Lead Score Case Study

Summary Report:

We started working on assignment by first understanding the problem statement. Once we were clear about business perspective and what type of outcome we have to give, we started understanding data set. We started analyzing data and we got to know that our dependent variable is 'Converted'.

Data had a lot of missing values. So, we seek out null values and removed features which have more than 40% of null values, they were not much useful from business perspective. Then we imputed some null values by mean, mode or other suitable values, some times we had to introduce a new category. We were analyzing and visualizing all feature while taking care of null values, it was sort of univariate analysis. After this point we started EDA, and we analyzed all features with respect to target variable. Here we found out very good and deep inferences, which may prove useful from business perspective.

Once we got done with EDA, a lot of features are just redundant and will just make model complex. So, All the features which are not useful got dropped here.

The data we had consists of two categories - Form generated data and Sales team generated, we had to neglect all the feature variables introduced by sales team.

Then We converted all binary variables into 1/0 form. And created dummy variables of remaining categorical variables. Then we split the data into train and test set, to train and test our final model. We also had to do feature scaling as some of the numerical values have different scales. We used RFE to automatically select important features to be used in models.

After all data preparation we started making models, and then removing feature in the basis of their p-value and VIF (variance inflation factor).

We build a logistic regression model for this problem as we had to predict a categorical variable. we used default cutoff around 0.4 and we found out our accuracy was about 82%, sensitivity was 78% , specificity was around 85.33%. But as Business problem stated they needed a model with high sensitivity (they want to increase conversion rate) so we analyzed which cutoff is better to use in this case. When we tried to get decent cutoff, we thought 0.3 would be a great cutoff as it balances all parameters i.e. accuracy, sensitivity and specificity. But we tend to get a little more sensitivity so we trade off some specificity with sensitivity and chose 0.25 as final cutoff.

This gave us around 85% sensitivity, and 79% specificity. Then we tried our model on test set and it was able to predict around 82 % correct values and sensitivity jumped up to 84%.

Finally we ended our work by creating a data table with lead score higher than cutoff and having all hot leads. This will help business in getting more leads and increasing conversion rate.

By-

A V Ramya Keerthana and **Kamesh Vishwakarma**