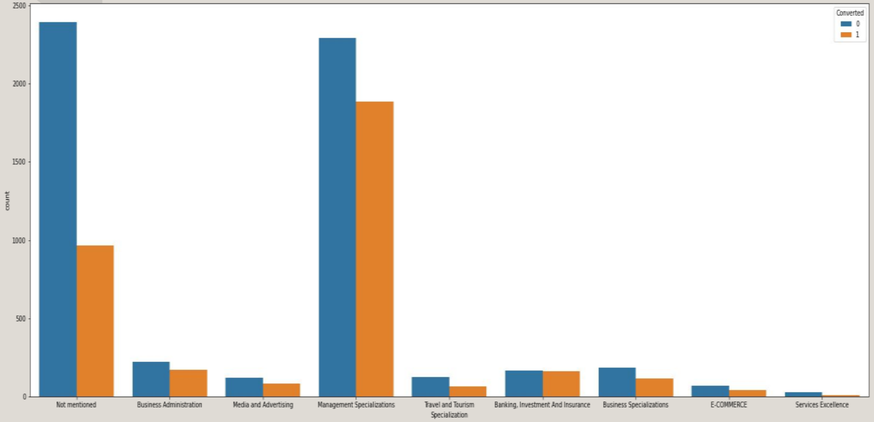
# Lead_Origin

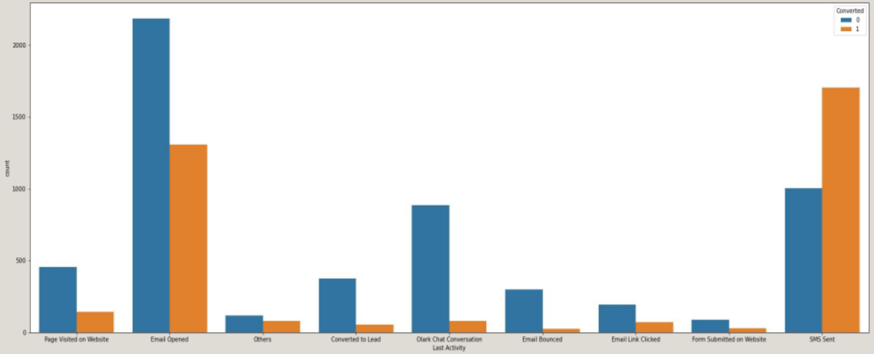
- 'Landing page submission' is the most common 'lead origin' and the number of converted leads is also higher for this level.
- The lead conversion rate is very high among the customers with 'Lead add form' origin.

# Lead_source

- Google and Direct Traffic are the two major lead sources.
- However 'Reference' has the highest lead conversion rate.

EDA RESULTS





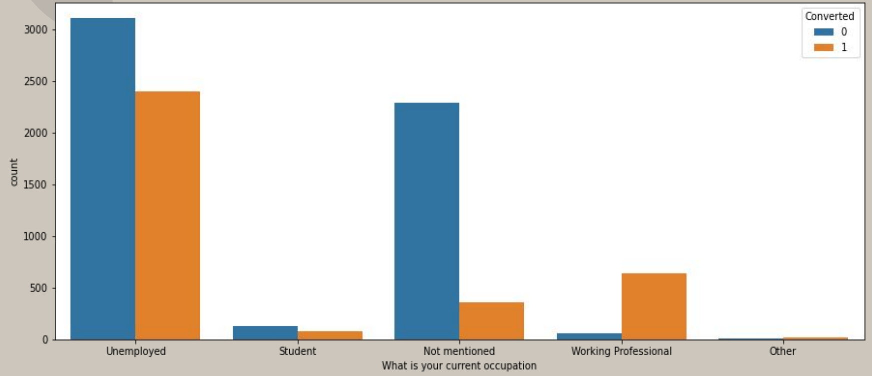
# Specialization

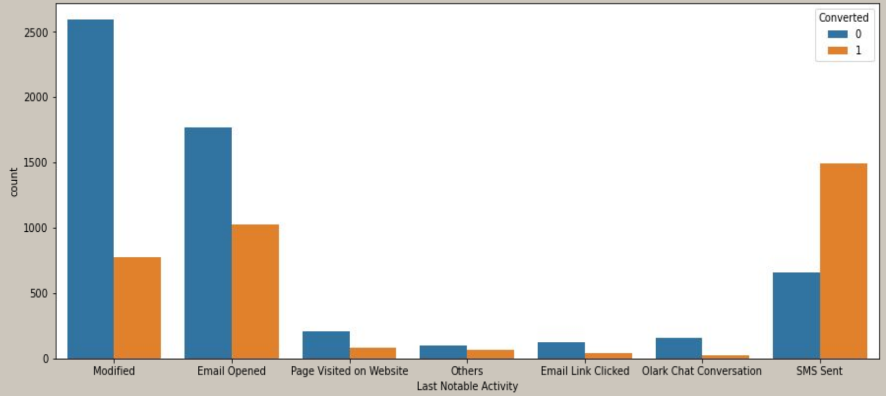
- Most of the customers prefer not to mention their specialization
- Among the customers who have mentioned their specializations,most of them fall into Management domain.

# Last Activity

- Lead conversion rate is very high among customers whose last activity is 'SMS sent'
- Lead conversion rate is very low among customers with last activity as 'Olark chat conversion'
- Most customer's last activity is 'Email opened'

EDA RESULTS





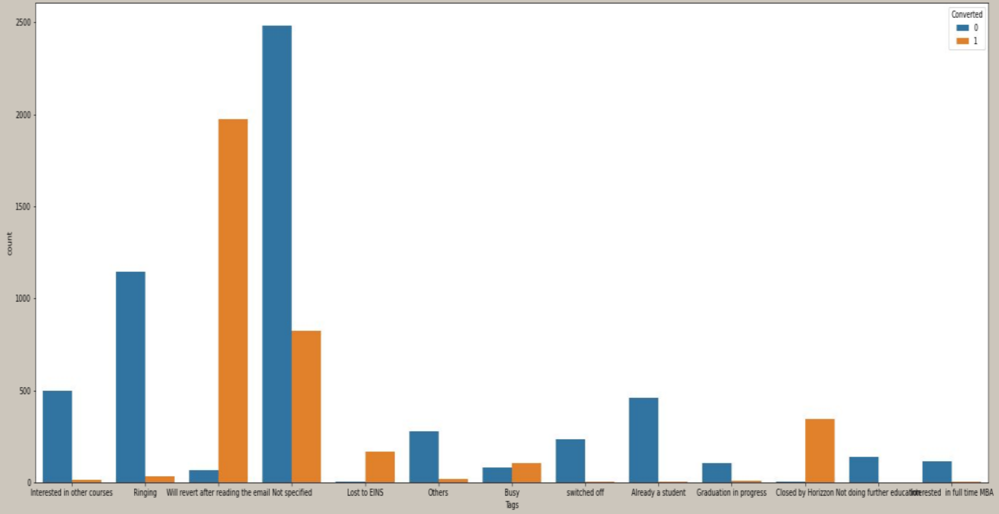
# What is your current occupation

- The number of unemployed people purchasing online courses are high.
- Most of the working professionals who check out the courses end up purchasing them.

# Last Notable Activity

- Most of the customers belong to 'modified' category of 'Last Notable Activity'
- The lead conversion rate is very high among people in 'SMS Sent' category

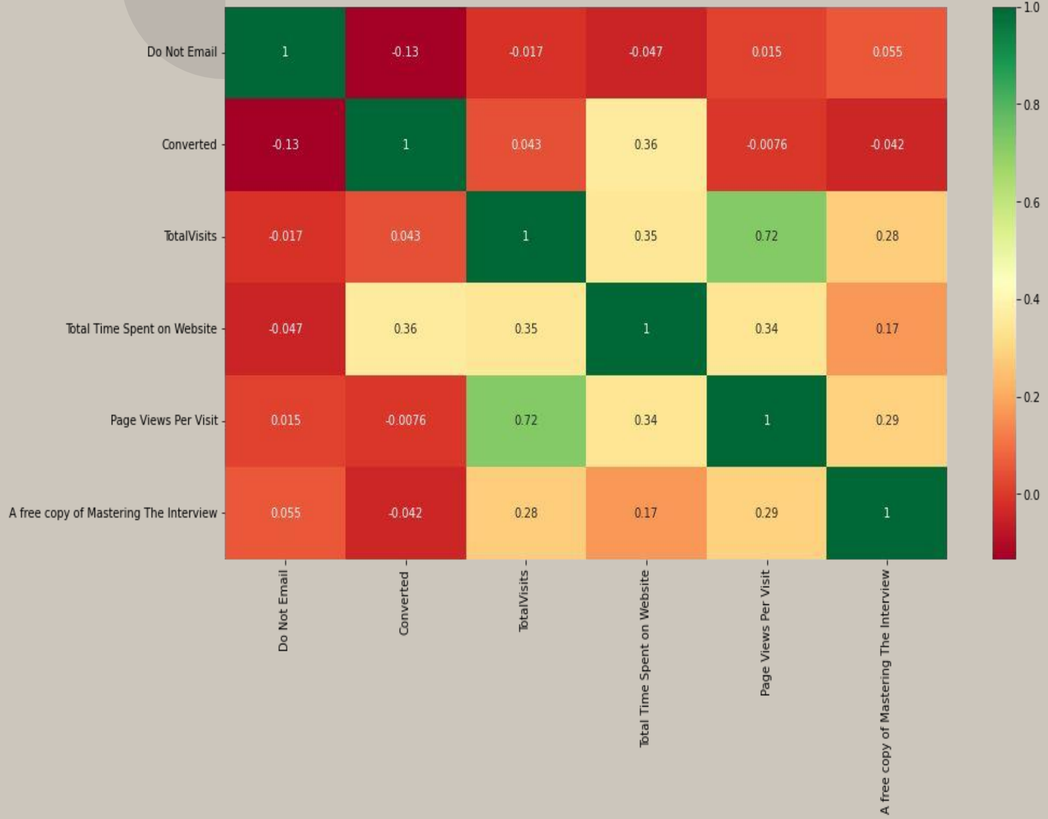
EDA RESULTS



# Tags

- Customers who are tagged as 'will revert after reading the email' or ‘Closed by Horizzon’ are more likely to purchase the courses.
- Leads in ‘Lost to EINS seems to have very high lead conversion rate.
- Leads who are tagged as ‘Switched off’,’Ringing’ and ‘Already a student’ are unlikely to purchase the course materials.

EDA RESULTS



	Do Not Email	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	A free copy of Mastering The Interview
Do Not Email	1	-0.13	-0.017	-0.047	0.015	0.055
Converted	-0.13	1	0.043	0.36	-0.0076	-0.042
TotalVisits	-0.017	0.043	1	0.35	0.72	0.28
Total Time Spent on Website	-0.047	0.36	0.35	1	0.34	0.17
Page Views Per Visit	0.015	-0.0076	0.72	0.34	1	0.29
A free copy of Mastering The Interview	0.055	-0.042	0.28	0.17	0.29	1

# Heat map with correlation coefficients of all the numerical variables

- Our target variable 'Converted' has a positive correlation with 'Total Time Spent On Website' with a correlation coefficient of 0.36
- The more a lead spends time on X Education’s website the more is the chances of lead conversion.
- Other numerical variables do not seem to have any strong correlation with our target variable ‘Converted’.

APPROACH

- After exploratory data analysis, we created dummy variables for categorical variables. Since k-1 dummies can take care of all the information contained in a categorical variable with k levels,We dropped one dummy from dummies created for each column. Our dataframe was later split into two,one for training the model and the other for validating the model. Numerical variables were normalized so that they wouldn't have any disproportionate effect on the model's results.We found that in the given data set there was a lead conversion rate of 38.45%.This is neither exactly 'balanced' (which a 50-50 ratio would be called) nor heavily imbalanced. So we didn't have to do any special treatment for this dataset .we had 54 features in in the dataset, since it was not feasible to use all the variables, we eliminated a few features using Recursive Feature Elimination (RFE) and selected 20 features for Model building

MODEL BUILDING

- We built a model using the features selected by RFE. In that model one feature was insignificant(had a p value less than 0.05), we dropped that feature and rebuilt the model in which all the features were statistically significant and had a VIF value below 2(there is no issue of multicollinearity). We used this model as our final model
- Using logistic regression curve's 'predict' function we computed the probabilities of lead conversion.
- The logistic curve gives just the probabilities and not the actual classification of lead conversion(converted/not converted). We chose 0.5 as an arbitrary cutoff wherein if the probability of a particular lead converting is less than 0.5,we would classify it as 'unconverted' and if it's greater than 0.5, we would classify it as 'Converted'

MODEL EVALUATION

- Since we were classifying the leads into two classes, we had some errors. The classes of errors that were there are: 'Converted' leads being (incorrectly) classified as 'not converted' and 'not converted' leads being (incorrectly) classified as 'Converted'
- To capture these errors, and to evaluate how well the model performs, we used 'Confusion Matrix'. Model sensitivity and specificity were calculated
- Our model had a sensitivity around 88.8% and specificity around 96.3%
- We plotted a ROC curve and found that our curve is towards the upper-left corner and there is a larger area under the curve (AUC) indicating that the model is good.
- Our model had very high specificity value but sensitivity value was relatively lower.So we needed to find an optimum cut-off to have a good balance between these two. For this we plotted accuracy sensitivity and specificity for various probabilities and picked 0.3 as optimum cut-off.
- lead score for every lead in the dataframe were calculated by multiplying their conversion probability with 100.

RESULTS & CONCLUSIONS

- Evaluation metrics before and after choosing the optimum cut-off for conversion probability.
  - Earlier with 0.5 cut-off
  - accuracy: 93.4%   sensitivity : 88.8% specificity : 96.3%
  - After choosing the optimum cut-off at 0.3
  - accuracy : 93% ->sensitivity : 93.7%   specificity : 92.6%
- Predictions were made on test set and model's performance was evaluated
  - accuracy : 93.3%   sensitivity : 93%   specificity : 93.5%

- Model is able to correctly predict ~93% of the converted labels

- Model is able to correctly predict ~93% of the unconverted labels

RECOMMENDATIONS

X education can use our model and give a lead score for every lead. Once the leads have been scored, the sales team can focus on leads with scores above 30(since our probability cut off was 0.3) to ensure higher lead conversion.

- Sales team doesn't want to miss calling leads tagged as Lost to EINS (coef 6.3616),Closed by Horizon (coef 5.5389) or lead source with Welingak Website (coef 3.7183) as they have higher coefficient values in our model. Higher coefficient values of these dummy variables indicate that there is a higher chance of lead conversion among leads with these dummies. Sales team may avoid calling leads who are tagged as switched off (coef -5.7080),Ringing (coef -4.5025) or Already a student (coef -4.3912),as their coefficient values are highly negative in our model and they will affect the conversion probability negatively.