Lead Scoring
Case Study Presentation

Logistic Regression

- Surya Kurapati
- Jalpa Vataliya
- Siddalingappa kadakol

Core Objective:

- An education company named ‘X Education’ sells online courses to industry professionals through their website.
- People land on the website and either 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.
- Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- Require to build a model that assigns 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

Dataset:

- Dataset of leads from the past with around 9000 data points with the target variable column as ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Goal:

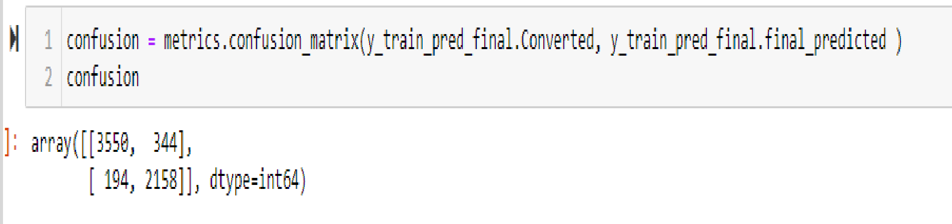
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean the lead is most likely to convert whereas a lower score would mean that the lead mostly not get converted.

Observations:

1. Given raw data set has 9240 rows and 37 columns with 17 columns having nulls in it.
2. Prospect ID and Lead Number are considered to be unique keys in the given dataset.
3. There are no exact duplicates found in above two columns and all 9240 records are found to be unique.
4. To clean the null data for further analysis and model building, firstly dropped the columns with null% greater than 45% and those are:
   [‘Lead Scoring Case Study’, ‘Asymmetrique Profile Score’, ‘Asymmetrique Activity Score’, ‘Asymmetrique Profile Index’, ‘Asymmetrique Activity Index’]
5. Now for the rest of the null columns, visually analyzed using univariate and bi-variate analysis and have come to the conclusion whether to remove or not.
6. Although [“Tags”] feature has got around 37% of null values, it has few values that has shown significant amount of lead conversion rate and felt the need of the variable for further modelling.
7. Majority (~75%) of the ['What matters most to you in choosing a course’] feature data belong to the value 'Better Career Prospects’ and ~20% to null data, hence dropped the column as there would be no intuition coming around it.
8. Working Professionals going for the course have high chances of joining the course, whereas unemployed leads are the most in terms of Absolute numbers.
9. Majority of the country data belong to 'India', hence dropped the column as there would be no intuition coming around it.
10. Specialization with Management in them have higher number of leads conversion rate, so this could be a significant variable and haven’t dropped.
11. Standardized and updated the values in [‘Lead Source’] feature w.r.t. google, facebook, and other category sources for easy modelling at later purpose. However, maximum number of leads are generated by Google and Direct traffic.
12. In addition, conversion rate of reference leads and leads through welingak website is high.
13. It is observed that customers with last activity as “SMS Sent” showcased higher lead conversion rate.
14. In order to improve overall lead conversion rate, it is observed to improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
15. After recursive feature elimination and building a final model, found that there are 761 leads who can be contacted and have high chance of getting converted as potential lead.

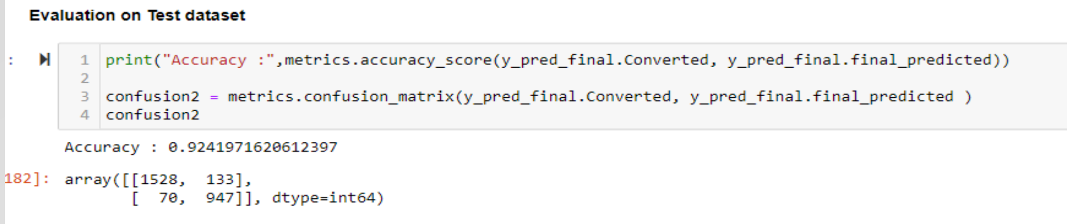
Metrics Evaluation of the final model:

1. Upon recursive feature elimination and final model building, below is the confusion matrix that helps in deriving following model evaluation metrics.



```
confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.final_predicted )
confusion
```

```
]: array([[3550,  344],
         [ 194, 2158]], dtype=int64)
```



Evaluation on Test dataset

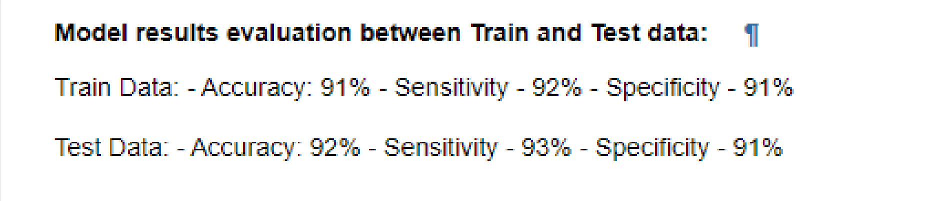
```
print("Accuracy :",metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted))

confusion2 = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.final_predicted )
confusion2
```

```
Accuracy : 0.9241971620612397

182]: array([[1528,  133],
             [  70,  947]], dtype=int64)
```

2. Below are the final model evaluation metrics derived for train and test dataset, explaining the accuracy, sensitivity where sensitivity coming out to be 92% mean, off total converted leads 92% of the data is correctly predicted as potential leads against train data set, this proves model confidence.

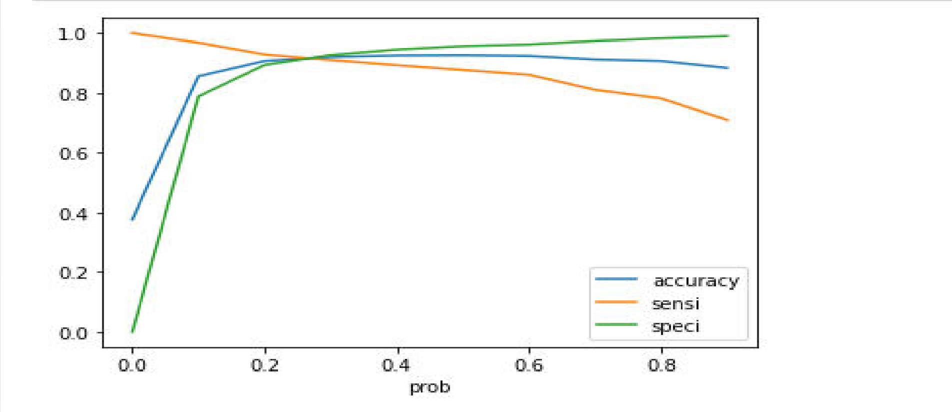


Model results evaluation between Train and Test data: ¶

Train Data: - Accuracy: 91% - Sensitivity - 92% - Specificity - 91%

Test Data: - Accuracy: 92% - Sensitivity - 93% - Specificity - 91%

3. Drew a graph between metrics such as accuracy, sensitivity and specificity and obtained an optimal point as 0.25 as cutoff score.



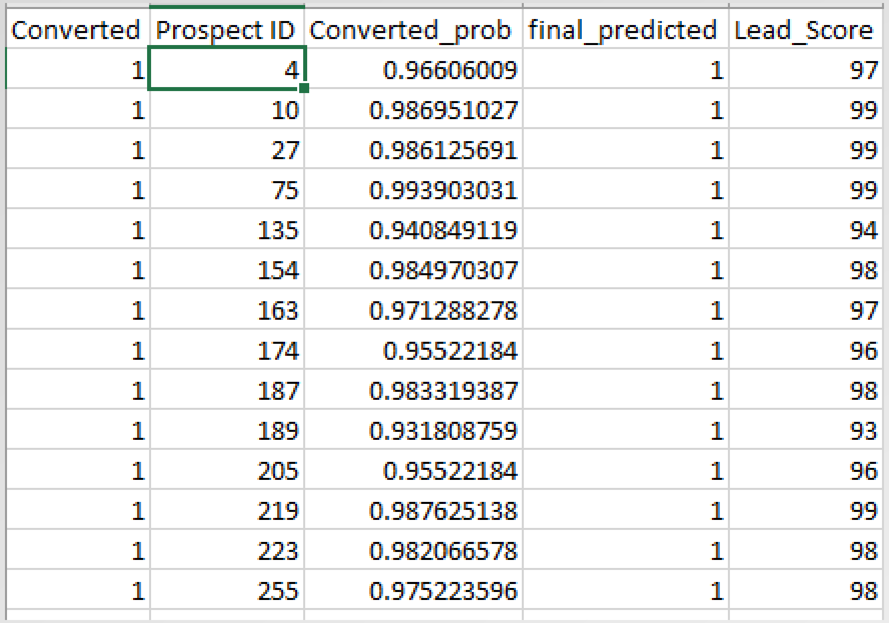
List of potential leads:

1. Attached below excel consists of all the 761 potential leads data

Microsoft Excel
ma Separated Valu

2. Sample data:

Converted: Whether customer is a potential lead or not to take the course, here 1 is potential lead and 0 is not a potential lead
Prospect ID: Index ID of every customer / Unique key
Converted_prob: Describes the probability score wherein further segregation can be obtained as below
Converted_prob > 0.95 -> Priority 1 customers
Converted_prob > 0.90 -> Priority 2 customers
Lead_Score: Converted_prob converted in percentage (%).



Converted	Prospect ID	Converted_prob	final_predicted	Lead_Score
1	4	0.96606009	1	97
1	10	0.986951027	1	99
1	27	0.986125691	1	99
1	75	0.993903031	1	99
1	135	0.940849119	1	94
1	154	0.984970307	1	98
1	163	0.971288278	1	97
1	174	0.95522184	1	96
1	187	0.983319387	1	98
1	189	0.931808759	1	93
1	205	0.95522184	1	96
1	219	0.987625138	1	99
1	223	0.982066578	1	98
1	255	0.975223596	1	98