LEAD SCORING CASE STUDY

Submitted by:

Deepak Krishna A.R & Nicky Paul Eapen

Problem Statement :

An Education Company, X Education which sells online courses to Industry Professionals aspire to identify and target the most promising leads for the company.

Understanding the Business Objective:

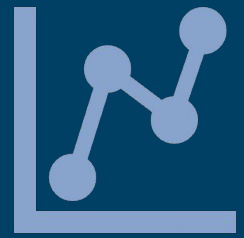
1. The company wants to know the most relevant factors influencing a potential/promising lead
2. The company look forward to increase the lead conversion rate by asking the sales team to focus more on communicating with the potential leads rather than making calls to everyone.
3. To achieve the ballpark of the target lead conversion rate ~80% in the long run

Analysis Goals:

1. To assign a score for each lead by building a logistic regression model to predict the probability that a particular lead could be a promising one
2. To select the features which are most important in determining lead conversion

APPROACH & METHODOLOGY

We used the industry agnostic CRISP-DM methodology to conduct our analysis on the given problem statement
All the underlying analysis have been performed on the below data set



1. leads.csv - base data

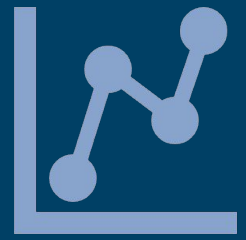
Data Inspection & Data Quality Checks::

1. Missing Values: Few of the columns were dropped. And for some important features, the rows of missing observations were removed.
2. Very less to no variance : Features like ‘Do not Call’ , ‘Get updates on DM content’, ‘Update me on supply chain Content’. These data fields are insignificant and thus removed
3. In total 24 features were removed and we could retain ~ 70% of the original dataset
4. Outlier Analysis: There were two numerical features with extreme outliers: ‘TotalVisits’ and ‘Page Views Per Visit”. These outliers were handled by capping the extreme value to the 0.99th percentile

Model Building- Steps involved:

1. The Dataset is split into train and test data in 70:30 ratio.
2. The numeric values are scaled using ‘MinMax Scaler’ from ‘sklearn’.
3. Correlation between variables was inspected by listing down the top correlated features and removing them thus avoiding any issues of Multicollinearity.
4. Feature Selection is done using Recursive Feature Elimination method to identify the top 15 features
5. The statistics of these selected features were further inspected by generating Generalized Linear Model Regression Result using statsmodel. A total of 3 models were generated by eliminating a feature with high P value in each step.
6. The final model was used to generate a predict the probability of the lead conversion and a random cut off of 50% probability was chosen to predict if the lead is converted or not.

APPROACH & METHODOLOGY



Model Evaluation- Steps involved:

1. To further evaluate the efficiency of the model we calculated the key evaluation metrics such as
    a. Accuracy
    b. Sensitivity
    c. Specificity

    using the confusion matrix and plotted an ROC curve to test the accuracy of the model.

2. An optimum cut off probability value was found out by plotting Accuracy, Sensitivity and Specificity i.e. 0.42
3. The predictions are now re-calculated and validated again by calculating Accuracy, Sensitivity and Specificity.
4. Prediction was done based on our final model and the optimum cut off 0.42. The evaluation metrics were recalculated and the values were very close to what we obtained in the train dataset.

The top 3 features obtained from the model are:

1. Lead Source_Welingak Website
2. Total Time Spent on Website,
3. Lead Source_Reference

Key Recommendation:

X Education should focus on leads that come through Reference and ‘Welingak’ website. Moreover, the total time a prospect spends on the company's portal is a strong indicator of a Hot Lead.

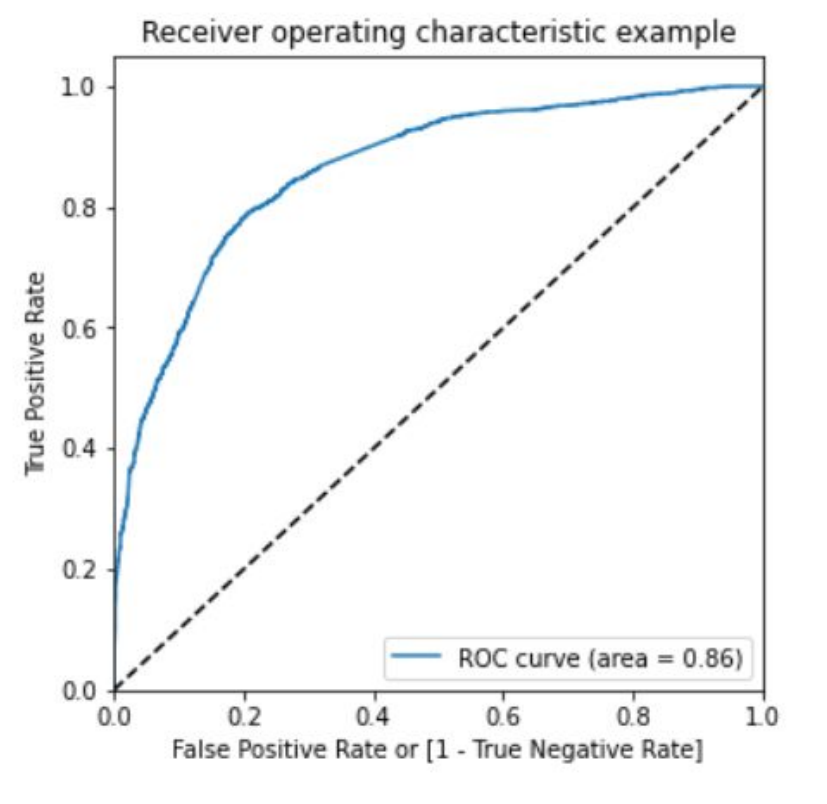
Model Evaluation

Metrics observed on the test dataset

- Accuracy=78.7%
- Sensitivity=78.8%
- Specificity=78.7%

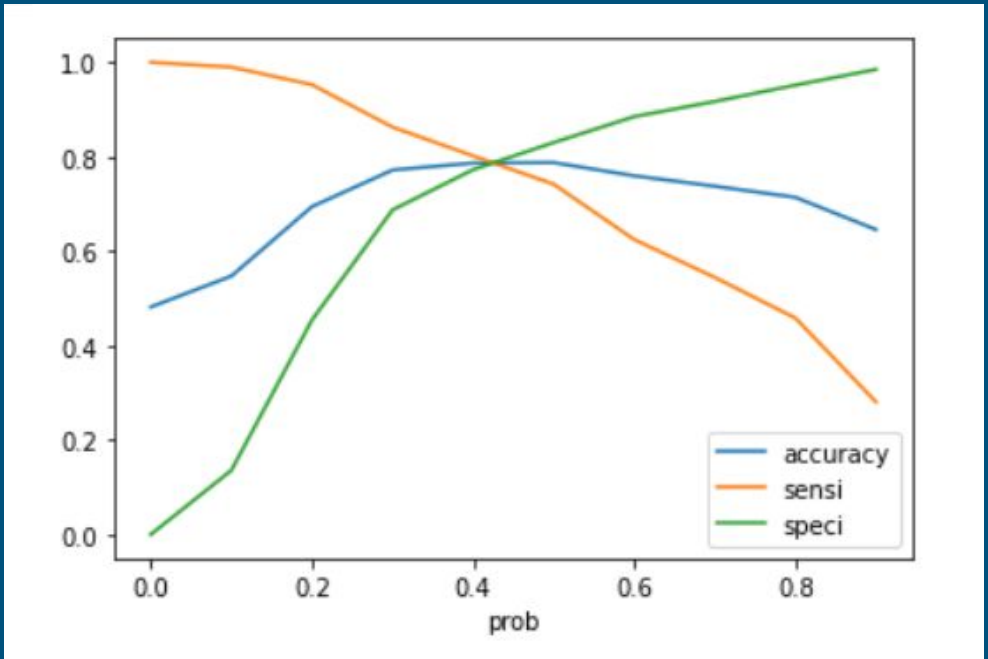
Metrics Observed from the train dataset:

- Accuracy=79%
- Sensitivity=79%
- Specificity=79%



1. Metrics obtained from the train and test datasets are optimal and almost equal.
2. The ROC curve also shows the accuracy of the model. The curve is observed to be close to the left and top borders and is having sufficient area under curve.

Arriving at the Optimum Cut-off Probability



From the Accuracy, Sensitivity and Specificity curve we have arrived at cut off probability of .42

Final Set of Features Identified with their coefficients in the logistics regression equation

const	-0.900209
Do Not Email	-1.437993
TotalVisits	1.370042
Total Time Spent on Website	4.378436
Lead Source_Direct Traffic	-1.787851
Lead Source_Google	-1.322714
Lead Source_Organic Search	-1.604672
Lead Source_Reference	2.830233
Lead Source_Referral Sites	-1.676567
Lead Source_Welingak Website	4.912116
Last Activity_Had a Phone Conversation	2.640942
Last Activity_SMS Sent	1.174794
What is your current occupation_Working Professional	2.573511
Last Notable Activity_Unreachable	2.770880

dtype: float64

Insights & Recommendations:

Key insights and Recommendations we arrived at from the Logistic Regression Model deployed:

1. The leads having a score greater than 42 are considered as hot leads.

2. Leads that come through Reference and Welingak websites are strong indicators of Hot leads. Such leads should be followed up actively

3. Leads who spend more time on the X Edu company's portal tend to be more interested and curious about the course and are highly likely to be a hot lead. The Sales team should focus on them,nurture those leads and can give them more insights about the course. They are promising leads and the conversion rate could be really high from this pool

4. Providing incentives and bonus offers for hot leads can increase their probability of conversion Working Professionals are more promising than unemployed people and have a higher chance of getting converted into the platform

Action Plan to convert promising leads through interns:

1. Hosting Events
   Conducting special online and offline events like webinars, workshops by giving special invites for Hot leads. The events can aim towards giving the audience better insight about the course and provide a platform to answer questions from customers.

2. Live Chat rooms
   The company website can make use of a live chat box to interact and guide with leads. Hot leads can be identified by collecting information from the customers and can be nurtured further.

3. Discount Offers
   Based on Lead Score Customers can be attracted by giving them personalised discount offers and referral Bonuses

4. The sales team, especially the interns can make use of social media to do sales campaign in social media platforms like linkedin, Facebook, twitter, whatsapp, etc. New interns can also be offered referral bonus attained through social media campaigning.

5. Free trials of the course or preview on offer to hot leads which would attract the lead to convert.

Action Plan to convert promising leads before Quarter End:

1.The Sales team should increase the cut off rate to a very high value and focus on high conversion probability customers.

2. Extensive focus on Referrals, as they have more chance of being a hot lead.

3. Identification of people who spent more time on their platform. Pursue them for lead conversion.

4. Target Working Professionals through their organization.

5. Marketing courses that are specific to an organization domain.

6. Social Media handles of the company can be used to write out informative blogs and writings. Testaments of previous learners can be posted on the same platform for a more reliable promotion.