Presented by
Vibhudendra S
upGrad & IIITB | Data Science Program - February 2024
Batch ID 5702

Dt. 16.10.2024

Summary report on Lead Scoring Assignment

Business Goal: 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Problem solving approach methodology: Data Science methodology, Machine Learning: Logistic Regression as this is a classification problem, Model evaluation, Insights & recommendations based on feature importance.

Steps followed are as under.

1. Importing the available data set
2. Data Understanding
3. Data Pre-processing
4. EDA
5. Data Preparation
6. Machine Learning/Model building
7. Model evaluation
8. Final model
9. Feature importance
10. Recommendations

1. **Importing the available data set:** 'Leads.csv' was the available dataset which we imported.

2. **Data Understanding:**

   This dataset has 9,240 records & 37 features
   There are 30 categorical features & 7 numerical features

3. **Data Pre-processing**

   - Checked & removed duplicate records, but in this case there was no duplicate records found.
   - There were few features which had ‘select’ as value which was made ‘null’ as applicable in this case.

- Null values handling: 7 features had > 40% null values which we dropped.10 more features had null values < 40% which we dropped few & imputed median & mode values for rest of them. We ended up having 25 features at this stage.
- Checked for class imbalance: Dropped 13 features which had highly imbalanced data.

4. EDA

I. Created numerical & categorical feature lists

II. Created pairplots, countplots, boxplots & heatmap

III. Insights & Recommendations are as under

- Lead Origins like 'API' & 'Landing Page Submission' have significant impact on lead conversions.
- Lead Sources like 'Google', 'Direct Traffic', 'Reference', 'Olark Chat' & 'Organic Search' have significant impact on lead conversions.
- Those who have opted for email have significant impact on lead conversions.
- If the Last Activity is either 'SMS sent' or 'Email opened', then they have significant impact on lead conversions.
- Those who have mentioned specialization as 'Others' have significant impact on lead conversions.
- Those who are 'Unemployed' have significant impact on lead conversions.
- Those who have not opted for 'A free copy of Mastering The Interview' have significant impact on lead conversions
- Those who have visited the website atleast 5 times have significant impact on lead conversion rate.
- Those who have spent total time on website of around 1200 units have significant impact on lead conversion rate.
- Those who have viewed Average number of pages on the website of atleast 3 have significant impact on lead conversion rate
- 'Page Views Per Visit' & 'Total Visits' have a high correlation of 0.72

IV. Outliers & its handling: 'Total Visits' & 'Page Views Per Visit' have outliers on the upper range & required Outlier Treatment. From describe function previously, it was observed that 75th percentile value of 'Total Visits' is 5 & for 'Page Views Per Visit' is 3. Hence capped maximum values to 99th percentile.

5. Data Preparation:

- Converted 2 features binary values (Yes/No) to 1/0
- Converted 5 features 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization' & 'What is your current occupation' into numerical features using dummies
- We had totally all 11 as numerical features at this stage which was good enough for model building.
- Feature Scaling: on 3 features using Standard Scaler from sklearn
- We had almost 38% lead converted rate.
- 'Do Not Email' & 'Last Activity_email bounced' have high correlation of 0.63.
- 'Lead Origin_Lead Add Form' and 'Lead Source_Referance' having higher correlation of 0.85.

- 'Lead Source_Facebook' and 'Lead Origin_Lead Import' having higher correlation of 0.98
- 'TotalVisits' and 'Page Views Per Visit' having correlation of 0.72

6. Machine Learning/Model building

- Used Logistic Regression from sklearn
- Feature selection using RFE(selected 15 important features)
- Used statsmodels to add constant & fit the model
- Used summary()/mainly p-values & VIF scores to tune the model
- Final model(Third model) had 13 features, all significant p-values & all VIF scores below 5.
- Using final model, predicted train dataset values.
- Found out Optimal cut-off point using accuracy, sensitivity & specificity scores as 0.35
- Tuned predicted values based on 0.35 threshold & added lead scores based on the same.
- The cut-off point of Precision Recall curve was around 0.45
- Plotted ROC curve & had AUC score of 0.89

7. Model Evaluation

- Model was evaluated using Precision Recall cut-off of 0.45 which reduced both sensitivity & Recall scores. This was not ideal for the nature of business problem had on hand & hence continued based on 0.35 threshold only.
- Made predictions on test set

Train data evaluation metrics
Accuracy : 81%
Sensitivity : 81%
Specificity : 80%
Precision: 72%
Recall: 81%

Test data evaluation metrics
Accuracy : 80%
Sensitivity : 81%
Specificity : 80%
Precision: 72%
Recall: 81%

8. Final Model

-0.2785 - 1.6051xDo Not Email + 1.0675xTotal Time Spent on Website - 0.8760xLead Origin_Landing Page Submission + 2.8909xLead Origin_Lead Add Form + 1.0952xLead Source_Olark Chat + 3.1667xLead Source_Welingak Website - 1.1395xLast Activity_Converted to Lead - 1.2742xLast Activity_Olark Chat Conversation + 1.3083xLast Activity_SMS Sent + 1.3908xLast Activity_Unsubscribed - 0.8366xSpecialization_Other -

1.1914xWhat is your current occupation_Other + 2.3946xWhat is your current occupation_Working Professional

9. Feature Importance

Features importance with respective percentages are as under

Lead Source_Welingak Website	100.000000
Lead Origin_Lead Add Form	91.291311
What is your current occupation_Working Professional	75.617920
Last Activity_Unsubscribed	43.917779
Last Activity_SMS Sent	41.314677
Lead Source_Olark Chat	34.584874
Total Time Spent on Website	33.709718
Specialization_Other	-26.417332
Lead Origin_Landing Page Submission	-27.661534
Last Activity_Converted to Lead	-35.984065
What is your current occupation_Other	-37.624009
Last Activity_Olark Chat Conversation	-40.236284
Do Not Email	-50.687170

10. Recommendations

Top 3 features contributing to lead conversions are

1. Lead Source_Welingak Website
2. Lead Origin_Lead Add Form
3. What is your current occupation_Working Professional

Top 3 features to be focussed more for having significant leads conversions are

1. Do Not Email
2. Last Activity_Olark Chat Conversation
3. What is your current occupation_Other

- Lead Origins like 'API' & 'Landing Page Submission' have significant impact on lead conversions.
- Lead Sources like 'Google', 'Direct Traffic', 'Reference', 'Olark Chat' & 'Organic Search' have significant impact on lead conversions.
- Those who have opted for email have significant impact on lead conversions.
- If the Last Activity is either 'SMS sent' or 'Email opened', then they have significant impact on lead conversions.
- Those who have mentioned specialization as 'Others' have significant impact on lead conversions.
- Those who are 'Unemployed' have significant impact on lead conversions.
- Those who have not opted for 'A free copy of Mastering The Interview' have significant impact on lead conversions

- Those who have visited the website atleast 5 times have significant impact on lead conversion rate.
- Those who have spent total time on website of around 1200 units have significant impact on lead conversion rate.
- Those who have viewed Average number of pages on the website of atleast 3 have significant impact on lead conversion rate

Few areas they can look into to create more potential leads are as under

a. *E-mail reminder campaigns*
b. *Improvise Marketing strategies around preferred lead sources like 'Google', 'Direct Traffic', 'Reference', 'Olark Chat' & 'Organic Search'.*
c. *Referral programs with eye-catching incentives.*
d. *Target customers whose occupation mentioned as 'others'(to understand more about them & filter out who among these might be really interested in persuing the learning/course)*