LEAD SCORING CASE STUDY

ESWARA REDDY KAMIREDDY

POOJA RATHOD

Problem Objective

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, although X Education gets a lot of leads, its lead conversion rate is very poor which is only 30 %. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Goals of the Case Study

There are quite a few goals for this 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. - The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Solution Methodology

Data Cleaning and Preparation

- Identify the data quality and clean based on requirement
- Handle null values, Data Imputation wherever necessary
- Outlier Analysis and treatment
- Dummy Variable creation

Solve problem

- Train Test Split data
- Logistic Regression Model Building using RFE
- GLM is used from stats
- Model Evaluation
- Model Prediction

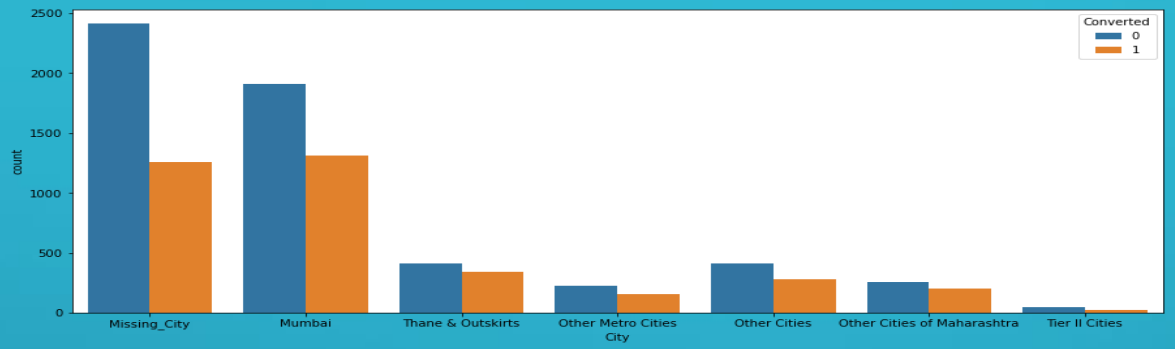
Identify influencing features

- Identify based on Logistic regression model
- Draw Conclusion and recommendations for modelData Cleaning and Preparation - Identify the data quality and clean based on requirement - Handle null values based on converted rate without removing data points - Data Imputation - Outlier Analysis and treatment Solve problem - Variable Processing - Univariate Analysis (EDA) - Train Test Split data - Logistic Regression Model Building Identify influencing features - Identify based on Logistic regression model - Draw Conclusion and recommendations for model.

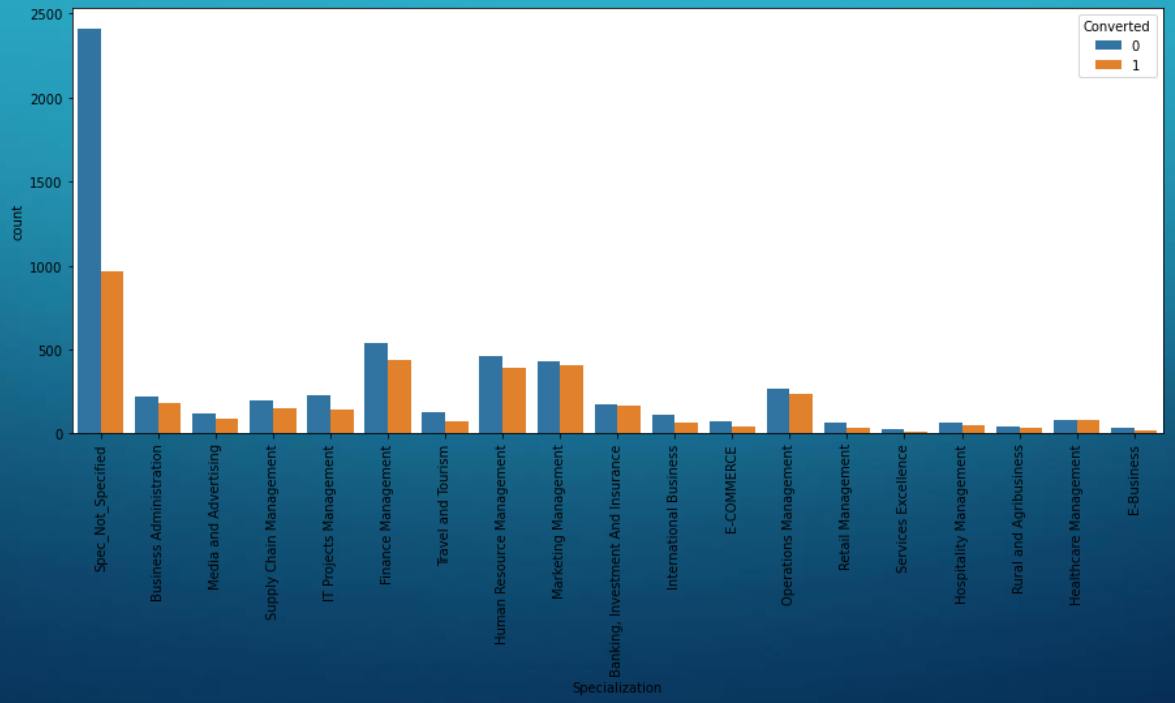
Data Cleaning and Preparation

- Replaced Select with NaN
- Dropping unnecessary columns and single unique feature columns.
- Imputed Values
  - with highest count in particular column
  - Added new value ‘not specified’ when missing value can lead to skewing the data
  - Added ‘Others’ when lot unique values are present.
- Highly skewed columns were dropped.
- Outliers are handled by capping at 99 percentile
- Median is used to impute numerical columns
- Dummy variables are created

Data Cleaning and Preparation

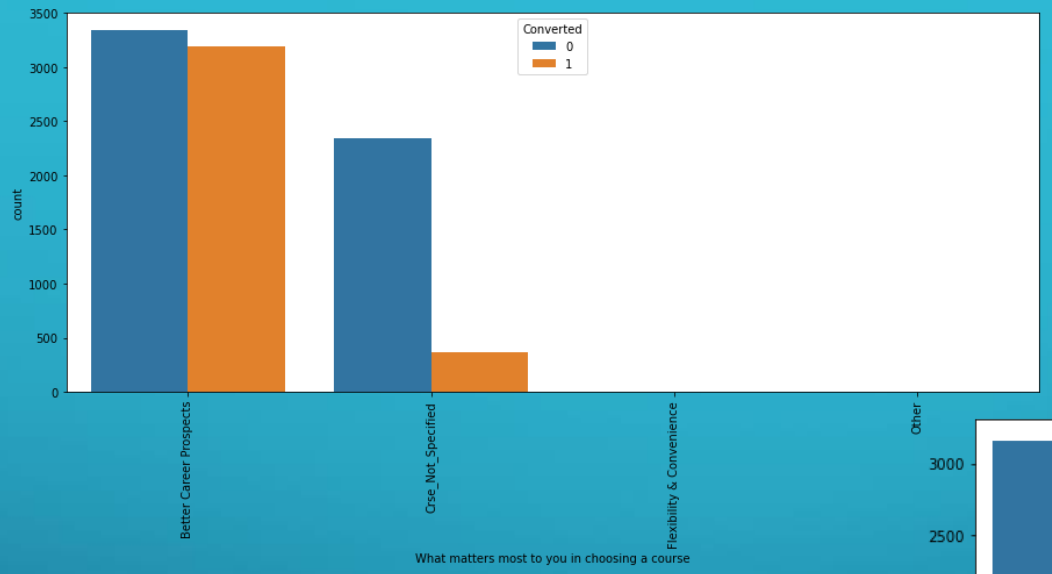


Maximum conversion is city of Mumbai



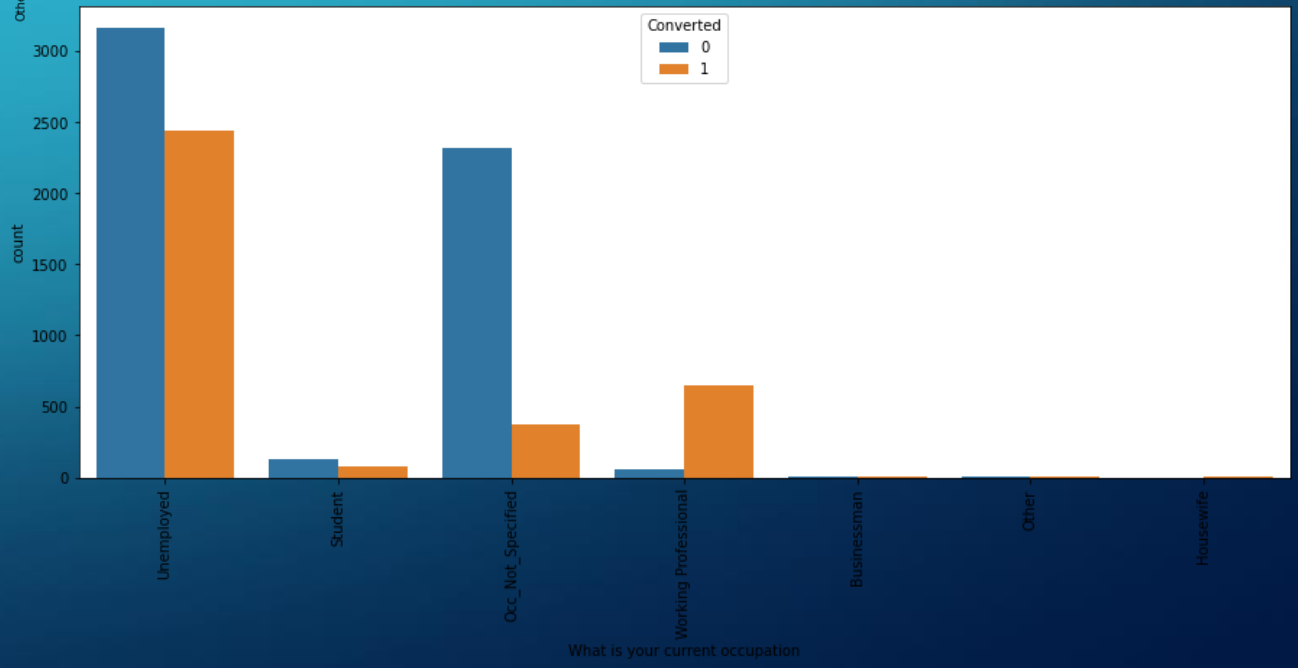
Maximum conversion is city of Mumbai

Data Cleaning and Preparation

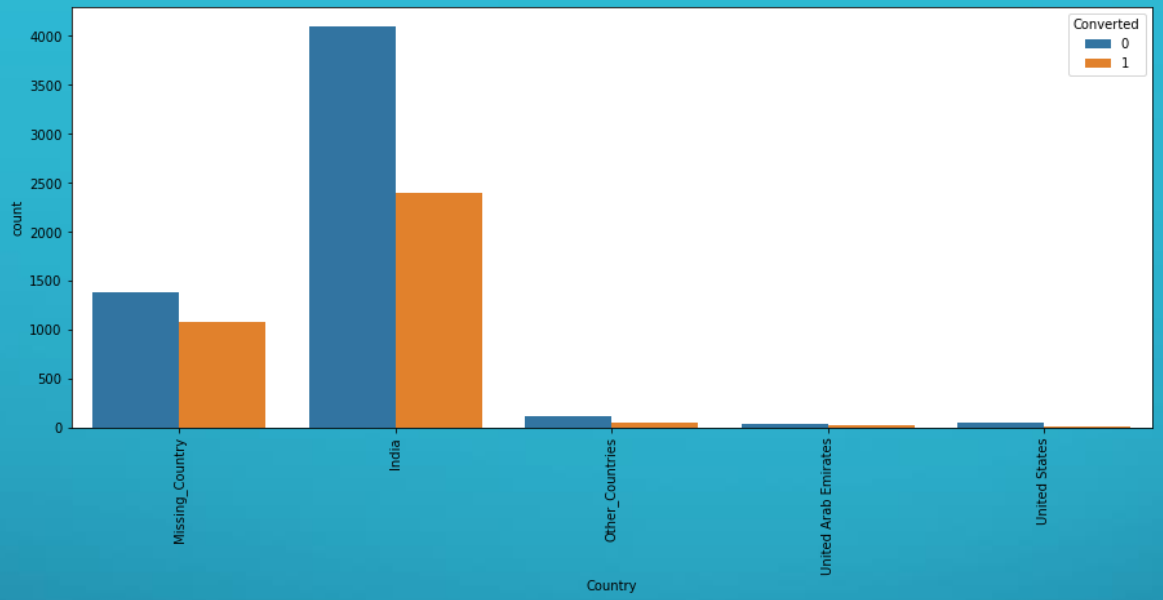


Lead looking for better career prospects have high probability of conversion

Unemployed and working profession have high probability of conversion

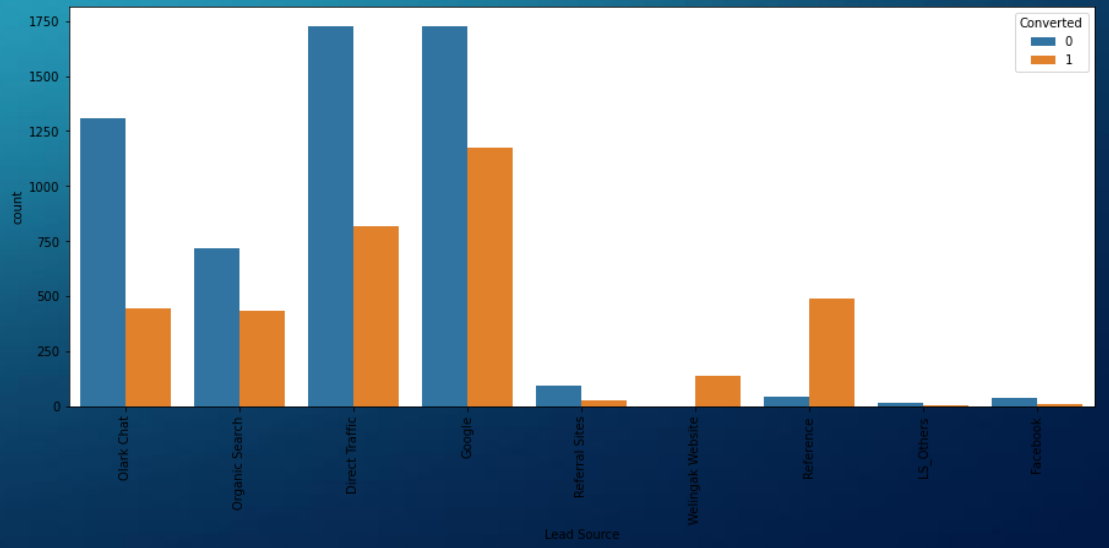


Data Cleaning and Preparation

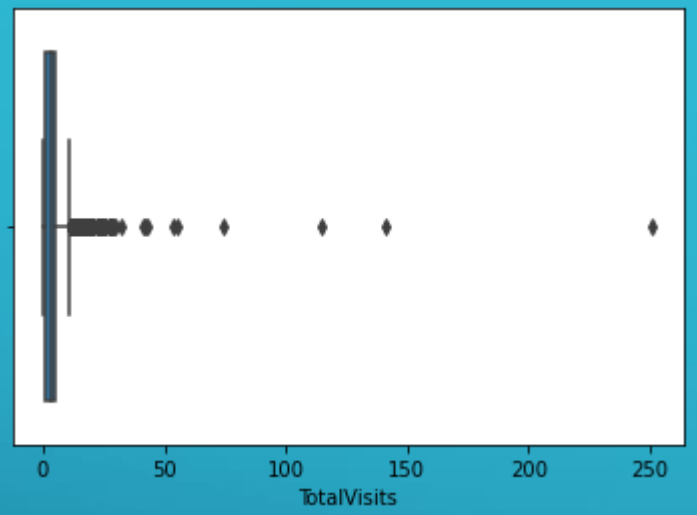


India has max conversion

Google is the best lead source among all categories in the lead source

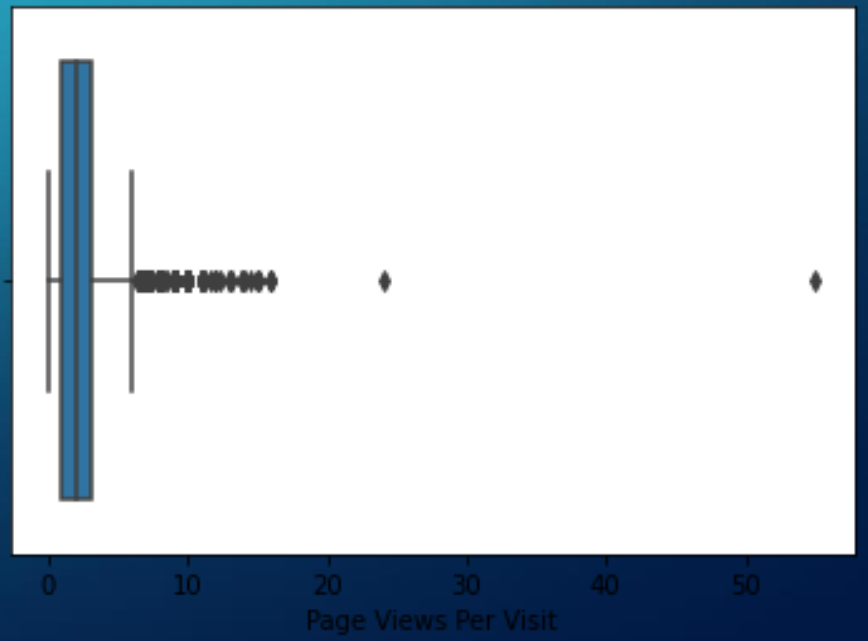


Data Cleaning and Preparation



Total visits has outliers and are handled using capping to 99%

Page Views Per Visit has outliers and are handled using capping to 99%



Model Building

- For Model building we need to scale and split data into train and test dataset.

  ```
  X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=100)
  ```

- We will be using Logistic Regression for building the model.

  ```
  logreg = LogisticRegression()
  ```

- Variable selection done through RFE(recursive feature elimination) and further we remove features with high p value and VIF value.
- GLM is used for Stats

  ```
  X_train_sm = sm.add_constant(X_train[col])
  logm1 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
  res = logm1.fit()
  res.summary()
  ```

- Analyzing various parameters for train dataset Specificity, Sensitivity, Accuracy, Precision and Recall for train data.
- Plot the ROC Curve which shows trade off between sensitivity and specificity

  ```
  fpr, tpr, thresholds = metrics.roc_curve( y_train_pred_final.Converted, y_train_pred_final.Converted_Prob, drop_intermediate = False )
  draw_roc(y_train_pred_final.Converted, y_train_pred_final.Converted_Prob)
  ```

Logistic Regression Model

Using RFE and Manual feature elimination for features having high P-value and high VIF. We reached a final model.

INITIAL MODEL

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6452
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1279.9
Date:	Tue, 12 Jul 2022	Deviance:	2559.8
Time:	20:20:30	Pearson chi2:	1.60e+04
No. Iterations:	22	Pseudo R-squ. (CS):	0.6069
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.9259	0.181	-16.141	0.000	-3.281	-2.571
Lead Source_Welingak Website	2.7216	0.753	3.613	0.000	1.245	4.198
Last Activity_SMS Sent	2.2484	0.121	18.545	0.000	2.011	2.486
What matters most to you in choosing a course_Cse_Not_Specified	-2.6311	0.148	-17.786	0.000	-2.921	-2.341
Tags_Busy	1.9295	0.277	6.974	0.000	1.387	2.472
Tags_Closed by Horizzon	9.4289	1.021	9.235	0.000	7.428	11.430
Tags_Lost to EINS	9.4761	0.766	12.368	0.000	7.974	10.978
Tags_Not doing further education	-1.3575	1.033	-1.314	0.189	-3.383	0.668
Tags_Ringing	-1.8406	0.280	-6.582	0.000	-2.389	-1.292
Tags_Tags_Not_Specified	3.4473	0.212	16.270	0.000	3.032	3.863
Tags_Will revert after reading the email	6.2606	0.237	26.454	0.000	5.797	6.724
Tags_Invalid number	-2.3810	1.037	-2.295	0.022	-4.414	-0.348
Tags_switched off	-2.3117	0.546	-4.237	0.000	-3.381	-1.242
Tags_wrong number given	-21.6894	1.3e+04	-0.002	0.999	-2.55e+04	2.54e+04
Last Notable Activity_Modified	-1.5260	0.121	-12.579	0.000	-1.764	-1.288
Last Notable Activity_Olark Chat Conversation	-1.1745	0.418	-2.810	0.005	-1.994	-0.355

Final MODEL

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6456
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1294.4
Date:	Tue, 12 Jul 2022	Deviance:	2588.7
Time:	20:20:31	Pearson chi2:	1.54e+04
No. Iterations:	8	Pseudo R-squ. (CS):	0.6051
Covariance Type:	nonrobust		

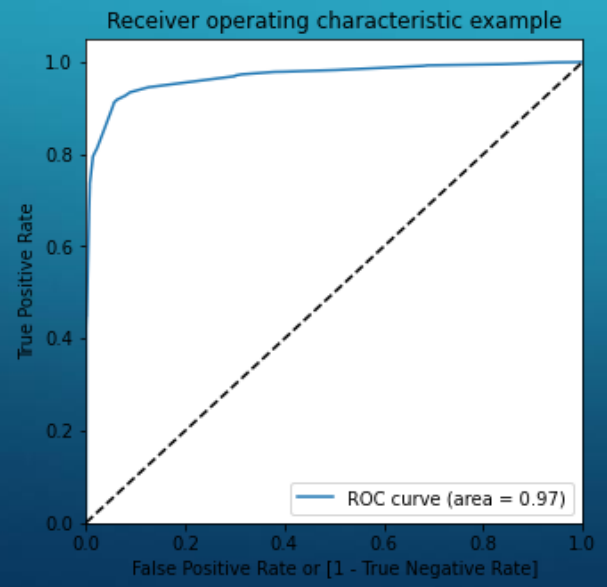
	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.2917	0.178	-18.449	0.000	-3.641	-2.942
Lead Source_Welingak Website	2.7459	0.752	3.653	0.000	1.273	4.219
Last Activity_SMS Sent	2.2303	0.119	18.783	0.000	1.998	2.463
What matters most to you in choosing a course_Cse_Not_Specified	-2.6132	0.146	-17.894	0.000	-2.899	-2.327
Tags_Busy	2.2952	0.271	8.462	0.000	1.764	2.827
Tags_Closed by Horizzon	9.7274	1.021	9.527	0.000	7.726	11.729
Tags_Lost to EINS	9.7491	0.766	12.726	0.000	8.248	11.251
Tags_Ringing	-1.4628	0.273	-5.363	0.000	-1.997	-0.928
Tags_Tags_Not_Specified	3.7538	0.209	17.989	0.000	3.345	4.163
Tags_Will revert after reading the email	6.5844	0.235	27.984	0.000	6.123	7.046
Tags_switched off	-1.9326	0.542	-3.567	0.000	-2.995	-0.871
Last Notable Activity_Modified	-1.4591	0.121	-12.101	0.000	-1.695	-1.223

Potting ROC Curve

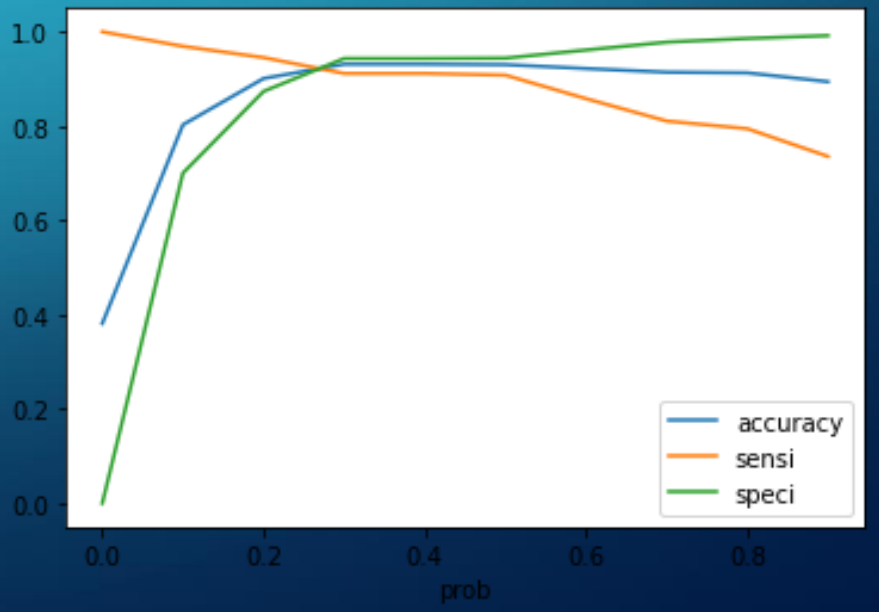
ROC curve demonstrates several things

- It shows trade off between sensitivity and specificity.
- The closer the curve follows left hand border and then the top border of the ROC space, this proves better accuracy of the test.
- The closer the curve comes to the 45-degree diagonal of the
- ROC space, the less accurate the test.

ROC Curve area is 0.97

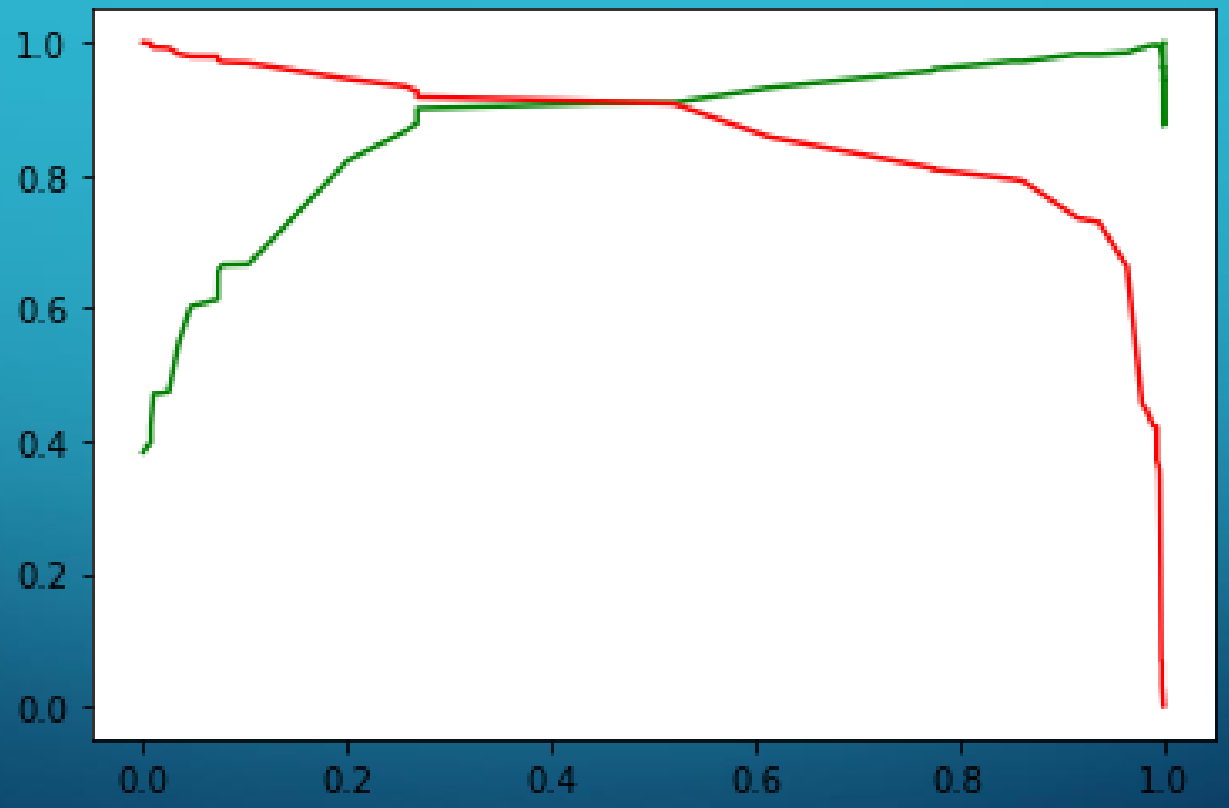


Optimum point is 0.25



Precision and recall tradeoff

- 0.42 is the cutoff from precision_recall_curve



Lead Score for varying cut off probability

- Lead score can used to predict incase if we want to change the cutoff

```
In [185]: y_pred_final['final_predicted'] = y_pred_final.Converted_Prob.map(lambda x: 1 if x > 0.42 else 0)
          y_pred_final['Lead Score'] = y_pred_final['Converted_Prob']*100
          y_pred_final.sort_values(by='Lead Score',ascending=False,inplace=True)
```

```
In [186]: y_pred_final.head()
```

Out[186]:

	ID	Converted	Converted_Prob	final_predicted	Lead Score
157	4830	1	0.999953	1	99.995343
520	3339	1	0.999952	1	99.995241
1329	4812	1	0.999952	1	99.995241
515	8412	1	0.999952	1	99.995241
2162	3736	1	0.999952	1	99.995241

Overall Accuracy

- Accuracy of the predicted model is 93.25

```
In [187]: # Let's check the overall accuracy.
          metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
Out[187]: 0.9325396825396826
```

Overall Accuracy

- Accuracy of the predicted model is 93.25

```
In [187]: # Let's check the overall accuracy.
          metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)

Out[187]: 0.9325396825396826
```

Recommendations

Top 3 variables that contribute the most towards the probability of a lead getting converted are:

- Tags_lost to EINS
- Tags_Closed by Horizzon
- Tags_will revert after reading the email

X education company needs to focus on following key aspects to improve overall conversion rate:

- Focus on the top 3 tags which are very positive for business.
- Focus on working professional & unemployed who have high conversion rate.
- Focus on rewards for Referrals as the conversion rate is high.