LEAD SCORE CASE STUDY

(MARKET PLACE BUSINESS MODEL)

Shubhang Sharma

Vishnu Sai Tej Nagabandi

Shubham Mojidar

PROBLEM STATEMENT

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%.

There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

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 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%.

coursera udemy

GOAL OF THE 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.

EXTERNAL SEARCH

The dataset used can be download from here.

The dataset consists of the following variables: -

Variables	Description
Prospect ID	A unique ID with which the customer is identified.
Lead Number	A lead number assigned to each lead procured.
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Do Not Email	An indicator variable selected by the customer wherein they select whether or not they want to be emailed about the course or not.
Do Not Call	An indicator variable selected by the customer wherein they select whether or not they want to be called about the course or not.
Converted	The target variable. Indicates whether a lead has been successfully converted or not.
TotalVisits	The total number of visits made by the customer on the website.
Total Time Spent on Website	The total time spent by the customer on the website.
Page Views Per Visit	Average number of pages on the website viewed during the visits.
Last Activity	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
Country	The country of the customer.
Specialization	The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form.
How did you hear about X Education	The source from which the customer heard about X Education.
What is your current occupation	Indicates whether the customer is a student, unemployed or employed.

What matters most to you in choosing this course	An option selected by the customer indicating what is their main motto behind doing this course.
Search	Indicating whether the customer had seen the ad in any of the listed items.
Magazine	
Newspaper Article	
X Education Forums	
Newspaper	
Digital Advertisement	
Through Recommendations	Indicates whether the customer came in through recommendations.
Receive More Updates About Our Courses	Indicates whether the customer chose to receive more updates about the courses.
Tags	Tags assigned to customers indicating the current status of the lead.
Lead Quality	Indicates the quality of lead based on the data and intuition the employee who has been assigned to the lead.
Update me on Supply Chain Content	Indicates whether the customer wants updates on the Supply Chain Content.
Get updates on DM Content	Indicates whether the customer wants updates on the DM Content.
Lead Profile	A lead level assigned to each customer based on their profile.
City	The city of the customer.
Asymmetrique Activity Index	An index and score assigned to each customer based on their activity and their profile
Asymmetrique Profile Index	
Asymmetrique Activity Score	
Asymmetrique Profile Score	
I agree to pay the amount through cheque	Indicates whether the customer has agreed to pay the amount through cheque or not.
a free copy of Mastering The Interview	Indicates whether the customer wants a free copy of 'Mastering the Interview' or not.
Last Notable Activity	The last notable activity performed by the student.

APPLICABLE REGULATIONS

Some of the applicable regulations for any education company providing education through an online platform like Udemy are as follows:

1. **Intellectual property laws:** Udemy is required to comply with copyright laws and ensure that all course materials uploaded to its platform do not infringe on the intellectual property rights of others.

2. **Data privacy laws:** Udemy is required to comply with data protection regulations such as the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the United States. This includes obtaining consent from users before collecting their personal data and implementing adequate security measures to protect their data.

3. **Consumer protection laws:** Udemy is subject to consumer protection regulations that require it to provide accurate and transparent information about its courses and services, offer fair pricing, and provide refunds or exchanges where necessary.

4. **Tax regulations:** Udemy is required to comply with tax regulations in the countries where it operates. This includes registering for and remitting taxes on behalf of instructors who earn income from the platform.

5. **Employment laws:** Udemy is required to comply with employment regulations in the countries where it has employees. This includes providing fair wages, benefits, and working conditions to its employees.

APPLICABLE CONSTRAINTS

As an online learning platform, Udemy is subject to various constraints that can affect its operations and growth. Some of the applicable constraints for Udemy are:

1. **Competition:** Udemy operates in a highly competitive market, with numerous other online learning platforms offering similar courses and services. This can constrain its

growth and profitability, as it needs to continually innovate and improve its offerings to stay ahead of the competition.

2. **Regulatory constraints:** As mentioned in the previous answer, Udemy is subject to various regulations that can constrain its operations and require it to invest in compliance measures.

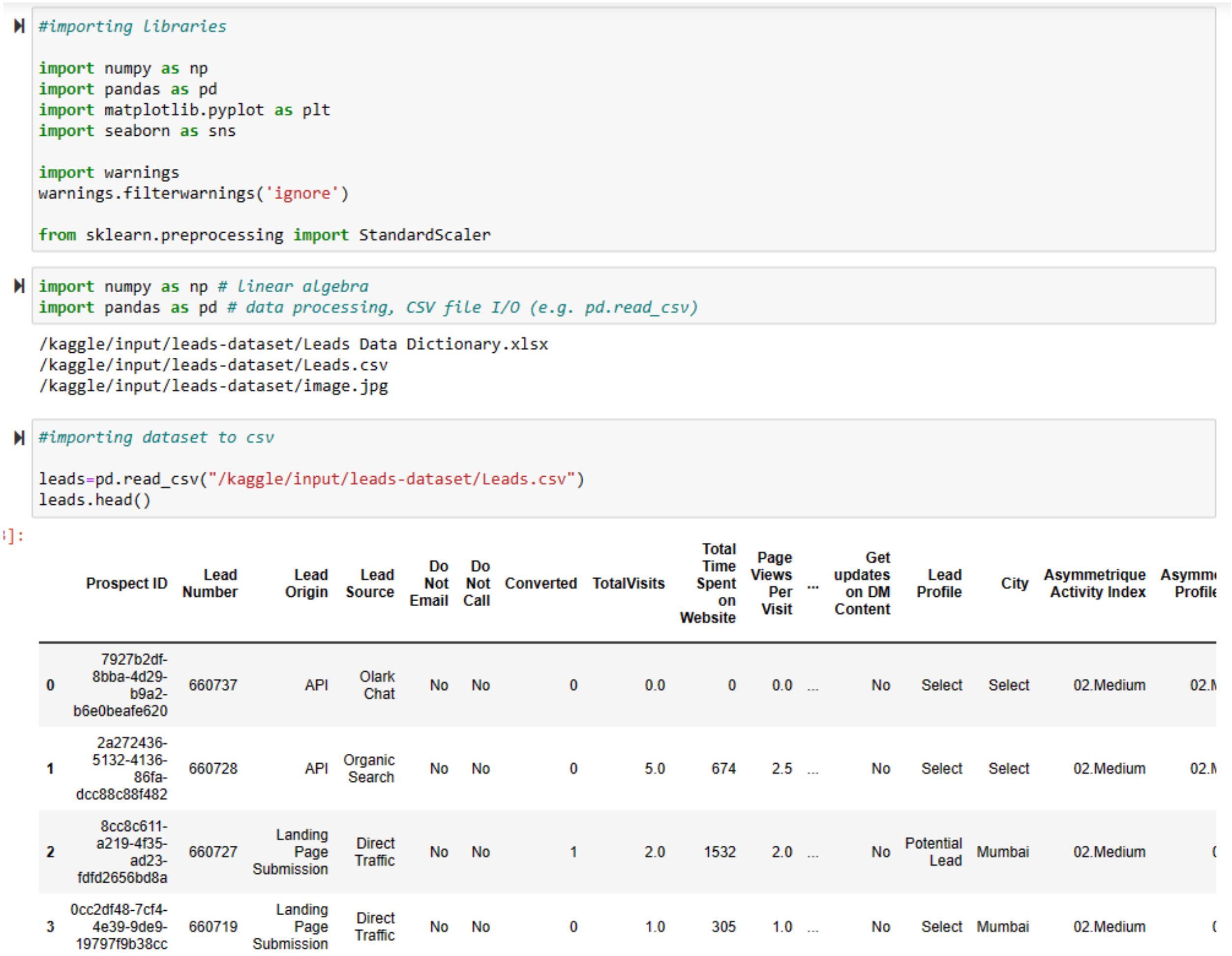
3. **Technical constraints:** Udemy relies on technology to deliver its courses and services, and technical issues such as downtime or slow loading times can affect user satisfaction and retention. Additionally, the platform may require significant investment in technology infrastructure and security measures to ensure that it operates smoothly and securely.

4. **Instructor quality and availability:** The quality and availability of instructors can also constrain Udemy's operations and growth. Udemy relies on its instructors to create and deliver high-quality courses, and a shortage of quality instructors or a high turnover rate can limit the number of courses available on the platform.

5. **User retention and engagement:** Finally, Udemy's success is also dependent on its ability to retain and engage users. High user churn rates or low engagement levels can constrain its growth and profitability, as it may struggle to attract new users and keep existing ones.

CONCEPT GENERATION

1. READING AND UNDERSTANDING THE DATA



```
#importing libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')

from sklearn.preprocessing import StandardScaler
```

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
```

```
/kaggle/input/leads-dataset/Leads Data Dictionary.xlsx
/kaggle/input/leads-dataset/Leads.csv
/kaggle/input/leads-dataset/image.jpg
```

```
#importing dataset to csv

leads=pd.read_csv("/kaggle/input/leads-dataset/Leads.csv")
leads.head()
```

]:

	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	Asymm Profile
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	Select	Select	02.Medium	02.M
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	...	No	Select	Select	02.Medium	02.M
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	(
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	(

2. DATA QUALITY CHECK

i. Checking for the null/missing values

There were no null/missing values in the dataset

ii. Duplicate check

The shape after running the drop duplicate command is same as the original dataframe. Hence, we can conclude that there were zero duplicate values in the dataset.

iii. Data Cleaning

There seems to be no Junk/Unknown values in the entire dataset.

iv. Removing redundant and unwanted columns

Based on the high-level look at the data and the data dictionary, the following variables can be removed from further analysis:

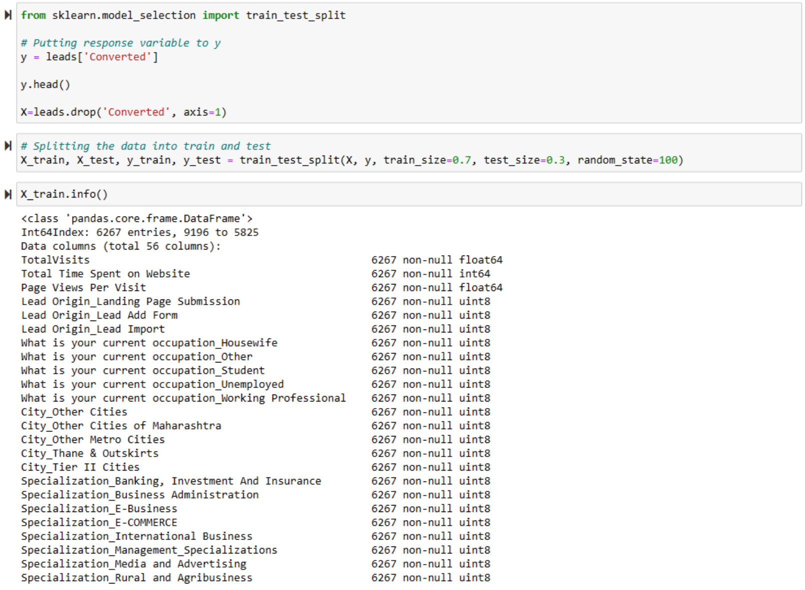
- **instant:** Its only an index value
- dteday: This has the date, Since we already have I columns for 'year' &
- 'month',hence, we could live without this column.
- casual & registered: Both these columns contain the count of bike
- booked by different categories of customers. Since our objective is to find the total count of bikes and not by specific category, we will ignore these two columns. Moreover, we have created a new variable to have the ratio of these customer types.
- We will save the new dataframe as bike_new, so that the original dataset is preserved for any future analysis/validation.

v. Creating dummy variables

We will create DUMMY variables for 4 categorical variables 'mnth', 'weekday', 'season' & 'weathersit'.

- Before creating dummy variables, we will have to convert them into 'category' data types.

3. SPLITTING OF DATA



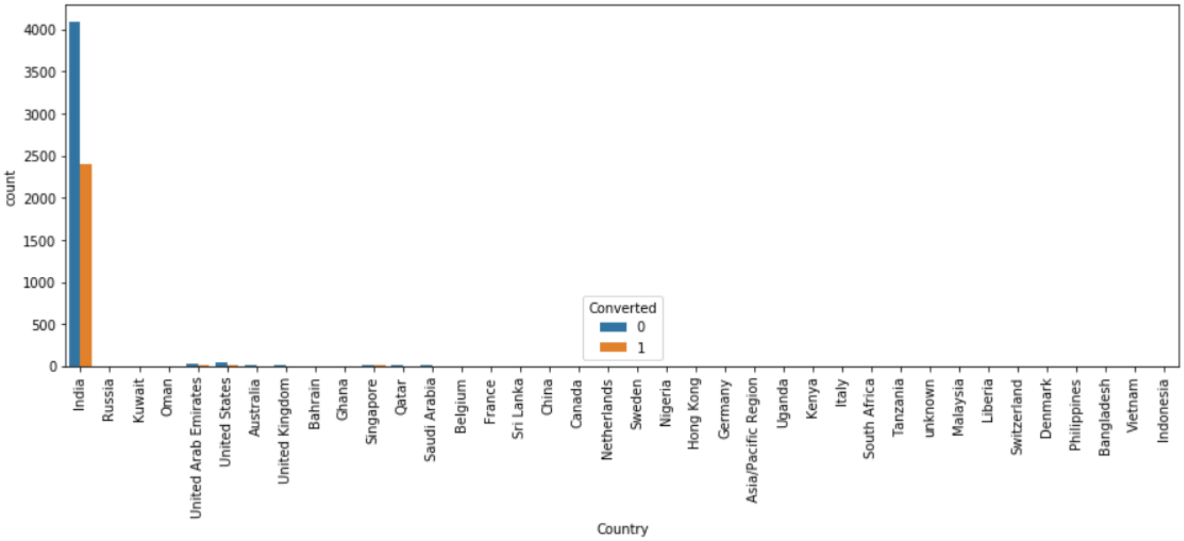
```
from sklearn.model_selection import train_test_split

# Putting response variable to y
y = leads['Converted']

y.head()

X=leads.drop('Converted', axis=1)
```

```
# Splitting the data into train and test
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=100)
```

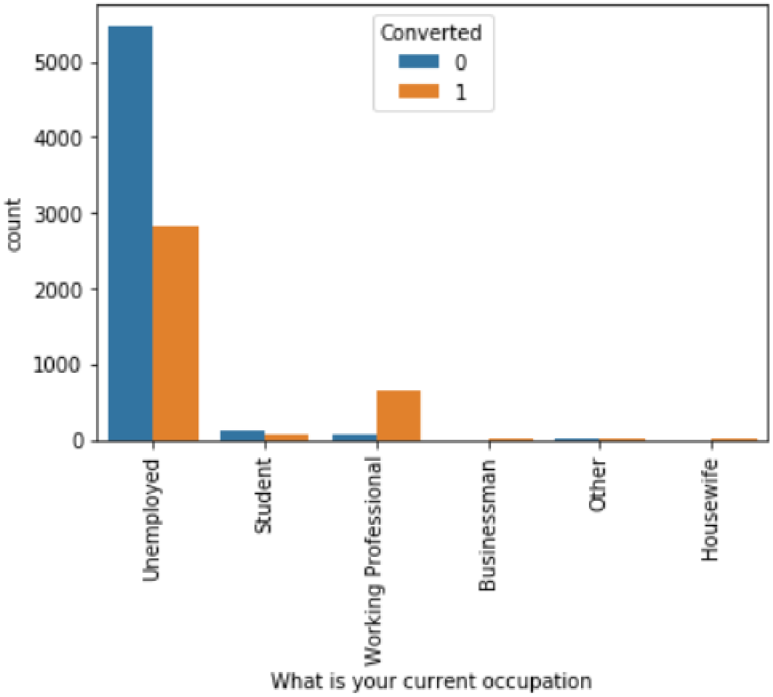
```
X_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6267 entries, 9196 to 5825
Data columns (total 56 columns):
TotalVisits                                         6267 non-null float64
Total Time Spent on Website                         6267 non-null int64
Page Views Per Visit                                6267 non-null float64
Lead Origin_Landing Page Submission                 6267 non-null uint8
Lead Origin_Lead Add Form                           6267 non-null uint8
Lead Origin_Lead Import                             6267 non-null uint8
What is your current occupation_Housewife           6267 non-null uint8
What is your current occupation_Other               6267 non-null uint8
What is your current occupation_Student             6267 non-null uint8
What is your current occupation_Unemployed          6267 non-null uint8
What is your current occupation_Working Professional    6267 non-null uint8
City_Other Cities                                   6267 non-null uint8
City_Other Cities of Maharashtra                    6267 non-null uint8
City_Other Metro Cities                             6267 non-null uint8
City_Thane & Outskirts                              6267 non-null uint8
City_Tier II Cities                                 6267 non-null uint8
Specialization_Banking, Investment And Insurance    6267 non-null uint8
Specialization_Business Administration              6267 non-null uint8
Specialization_E-Business                           6267 non-null uint8
Specialization_E-COMMERCE                           6267 non-null uint8
Specialization_International Business               6267 non-null uint8
Specialization_Management_Specializations           6267 non-null uint8
Specialization_Media and Advertising                6267 non-null uint8
Specialization_Rural and Agribusiness               6267 non-null uint8
```

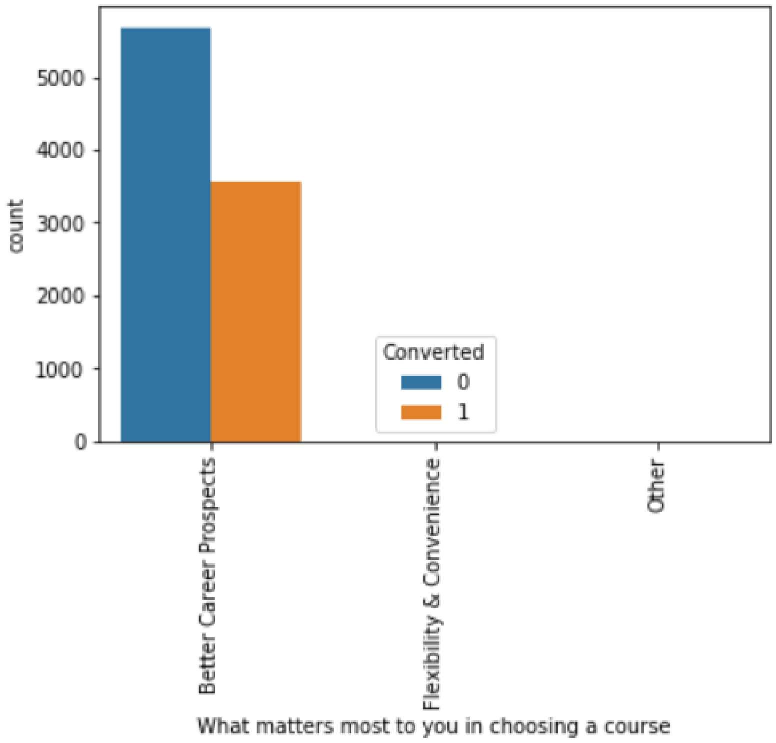
4. EXPLORATORY DATA ANALYSIS

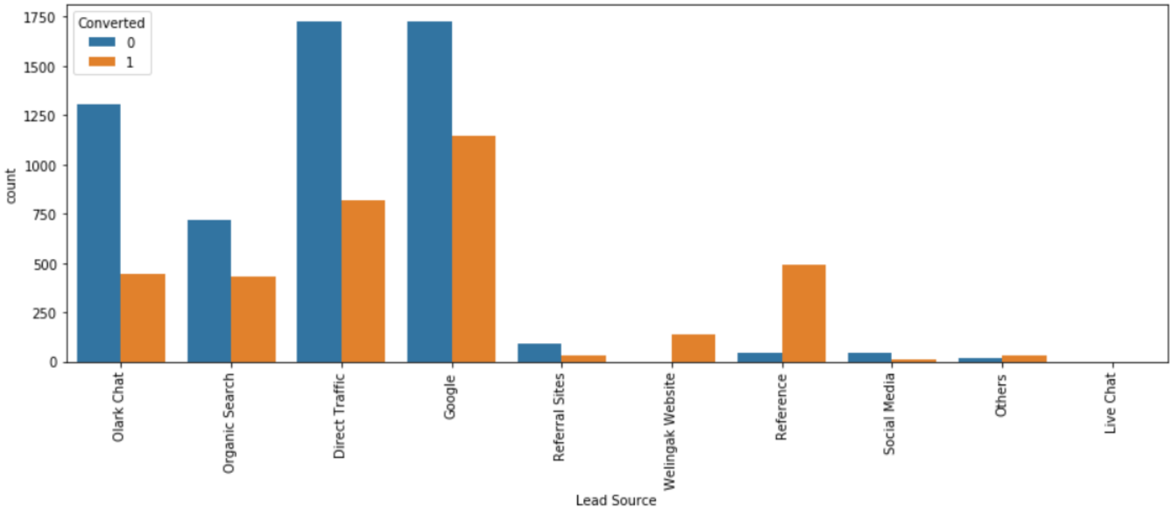


- As we can see the number if values for India are quite high(about 97% of the data), this column can be dropped.

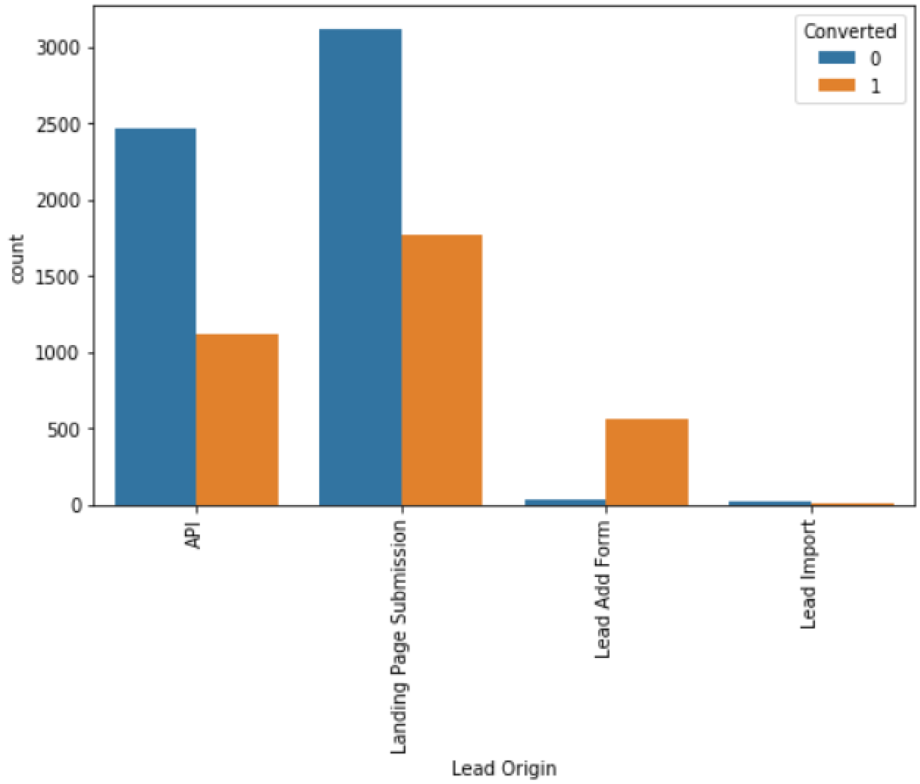


- Working Professionals going for the course have high chances of joining it.
- Unemployed leads are the most in terms of Absolute numbers.

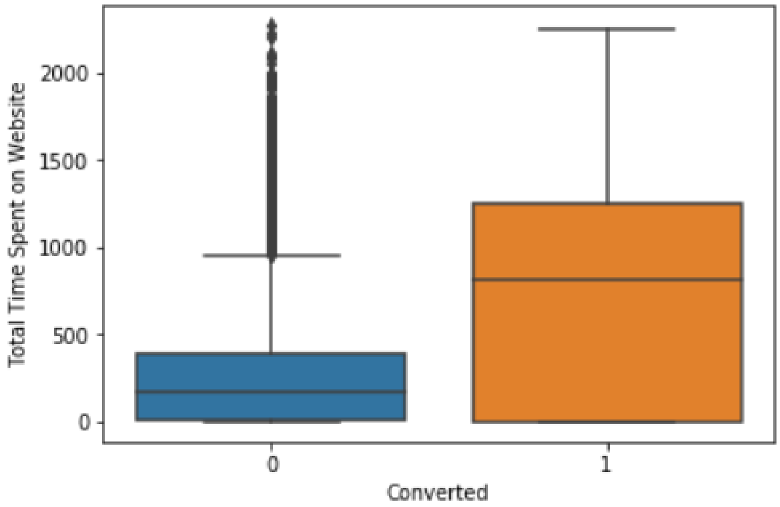




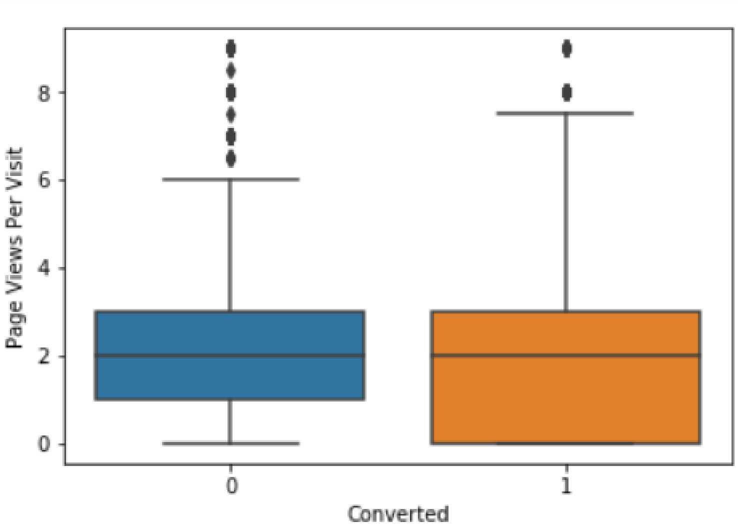
- Maximum number of leads are generated by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead conversion of Olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.



- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form get very few leads.
- In order to improve overall lead conversion rate, we have to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

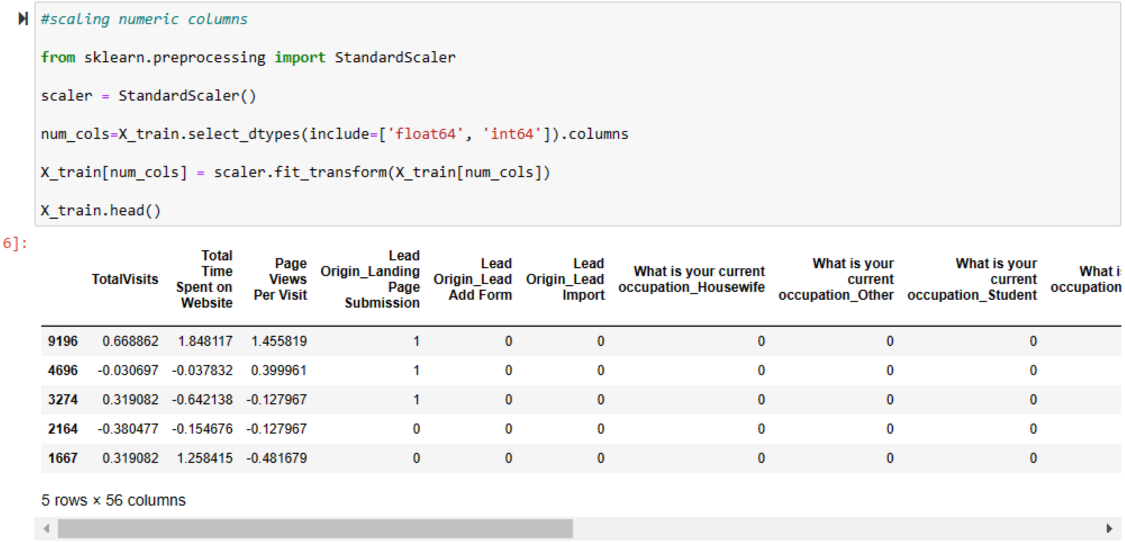


- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.



- Median for converted and unconverted leads is the same.
- Nothing can be said specifically for lead conversion from Page Views Per Visit

5. SCALING OF THE FEATURES



```
#scaling numeric columns

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

num_cols=X_train.select_dtypes(include=['float64', 'int64']).columns

X_train[num_cols] = scaler.fit_transform(X_train[num_cols])

X_train.head()
```

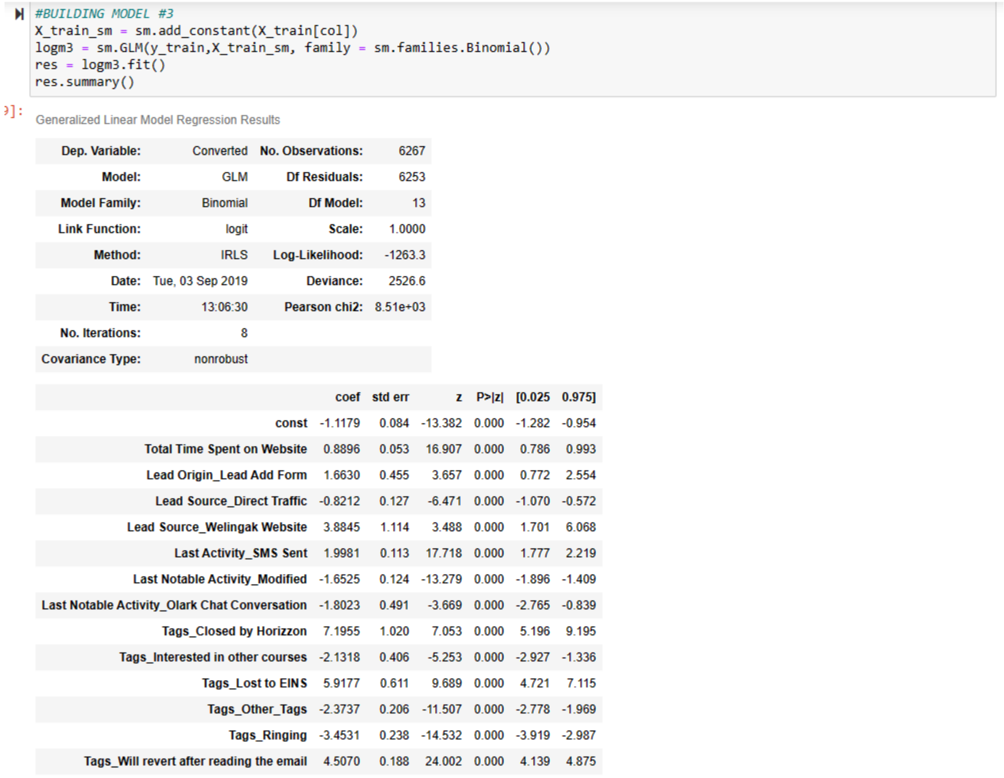
6]:

	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	What is your current occupation_Housewife	What is your current occupation_Other	What is your current occupation_Student	What i occupation
9196	0.668862	1.848117	1.455819	1	0	0	0	0	0	
4696	-0.030697	-0.037832	0.399961	1	0	0	0	0	0	
3274	0.319082	-0.642138	-0.127967	1	0	0	0	0	0	
2164	-0.380477	-0.154676	-0.127967	0	0	0	0	0	0	
1667	0.319082	1.258415	-0.481679	0	0	0	0	0	0	

5 rows × 56 columns

6. BUILDING A MODEL USING 'STATS MODEL'

Various Stats Model were trained to get the best fit model. In which 6 the model fit the best over the dataset provided. In these models, VIF values were checked so as to neglect the high valued predictors causing high multicollinearity.



```
#BUILDING MODEL #3
X_train_sm = sm.add_constant(X_train[col])
logm3 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm3.fit()
res.summary()
```

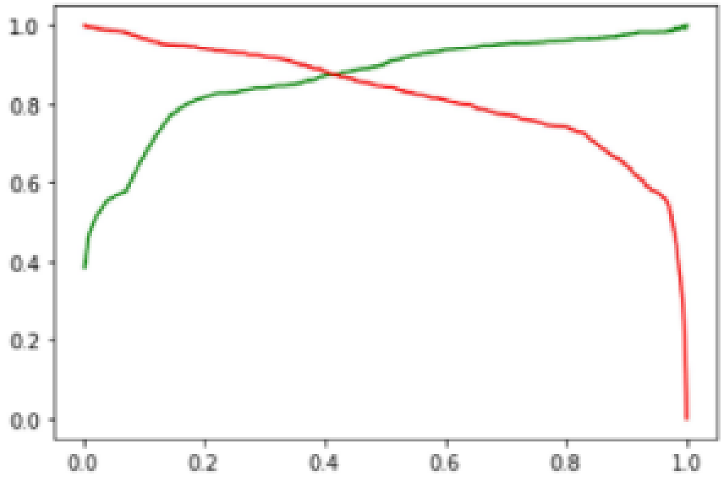
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6267
Model:	GLM	**Df Residuals:**	6253
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1263.3
Date:	Tue, 03 Sep 2019	**Deviance:**	2526.6
Time:	13:06:30	**Pearson chi2:**	8.51e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1179	0.084	-13.382	0.000	-1.282	-0.954
Total Time Spent on Website	0.8896	0.053	16.907	0.000	0.786	0.993
Lead Origin_Lead Add Form	1.6630	0.455	3.657	0.000	0.772	2.554
Lead Source_Direct Traffic	-0.8212	0.127	-6.471	0.000	-1.070	-0.572
Lead Source_Welingak Website	3.8845	1.114	3.488	0.000	1.701	6.068
Last Activity_SMS Sent	1.9981	0.113	17.718	0.000	1.777	2.219
Last Notable Activity_Modified	-1.6525	0.124	-13.279	0.000	-1.896	-1.409
Last Notable Activity_Olark Chat Conversation	-1.8023	0.491	-3.669	0.000	-2.765	-0.839
Tags_Closed by Horizzon	7.1955	1.020	7.053	0.000	5.196	9.195
Tags_Interested in other courses	-2.1318	0.406	-5.253	0.000	-2.927	-1.336
Tags_Lost to EINS	5.9177	0.611	9.689	0.000	4.721	7.115
Tags_Other_Tags	-2.3737	0.206	-11.507	0.000	-2.778	-1.969
Tags_Ringing	-3.4531	0.238	-14.532	0.000	-3.919	-2.987
Tags_Will revert after reading the email	4.5070	0.188	24.002	0.000	4.139	4.875

This model 3 looks good, as there seems to be VERY LOW Multicollinearity between the predictors and the p-values for all the predictors seems to be significant. For now, we will consider this as our final model (unless the Test data metrics are not significantly close to this number).

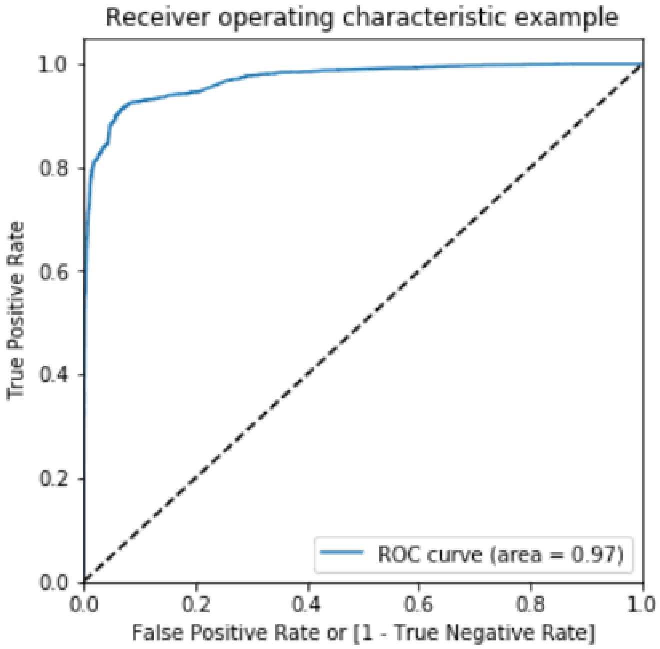
7. MODEL EVALUATION



The graph depicts an optimal cut off 0.42 based on Precision and Recall.

- Precision = 84.12%
- Recall = 92.05%

8. MAKING PREDICTIONS



- The ROC Curve should be a value close to 1. We are getting a good value of 0.97 indicating a good predictive model.

Observation:

1. So, as we can see above the model seems to be performing well. The ROC curve has a value of 0.97, which is very good. We have the following values for the Train Data:

   - Accuracy: 92.29%
   - Sensitivity: 91.70%
   - Specificity: 92.66%

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

   - Accuracy: 92.78%
   - Sensitivity: 91.98%
   - Specificity: 93.26%

CONCEPT DEVELOPMENT

Concept development for Udemy would involve identifying opportunities to improve and expand the platform’s offerings and user experience. Here are a few ideas:

1. **Personalized learning paths:** Udemy could develop algorithms that analyze user data and provide personalized learning paths based on each user’s interests, learning style, and progress. This could increase user engagement and satisfaction by providing a more tailored and efficient learning experience.

2. **Collaboration tools:** Udemy could develop tools that enable users to collaborate on projects and assignments, either with other users or with instructors. This could foster a sense of community and collaboration, and provide users with opportunities to practice and apply their new skills.

3. **Integration with other learning tools:** Udemy could explore partnerships or integrations with other learning tools, such as language learning apps, coding platforms, or project management software. This could expand Udemy’s reach and provide users with a more comprehensive learning experience.

4. **Certification programs:** Udemy could develop certification programs that enable users to demonstrate their mastery of specific skills or topics. This could increase the value of Udemy courses for both users and employers, and provide a new revenue stream for the platform.

5. **Corporate training solutions:** Udemy could develop customized training solutions for corporate clients, either by partnering with employers or by developing a separate platform for enterprise users. This could provide a new market for Udemy’s courses and services, and enable employers to provide their employees with relevant and up-to-date training.

CODE IMPLEMENTATION

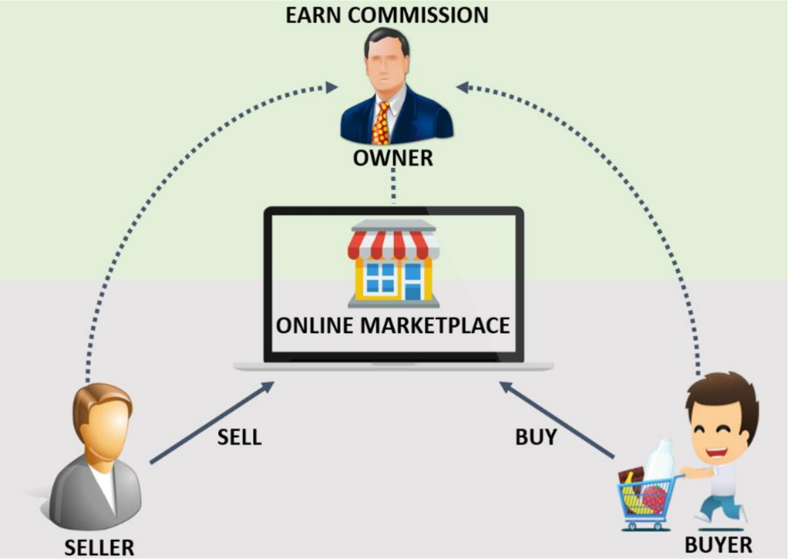
The code of the model proposed can be found here.

CONCLUSION

1. While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
2. Accuracy, Sensitivity and Specificity values of test set are around 91%, 91.41% and 90.62% which are approximately closer to the respective values calculated using trained set.
3. Lead score calculated shows the conversion rate on the final predicted model is around 92.05% (in train set) and 91.41% in test set
4. The top variables that contribute for lead getting converted in the model are
   - Total time spent on website
   - What is your current occupation
   - Lead Add Form from Lead Origin
   - Had a Phone Conversation from Last Notable Activity
5. Hence overall this model seems to be good.

BUSINESS MODEL

The marketplace business model has revolutionized the way we shop and conduct business in the digital age. In this model, a digital platform serves as a middleman connecting buyers and sellers, allowing them to transact with each other and exchange goods and services. This model has disrupted traditional retail and commerce by providing a more efficient and convenient way for consumers to access products and services from a variety of sellers in one place. The success of companies like Amazon, eBay, and Etsy has proven the viability of this model, and today, marketplaces exist in nearly every industry, from transportation to real estate to freelance services. This model has also created new opportunities for entrepreneurs and small businesses to reach a wider audience, and has fostered a culture of innovation and competition in the digital economy.



FEASABILITY:

- Udemy is an online learning platform that allows users to access a wide range of courses on various topics. The platform has been around for over a decade and has grown significantly over the years. In terms of feasibility, Udemy is a very feasible platform for both learners and instructors.
- **For learners,** Udemy provides a convenient way to access high-quality courses on various topics from anywhere in the world. The platform offers courses in a variety of formats, including video lectures, quizzes, and assignments. Additionally, courses on Udemy are often affordable compared to traditional education options, making it accessible to a wide range of learners.
- **For instructors,** Udemy offers an excellent opportunity to monetize their expertise and share their knowledge with others. The platform provides tools and resources to help instructors create and publish high-quality courses. Instructors can also earn money by selling their courses on Udemy, with the platform taking a percentage of the revenue.

VIABILITY:

- Udemy has been a very successful online learning platform, having served over 50 million students worldwide and offering over 155,000 courses in various topics. The platform has proven to be viable in terms of its business model and financial success.
- One reason for Udemy’s viability is its revenue-sharing model, which allows instructors to earn a percentage of the revenue generated from their courses. This incentivizes instructors to create high-quality courses and attract students to their content, while also providing Udemy with a steady stream of revenue. Udemy takes a percentage of each course sale, but the revenue-sharing model ensures that both Udemy and its instructors benefit from the platform’s success.

- Additionally, Udemy has been successful in expanding its user base by offering courses in multiple languages and partnering with universities and companies to offer specialized training programs. The platform has also invested in improving its technology, such as its mobile app and personalized learning features, to provide a better learning experience for its users.

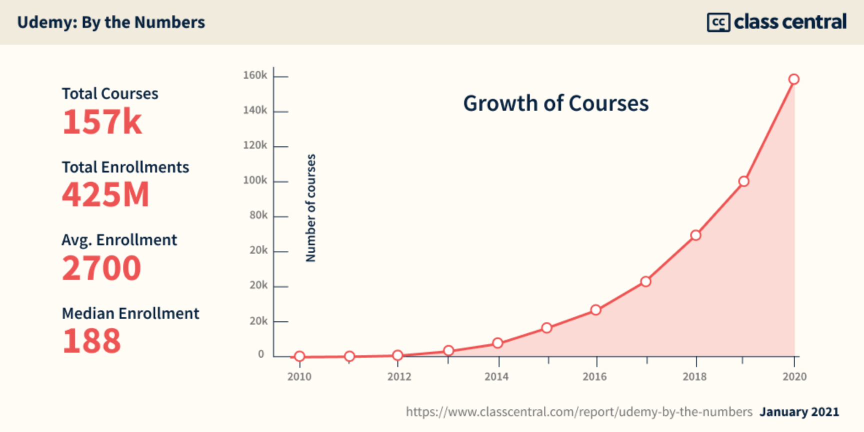
MONETIZATION:

- Udemy generates revenue through a variety of monetization strategies. The primary source of revenue for Udemy is through the sale of courses on its platform. When instructors publish their courses on Udemy, they have the option to price their courses and earn a percentage of the revenue generated from the sale of their courses. Udemy takes a percentage of the revenue generated from each course sale, typically around 50%, although this percentage may vary depending on the instructor's relationship with Udemy.
- In addition to course sales, Udemy also generates revenue through its Udemy for Business platform, which provides companies with access to its courses for employee training and development. Udemy for Business operates on a subscription-based model, with companies paying a monthly fee for access to its courses.
- Udemy also generates revenue through its affiliate program, which allows individuals or companies to earn a commission by promoting Udemy courses on their website or social media channels.

FINANCIAL EQUATION

The financial equation for Udemy is as follows:

Revenue = (Number of courses sold x Average course price) + (Number of Udemy for Business subscriptions x Monthly subscription fee) + (Affiliate marketing revenue) + (Advertising and sponsorships revenue)



https://www.classcentral.com/report/udemy-by-the-numbers **January 2021**