Lead Scoring Case Study Summary

Problem statement: - X Education sells online courses. Leads generated from various sources are captured. There is other metadata around lead that is captured for each lead. Team is assigned to nurture hot leads and convert such potential leads to confirmed opportunity (leads).

Solution –

Effective way of working on the leads is to start with hot leads
i.e. leads that have higher probability of getting converted. This will not only result in higher conversion ratio but also effective use of time. Time spent on nurturing hot leads can be increased whereas time spent on leads with low score (cold leads) can be minimized.

Determining hot and cold leads is done using logistic regression model. Using the meta data provided for each lead, we built a logistic regression model and assigned lead score to each lead.

Data Analysis –

We found that there are lot of columns with missing values in the data. Also, there are columns where default value "Select" is populated. We will be initially considering this as missing values and apply the same missing value treatment for such values.

Data preparation--

Data is prepared for Modelling. Columns with Null Values are either dropped or imputed/replaced with values. Outliers are capped. EDA is performed to find out the key features that could be potentially converted into Hot Leads. Scaling is also performed for uniform data range of columns. Highly correlated column is identified and dropped

Model Building

Data is split into Train and Test data on 70-30% ratio.
RFE is used to identify top 15 variables for modelling. Then we started building the model using Stats Model . All high p-value variables are dropped one by one. VIF are under control. Finally, after 5 iterations, a stable model is built.

Making prediction

Using the model from fifth iteration the optimal cutoff is found using ROC curve and different probabilities, the predictions are made on test dataset.

Model Selection and Lead Score

This model is finally used to assign lead score to all the potential leads with **optimal cut-off probability 0.3.**

Conclusion

Using Metrics like Sensitivity-Specificity as well as Precision and Recall, We came up with an optimal cut off for calculating the final predictions.
- Accuracy, Sensitivity and Specificity values of test set which came out to around 80% are very close to the values calculated using train set.
- Also the lead score calculated in the training and test set of data shows the conversion is around 80% as requested by the CEO of X-Education
- Overall, a good model has been built.

Using the Model built above, we figured out the top 3 features to focus on for turning into potential leads is as follows

Total Time Spent on Website
Lead Add Form (from Lead Origin)
Welingak Website (from Last Source)

Picking up on these leads and working on them would result in time well spent on nurturing and converting them into potential hot leads thereby increasing the revenue.