Lead Scoring Case Study Assignment

Tri Dao

Agenda

Data Input, Brief Overview and Cleaning

- Data Overview
- Data Input
- Data Cleaning

Visualization

- Some Visualizations for univariate and bivariate

Model Evaluation

Conclusion

- Scatter Plot

Data Input and Brief Overview

Data Input, Brief Overview and Data Cleaning

Data Overview

- 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. *When these people fill up a form providing their email address or phone number, they are classified to be a lead.* The company requires to build a model wherein we need to assign 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%.

- In other words, the company wants to understand the driving factors (or driver variables) behind lead conversion. The company can utilize this knowledge for its management action to optimize the business performance.

- Three csv files are provided:
  - *'Leads.csv' contains all the information of the client classified as leads.*
  - *'Leads Data Dictionary.xlsx' is data dictionary which describes the meaning of the variables.*

Data Input , Brief Overview and Data Cleaning

DATA INPUT

The two data csv files are imported using pandas.read_csv() function. It's also a good idea to see the data dimensions for further steps.

Reading and Understanding the Data

```
In [3]: df_leads = pd.read_csv('Leads.csv') # Import the input data
        df_leads.head()
```

Out[3]:

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	Get updates on DM Content	Lead Profile	City	Asymmetrique Activity Index	Asymmetriqu Profile Inde
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	Select	Select	02.Medium	02.Mediu
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	...	No	Select	Select	02.Medium	02.Mediu
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	...	No	Potential Lead	Mumbai	02.Medium	01.Hig
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	...	No	Select	Mumbai	02.Medium	01.Hig
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	...	No	Select	Mumbai	02.Medium	01.Hig

5 rows × 37 columns

Data Inspection

```
In [4]: df_leads.shape # Examine the data shape/dimensionality
Out[4]: (9240, 37)
```

In the next step, a brief overview output of provided data is a good start using common statistical figures while also checking for data types and null values.

```
In [5]: df_leads.describe() # Gain the first impression of data distributions
```

Out[5]:

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

```
In [6]: df_leads.info() # Checking on data type and potential null errors
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object
 11  Country                                        6779 non-null   object
 12  Specialization                                 7802 non-null   object
 13  How did you hear about X Education             7033 non-null   object
 14  What is your current occupation                6550 non-null   object
 15  What matters most to you in choosing a course  6531 non-null   object
 16  Search                                         9240 non-null   object
 17  Magazine                                       9240 non-null   object
 18  Newspaper Article                              9240 non-null   object
 19  X Education Forums                             9240 non-null   object
 20  Newspaper                                      9240 non-null   object
 21  Digital Advertisement                          9240 non-null   object
 22  Through Recommendations                        9240 non-null   object
 23  Receive More Updates About Our Courses         9240 non-null   object
 24  Tags                                           5887 non-null   object
 25  Lead Quality                                   4473 non-null   object
 26  Update me on Supply Chain Content              9240 non-null   object
```

Data Input, Brief Overview and Data Cleaning

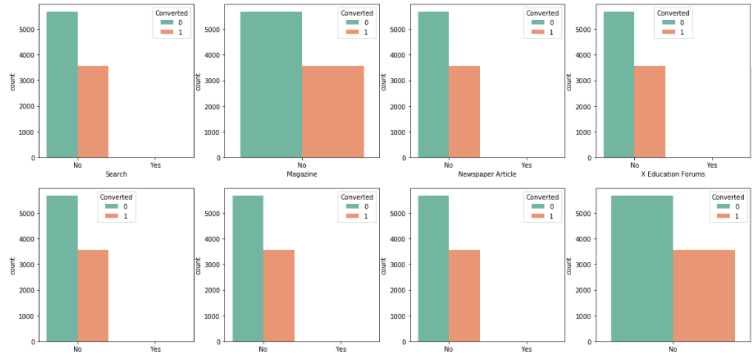
DATA CLEANING

This step includes the following actions:

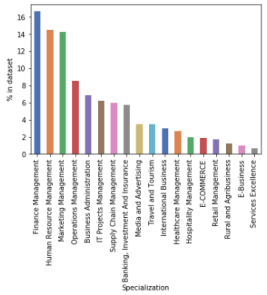
- Check for null and missing values with proposed imputation (e.g., mean and mode imputation for numerical/categorical correspondingly, except for the case of high frequency categorical data (e.g., Specialization)).
- Check for imbalanced variables (e.g., Search, Magazine, etc.).
- Check for outliers with corresponding proposals for actions (e.g., columns TotalVisits, Page Views Per Visit).

Imbalanced Variables

```
In [25]: # Visualzing  variables for imbalancing
fig, axs = plt.subplots(3, 4, figsize = (20, 15))
sns.countplot(x = "Search", hue = "Converted", data = df_leads, ax = axs[0,0],palette = 'Set2')
sns.countplot(x = "Magazine", hue = "Converted", data = df_leads, ax = axs[0,1],palette = 'Set2')
sns.countplot(x = "Newspaper Article", hue = "Converted", data = df_leads, ax = axs[0,2],palette = 'Set2')
sns.countplot(x = "X Education Forums", hue = "Converted", data = df_leads, ax = axs[0,3],palette = 'Set2')
sns.countplot(x = "Newspaper", hue = "Converted", data = df_leads, ax = axs[1,0],palette = 'Set2')
sns.countplot(x = "Digital Advertisement", hue = "Converted", data = df_leads, ax = axs[1,1],palette = 'Set2')
sns.countplot(x = "Through Recommendations", hue = "Converted", data = df_leads, ax = axs[1,2],palette = 'Set2')
sns.countplot(x = "Receive More Updates About Our Courses", hue = "Converted", data = df_leads, ax = axs[1,3],palette = 'Set2')
sns.countplot(x = "Update me on Supply Chain Content", hue = "Converted", data =df_leads, ax = axs[2,0],palette = 'Set2')
sns.countplot(x = "Get updates on DM Content", hue = "Converted", data = df_leads, ax = axs[2,1],palette = 'Set2')
sns.countplot(x = "I agree to pay the amount through cheque", hue = "Converted", data = df_leads, ax = axs[2,2],palette = 'Set2
sns.countplot(x = "A free copy of Mastering The Interview", hue = "Converted", data = df_leads, ax = axs[2,3],palette = 'Set2')
plt.show()
```

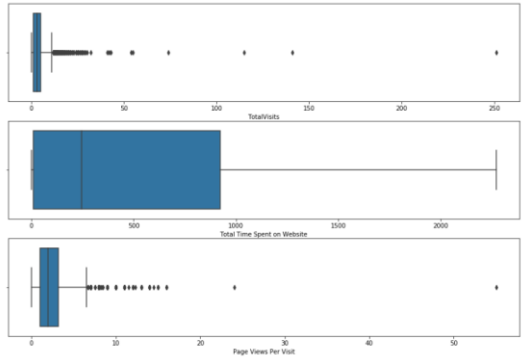



```
In [31]: # For 'Specialization'
percent_plot('Specialization')
```



There are a lot of different values of specialization and imputation as mode is not an ideal choice for such high frequency column. However, it is still possible that a person does not have a specialization or not listed in the provision.

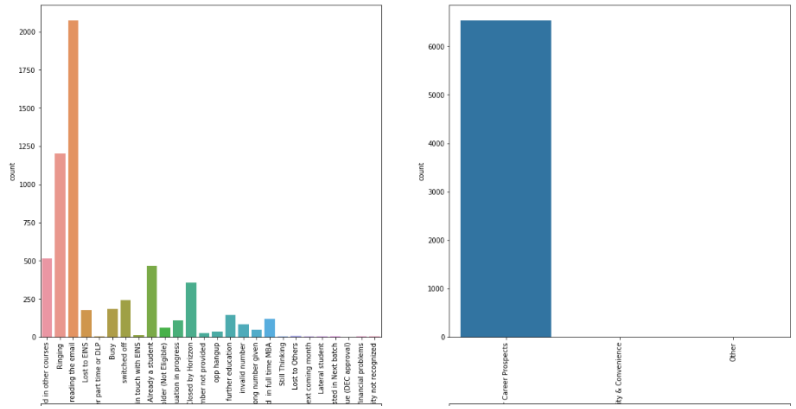
```
# Boxplots
num_var = ['TotalVisits','Total Time Spent on Website','Page Views Per Visit']
plt.figure(figsize = (15, 10))
for var in num_var:
    plt.subplot(3,1,num_var.index(var)+1)
    sns.boxplot(df_leads[var])
plt.show()
```




```
# Create a 2x2 grid of subplots
fig, axes = plt.subplots(2, 2, figsize=(20, 20))

# Create percent plots in each subplot
sns.countplot(data = df_leads, x = 'Tags', ax=axes[0, 0]).set_xticklabels(sns.countplot(data = df_leads, x = 'Tags', ax=axes[0,
sns.countplot(data = df_leads, x = 'What matters most to you in choosing a course', ax=axes[0, 1]).set_xticklabels(sns.countplot(
sns.countplot(data = df_leads, x = 'What is your current occupation', ax=axes[1, 0]).set_xticklabels(sns.countplot(data = df_lea
sns.countplot(data = df_leads, x = 'Country', ax=axes[1, 1]).set_xticklabels(sns.countplot(data = df_leads, x = 'Country', ax=axe

plt.show()
```



Data Visualization

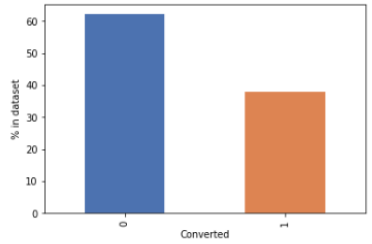
Data Visualization

This step includes the following actions:

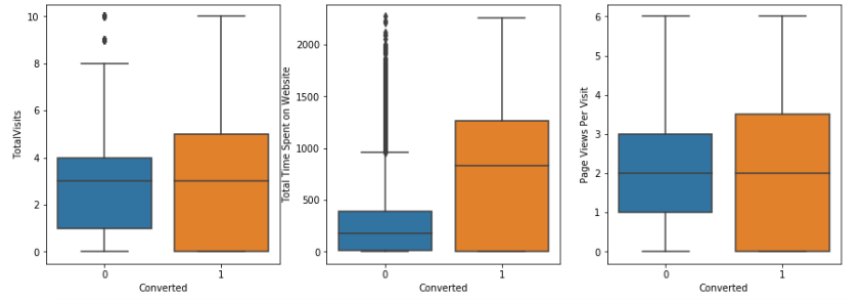
- Visualize the target variables for high-level sense of data distributions.
- Investigate the target variables with different other categorical/numerical variables with some comprehensive implication such as people spending more time on the website are more likely to be converted with also highest correlation coefficient.

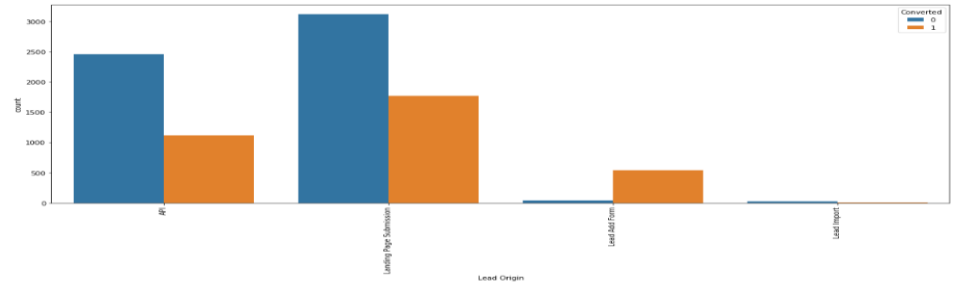
```
percent_plot('Converted') # Barplot for the target variable 'Converted'
print((sum(df_leads['Converted'])/len(df_leads['Converted'].index))*100) # check for current average conversion rate (e.g., 38%)
```

```
37.85541106458012
```




```
# Plot numerical variables against target variable to look for further relations
plt.figure(figsize=(15, 5))
for var in num_var:
    plt.subplot(1, 3, num_var.index(var)+1)
    sns.boxplot(y = var , x = 'Converted', data = df_leads)
plt.show()
```





Observations for **Lead Origin** :
'API' and 'Landing Page Submission' generate the most leads but have less conversion rates of around 30%. Whereas, 'Lead Add Form' generates less leads but conversion rate is great. We should **try to increase conversion rate for 'API' and 'Landing Page Submission', and increase leads generation using 'Lead Add Form'** while 'Lead Import' is very immaterial.

```
#Checking correlations of numeric values using heatmap
plt.figure(figsize = (10,8))
sns.heatmap(df_leads.corr(), cmap="YlGnBu", annot=True)
plt.show()
```

Model Evaluation

Model Evaluation – Final Model with all significant

Model 4

```
col3 = col2.drop('Tags_wrong number given', 1)

X4, logm4 = build_model(X_train[col3], y_train)
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                Converted   No. Observations:                 6351
Model:                            GLM   Df Residuals:                     6338
Model Family:                Binomial   Df Model:                           12
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1601.0
Date:                Tue, 16 Jan 2024   Deviance:                       3202.0
Time:                        22:44:53   Pearson chi2:                 3.48e+04
No. Iterations:                     8
Covariance Type:            nonrobust
==============================================================================================
                                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                                         -1.9192      0.211     -9.080      0.000      -2.333      -1.505
Do Not Email                                  -1.2835      0.212     -6.062      0.000      -1.698      -0.868
Lead Origin_Lead Add Form                      1.2035      0.368      3.267      0.001       0.482       1.925
Lead Source_Welingak Website                   3.2825      0.820      4.002      0.000       1.675       4.890
Tags_Busy                                      3.8043      0.330     11.525      0.000       3.157       4.451
Tags_Closed by Horizzon                        7.9789      0.762     10.467      0.000       6.485       9.473
Tags_Lost to EINS                              9.1948      0.753     12.209      0.000       7.719      10.671
Tags_Ringing                                  -1.8121      0.336     -5.401      0.000      -2.470      -1.154
Tags_Will revert after reading the email       3.9906      0.228     17.508      0.000       3.544       4.437
Tags_switched off                             -2.4456      0.586     -4.171      0.000      -3.595      -1.297
Lead Quality_Not Sure                         -3.5218      0.126    -28.036      0.000      -3.768      -3.276
Lead Quality_Worst                            -3.9106      0.856     -4.567      0.000      -5.589      -2.232
Last Notable Activity_SMS Sent                 2.7395      0.120     22.907      0.000       2.505       2.974
==============================================================================================
```

All of the features have p-value close to zero i.e. they all seem significant.

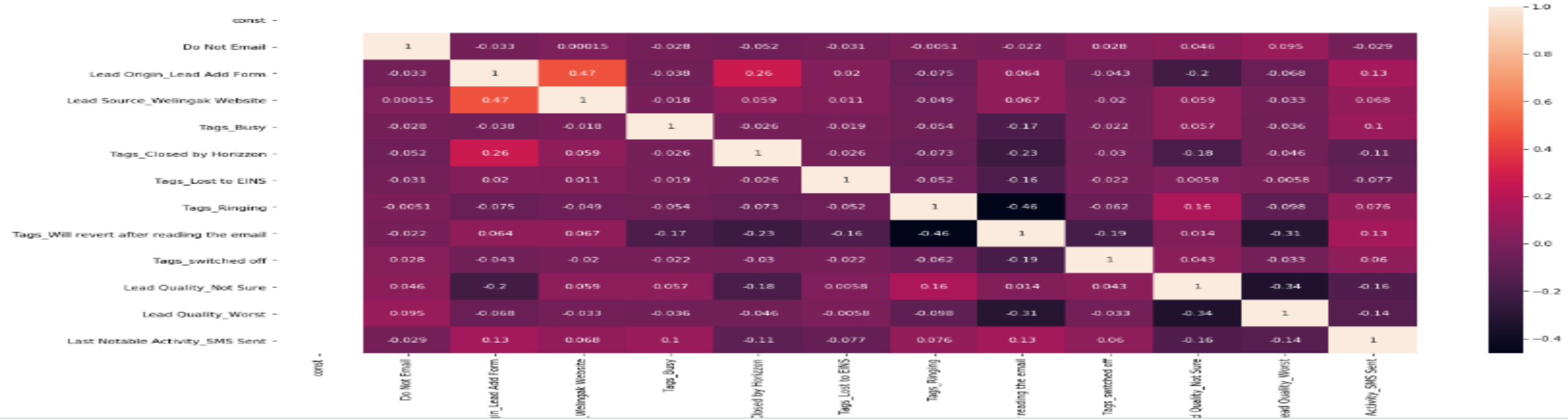
Model Evaluation – VIF and Heatmap Correlation

```
check_VIF(X4)
```

	Features	VIF
9	Lead Quality_Not Sure	2.62
7	Tags_Will revert after reading the email	2.57
1	Lead Origin_Lead Add Form	1.58
6	Tags_Ringing	1.52
11	Last Notable Activity_SMS Sent	1.51
2	Lead Source_Welingak Website	1.34
4	Tags_Closed by Horizzon	1.13
0	Do Not Email	1.10
3	Tags_Busy	1.10
8	Tags_switched off	1.10
5	Tags_Lost to EINS	1.04
10	Lead Quality_Worst	1.03

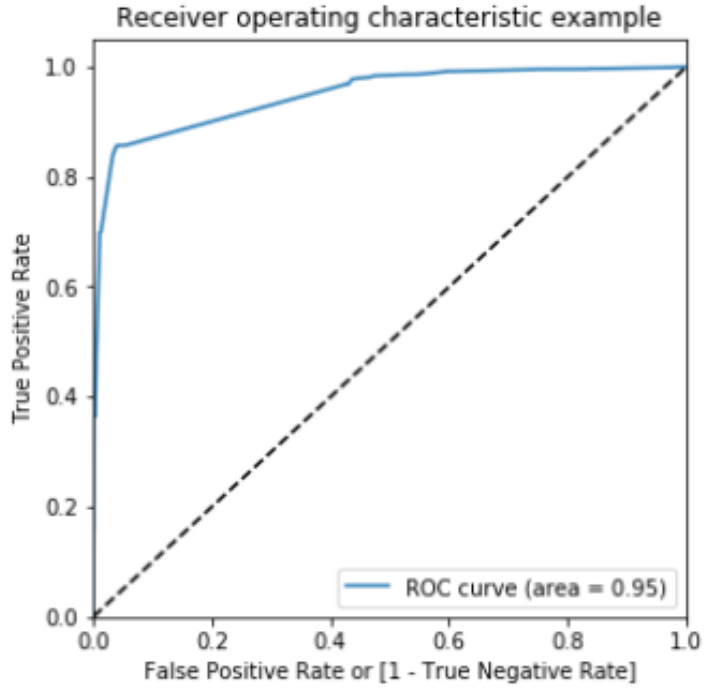
```
# To plot correlations
plt.figure(figsize = (20,10))
sns.heatmap(X4.corr(),annot = True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2422b6dd988>
```



Model Evaluation – ROC, Optimal Threshold, and Classification Report

```
# To plot ROC
plot_roc(y_train_pred_final.Converted, y_train_pred_final.Converted_prob)
```



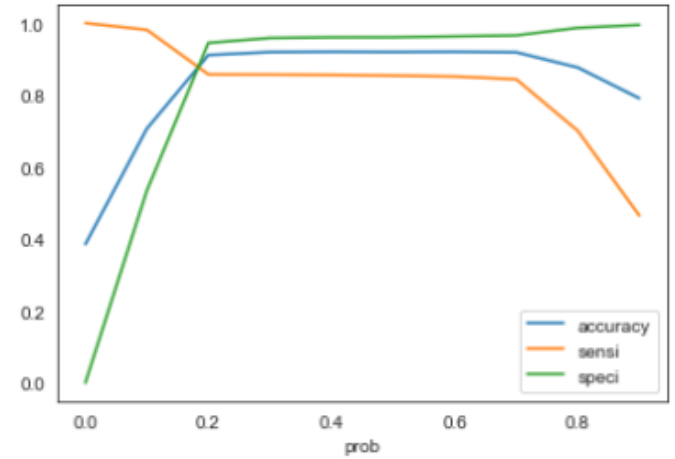
Classification Report

```
from sklearn.metrics import classification_report

print(classification_report(y_train_pred_final.Converted, y_train_pred_final.final_predicted))
```

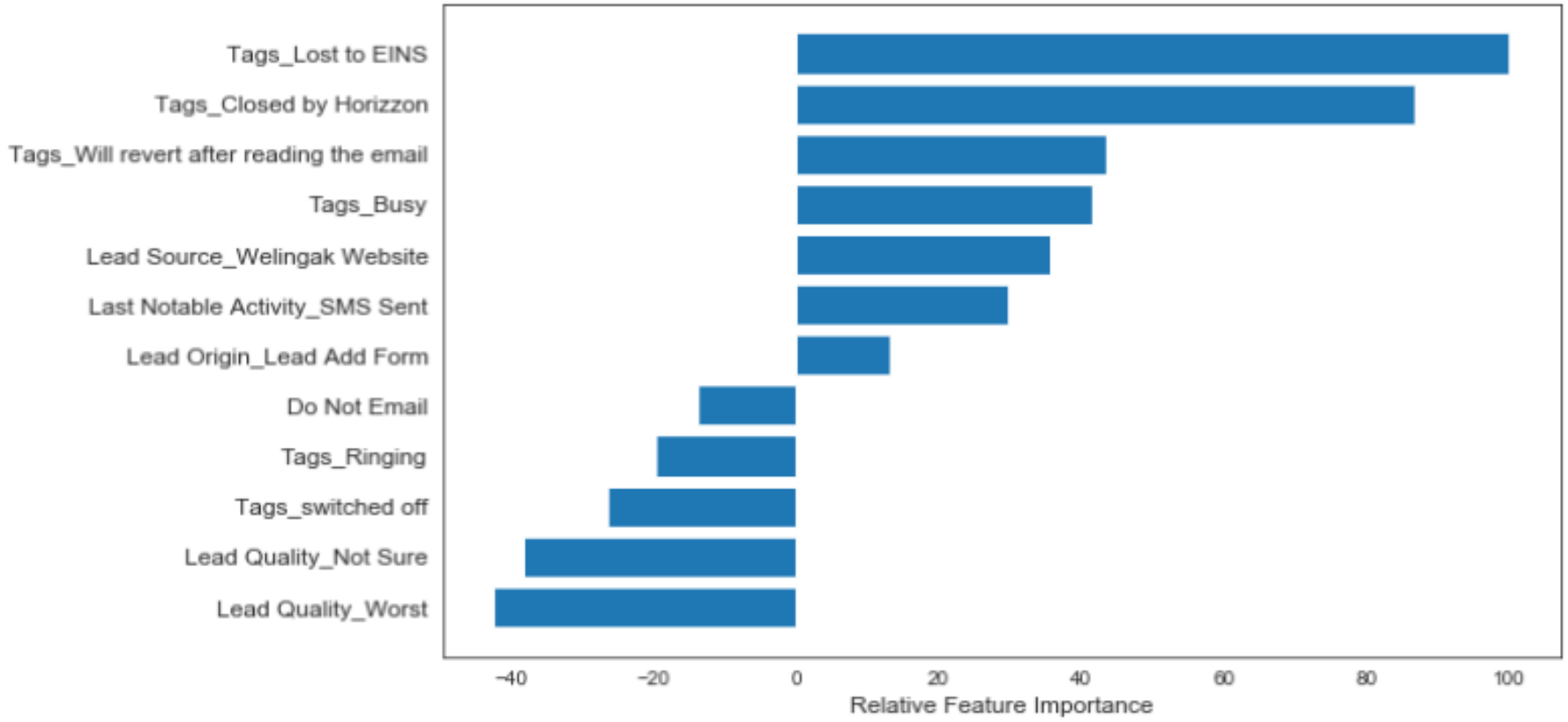
	precision	recall	f1-score	support
0	0.91	0.96	0.94	3905
1	0.93	0.86	0.89	2446
accuracy			0.92	6351
macro avg	0.92	0.91	0.91	6351
weighted avg	0.92	0.92	0.92	6351

```
# To plot accuracy, sensitivity and specificity for various probabilities
sns.set_style('white')
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```



Model Evaluation – Feature Importance

```
fig = plt.figure(figsize=(10,6))
ax = fig.add_subplot(1, 1, 1)
pos = np.arange(sorted_idx.shape[0])
ax.barh(pos, feature_importance[sorted_idx])
ax.set_yticks(pos)
ax.set_yticklabels(np.array(X_train[col3].columns)[sorted_idx], fontsize=12)
ax.set_xlabel('Relative Feature Importance', fontsize=12)
plt.show()
```



Conclusions

Conclusions

Three variables which contribute most towards the probability of a lead conversion in decreasing order of impact are:

- Tags_Lost to EINS
- Tags_Closed by Horizzon
- Tags_Will revert after reading the email

These are dummy features created from the categorical variable and contribute positively towards the probability of a lead conversion.

These results indicate that the company should focus more on the leads with these three tags.

Lastly leads with high total time spent on website might be a valuable source for increasing lead scoring.

Thank you for viewing

Q&A