LEAD SCORING CASE STUDY

By: *Sachin Jangid*
Divesh Bharti
Nizamulla Sharieff
Batch: DS C46 July 2022

Business Understanding:

- An education company named **X Education sells online courses to industry professionals**. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- 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%.

Problem Statement:

- Although, X Education gets a lot of leads, its lead conversion rate is very poor. X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Goal 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.*

Solution Approach

Reading & Understanding Data

- Check shape, missing values, and summary statistics
- Treat for missing values
- Outlier treatment
- Check for Target Imbalance (Converted Column)

EDA

- For Numeric and Categorical features
  - Univariate Analysis
  - Bivariate Analysis v/s Converted (response)
- Correlation matrix between numeric features

Preparing data for Modelling (feature Engineering)

- Convert Categorical feature to numeric:
  - Binary categorical -> 0 & 1 encoding
  - Multilevel categorical -> Dummy variable creation
- Train and Test Split (70% Train & 30% Test data)
- Numeric feature -> Standardize to have all features on same scale

Model Building

- Build Logistic Regression Model on Train Data
- RFE- Auto Tune to select 20 features
- Use Stats model to fine tune model

Model Evaluation and Prediction

On Training data

- Make Predictions using model built
- ROC Curve
- Tune cut off point for best performance
- Make predictions based on selected cut off
- Check performance metrics

On Unseen/Test data

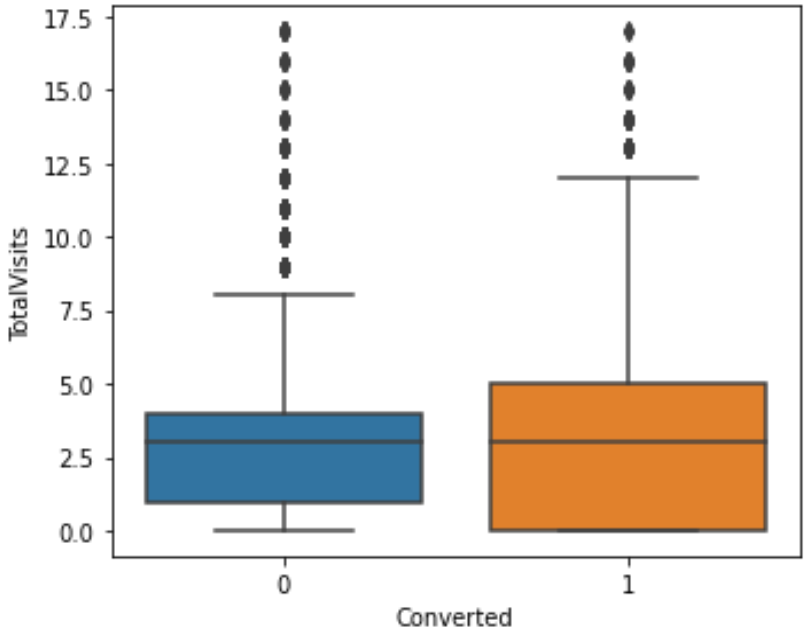
- Make Predictions on Test set
- Use optimal cut off point to predict
- Compare model performance on Train and Test data

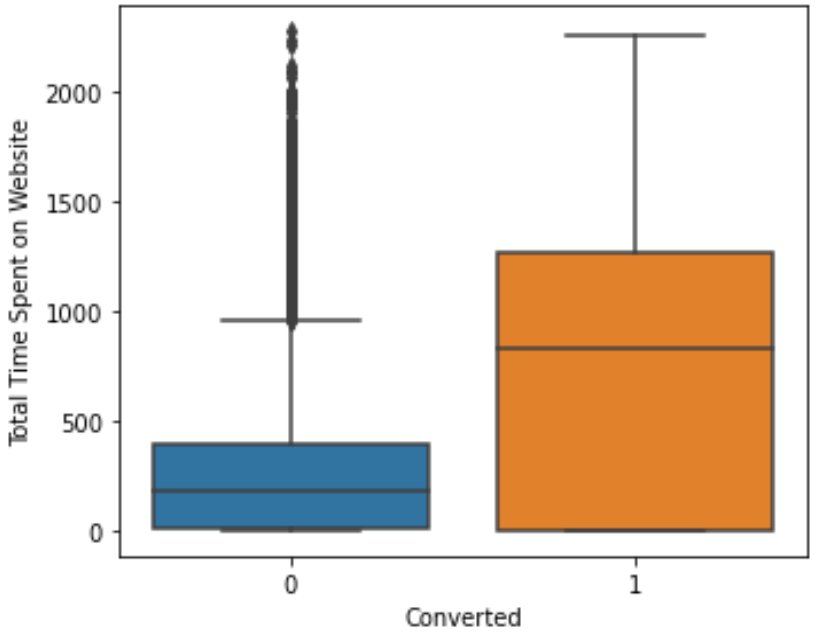
Interpreting Results in terms of Business

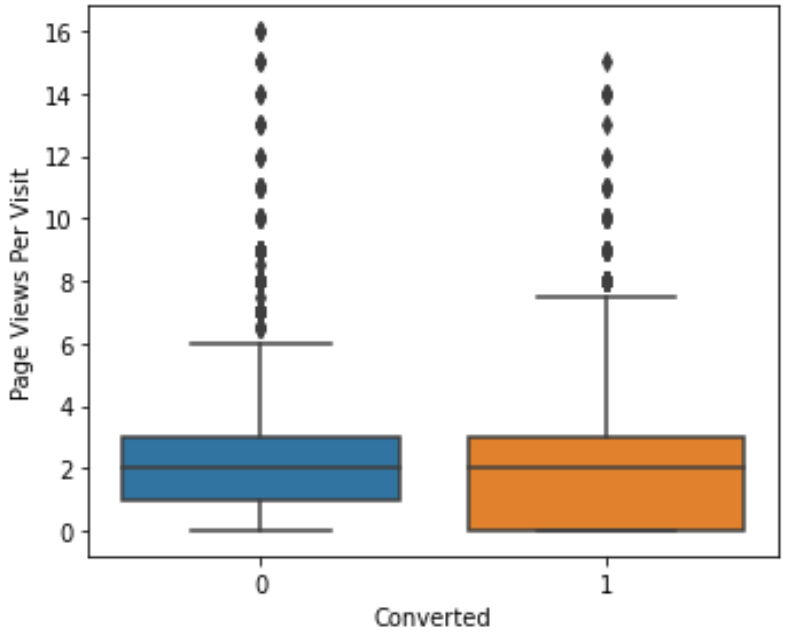
Data Cleaning to Solve Business Problem

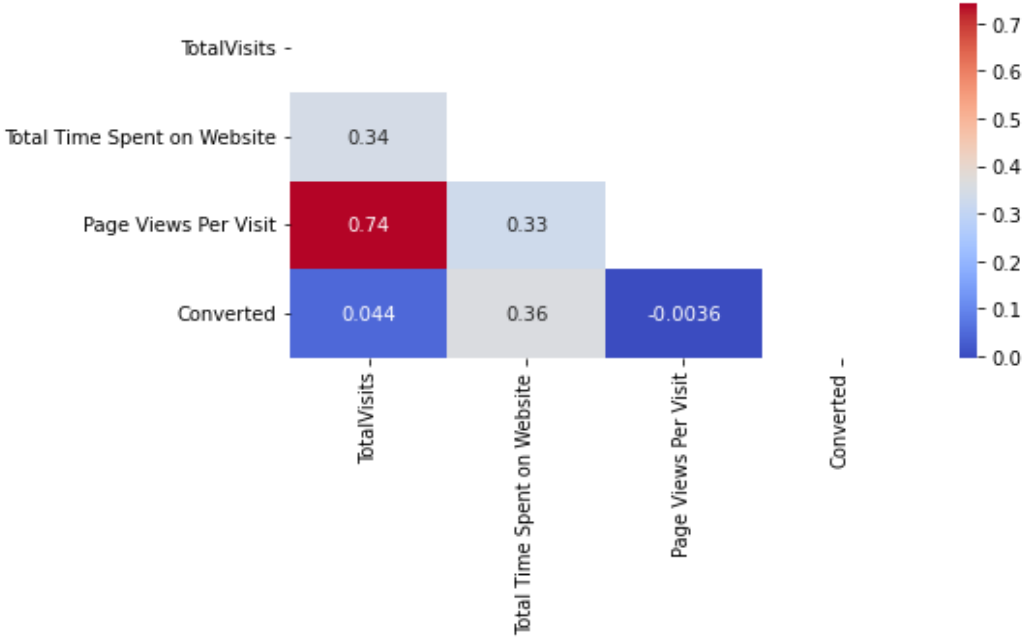
- **Raw Data:** 9240 rows and 37 features, of which “Converted” is the Target column
- Treated “Select” as missing in the entire dataset
- Checked for duplicates- not found
- Dropped features with >90% same class, as this would not help predict target
- **Missing value Treatment:** There are multiple columns with missing values
  - Dropped columns with >45% values missing
  - And for the rest imputed with mode/ separate class (categorical features), median (numeric features)
- **Outlier Treatment:** Two of Three numeric features have outliers (TotalVisits & Page Views Per Visit)
  - Removed upper & bottom 1% of values
- **Clean Data: 9157 rows remaining- 12 predictors and 1 target column**

EDA: Numeric Features



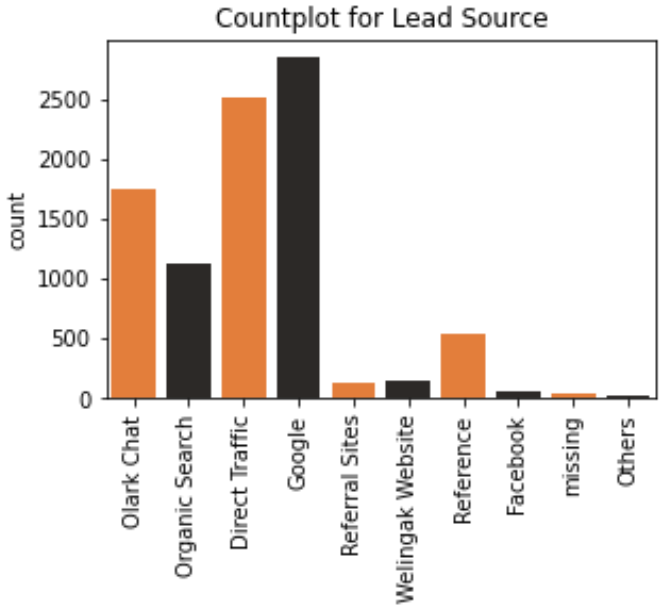


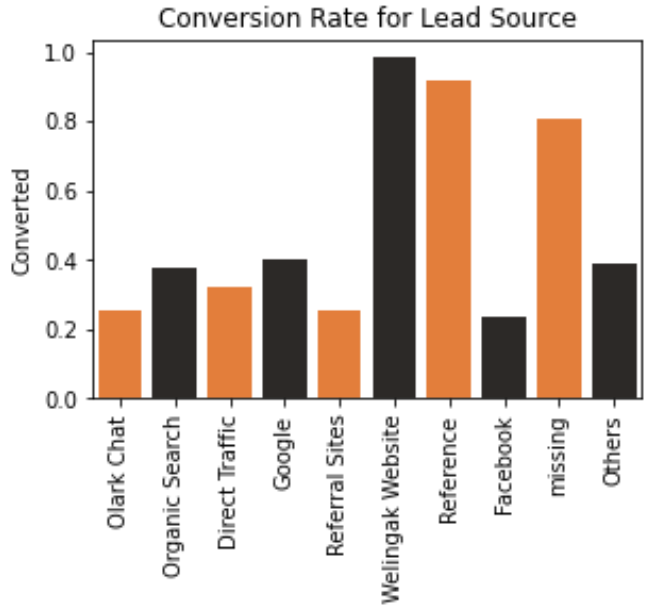




- Total Time Spent on website- converted leads spend more time on website. Therefore, website should have engaging content as it drives conversion
- Total visits & Page views per visit- Medians are same for converted and not converted leads, no significant pattern observed
- **Correlation Matrix:** High positive correlation between Page Views Per Visit and TotalVisits. *And, only Total time spent on website seems to have positive correlation with Converted response*

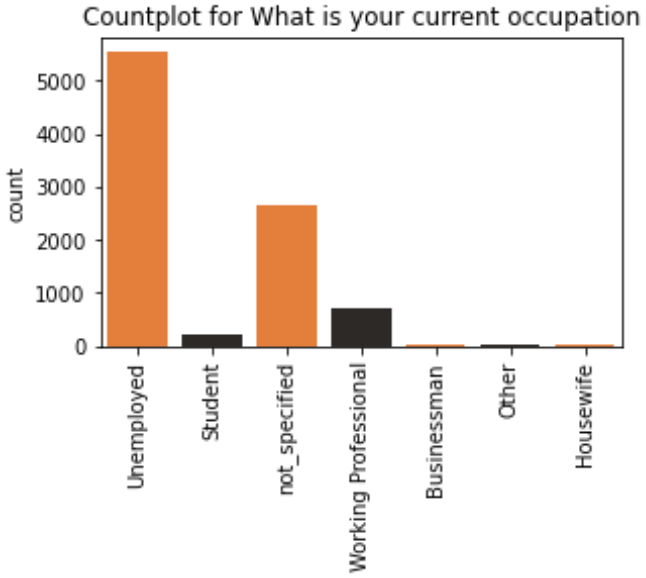
EDA: Categorical Features

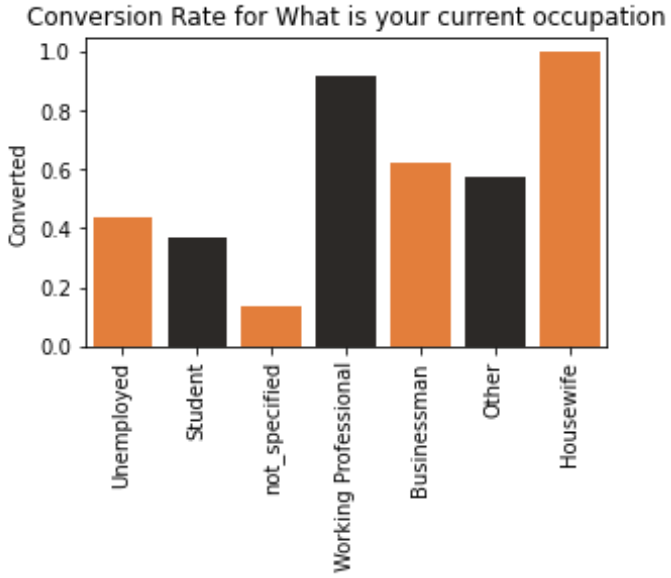




Lead Source:

- Reference and Welingak website has highest amount of conversions but lead volume is low
- So, increase conversion for Google, Direct traffic, Organic search, Olark chat leads, and increase lead volumes for Reference and Welingak website

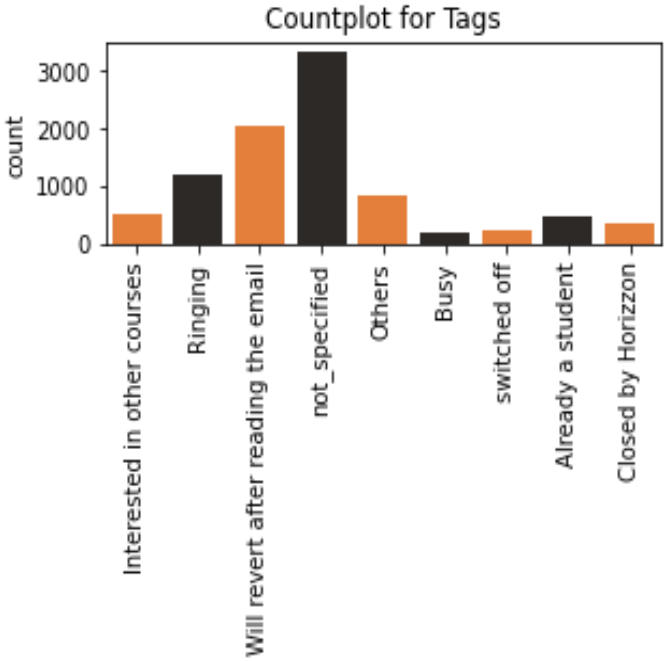


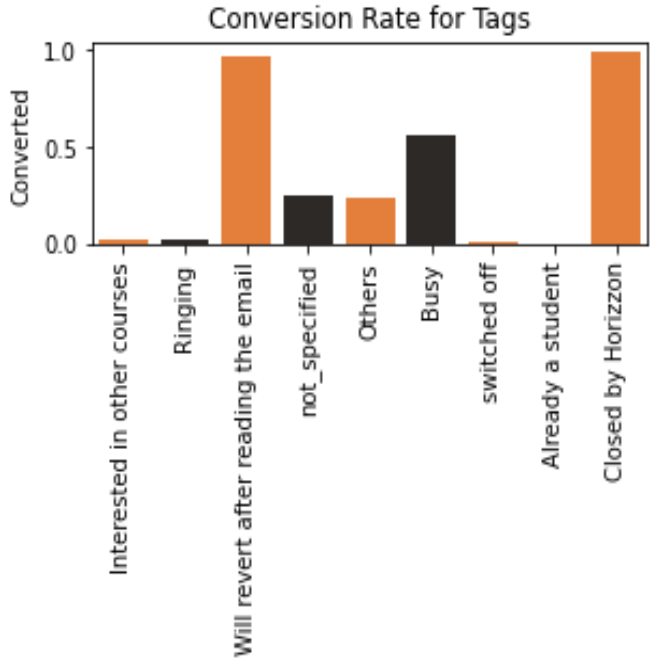


What is your current occupation:

- Working professionals and housewife have highest conversions but less amount of leads
- So, we should try improve conversions for unemployed and get more leads for Working professional/Housewife buckets

EDA: Categorical Features

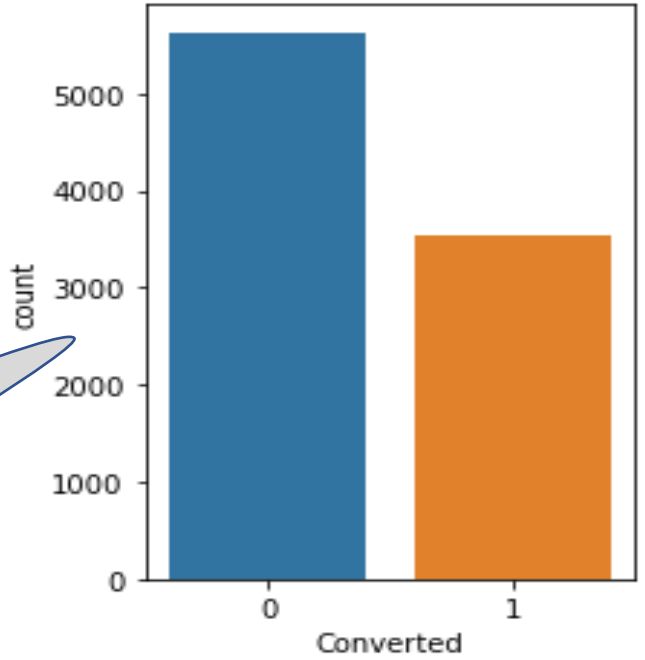


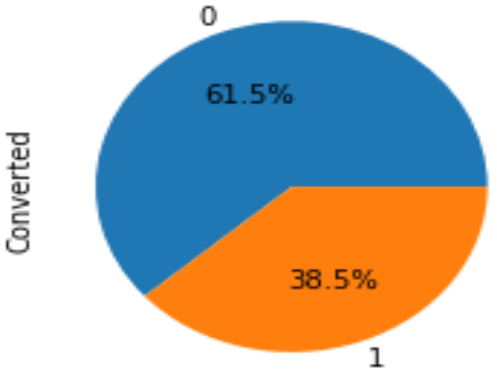


Tags:

- Avoid leads tagged Already Student- as course is intended for working professionals as students would not have paying capacity
- Focus on leads tagged “will revert after reading email” & “Closed by Horizon” as these have high conversion

Target Imbalance Check





38.5% of data has converted leads.
Looks good for modelling!!!

Feature Engineering & Model Building

Categorical Features:

- Binary class- encode with 0's & 1's
- Multiclass- Create n-1 Dummy features (n= number of classes)

Train & Test Split: 70% Train and 30% Test data. And, separate target column as y and predictors as X

Standardize Numeric Features: Different features have varied ranges. Hence, for the model to converge faster to global minima and easy interpretation we bring all features to similar range (-3 to 3)

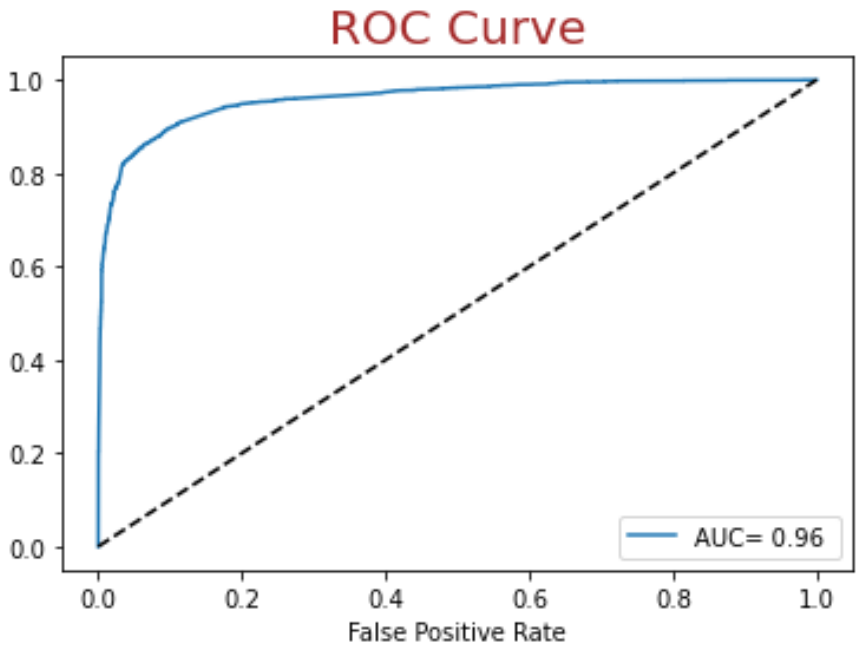
Finally, after all data preparation we have 59 predictors. Used RFE to select 20 significant features.
This is further fine tuned using stats model until we have features with p-value <5% and VIF <5

Model Evaluation using ROC:

Predict Converted probabilities for the train data and plot ROC Curve.

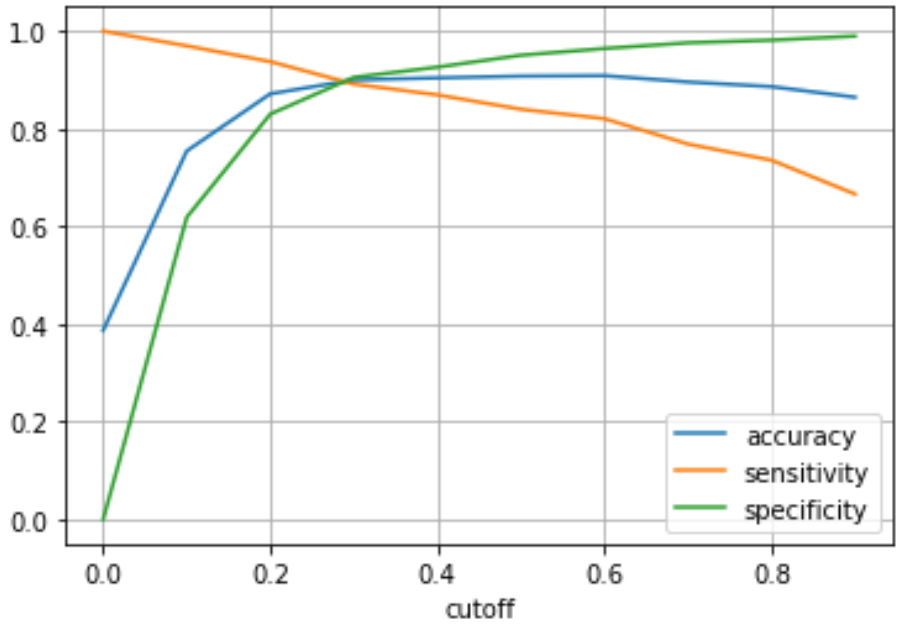
Area under curve is 96%. This indicates model built is good.

Assign Lead Score between 0-100 (prob. multiplied by 100)



Find Optimal Threshold | Evaluate Model on Unseen/Test Data

Finding optimal threshold



Model Perf. on Train data	Model Perf. on Test Data
Accuracy- 90%	Accuracy- 90.3%
Sensitivity- 89.2%	Sensitivity- 90%
Specificity- 90.5%	Specificity- 90.5%
Precision 85.6%	Precision 85.4%

Find optimal threshold for better sensitivity:

- From the curve we see 0.3 is optimal point where accuracy, sensitivity and specificity are balanced

Model has generalized well as holds good on Test Data. And, can be deployed in production to increase conversions!!

Interpreting Model Results and Business Impact

features	coef
Tags_Closed by Horizzon	5.766688
Lead Source_Welingak Website	5.704647
Tags_switched off	-4.360778
Tags_Will revert after reading the email	3.985228
Tags_Already a student	-3.650442
Tags_Ringing	-3.552854
Tags_Interested in other courses	-2.813063
Last Notable Activity_SMS Sent	2.776233
const	-2.373553
Lead Source_Reference	1.627135
Lead Source_Olark Chat	1.151420
Total Time Spent on Website	1.062775
What is your current occupation_Working Professional	0.886492
Last Activity_Email Opened	0.815780

Modelled Features (13 var) impacting Target variable in decreasing order
(Sign is indicative of positive or negative impact)

Top 3 features contributing towards probability of conversion:

- *Tags*- a) Positive impact- Closed by Horizon, Will revert after reading the email b) Negative impact- Already a student, switched off, Ringing, Interested in other courses
- *Lead Source*- Positive impact- Welingak Website, Reference, Olark Chat
- *Last Notable Activity*- Positive impact- SMS Sent

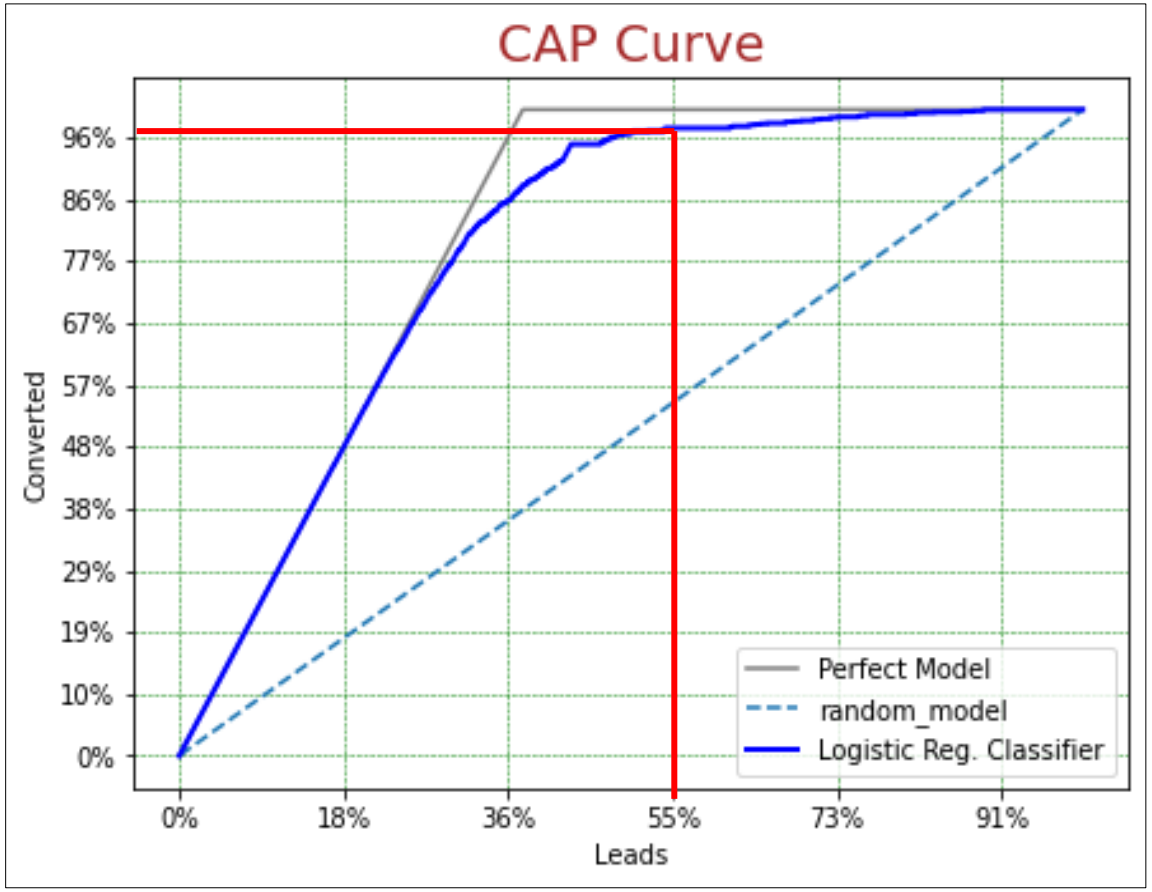
Top 3 features to increase probability of conversion:

- *Tags_Closed by Horizon*
- *Lead Source_Welingak Website*
- *Tags_Will revert after reading the email*

Interpreting Model Results and Business Impact (contd..)

Business Recommendations:

- Aggressive follow up on leads tagged *Will revert after reading the email*
- Target leads from Welingak website and Reference Lead Sources
- As X-Education offers professional courses- target leads who marked themselves as working professionals
- Target leads that have last notable activity as SMS Sent, and leads that spend more time on website
- Avoid leads with student's tag- maybe they do not have paying capacity (early stage)
- Avoid leads tagged- *Interested in other courses/ Switched off/ Ringing*



Business Impact driven by Model:
Just by attempting 55% of Leads prioritized by model. We can capture ~96% of Sales. So, Logistic classifier helps optimize sales team accordingly, and increase efficiency on lead churning.

Thank You