

Lead Scoring Case Study Summary

Submitted By:

Siddharth Aher (siddaher479@gmail.com)

Prerna Gupta (prerna.gupta1792@gmail.com)

Ruchi Rawat (ruchi93rawat@gmail.com)

Objective:

This case study is done for X Education company as they get a lot of leads, but its lead conversion rate is very poor at around 30%. The company wanted us to build a logistic regression model wherein we 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. CEO's target for lead conversion rate is around 80%.

Data Manipulation:

- Many values in some columns were marked as "Select". There was no such specific option given so we replaced the select values as "Null".
- Columns with more than 40% missing values were dropped.
- Categorical data was imputed with mode and columns with only one unique response from the customer were dropped.
- In column "Specialization", it may be possible that the leads may leave this column blank if he/she is a student or their specialization is not mentioned in the options given.
- For columns where missing values were less than 2%, the corresponding rows with missing values were dropped.
- Outliers were removed from the numerical columns using 1.5 IQR rule.

Exploratory Data Analysis:

- Data imbalance checked- only 38% of leads converted.
- Performed univariate and bivariate analysis for categorical and numerical variables. 'Lead Origin', 'Current occupation', 'Lead Source', etc. provide valuable insight into the effect on target variable 'Converted' which indicates whether lead was converted or not.
- Time spent on the website shows a positive impact on lead conversion.

Model Building:

- Used RFE to reduce variables to 20. This will make data frame more manageable.
- The First basic test for regression was performing a train test split, we had chosen a 70:30 ratio.
- Built multiple Models by removing the variable whose p-value is greater than 0.05 and whose VIF value is greater than 5 one by one.
- Used ROC curve, hyper parameter tuning to find the optimal cut off point
- Overall Accuracy of final model: 80%
- Recall or Sensitivity of final model: 80%

Recommendations:

- More budget/spending can be done on the Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing references that convert to lead, encourage to provide more references.
- Working professionals to be aggressively targeted as they have a high conversion rate and will have better financial situations to pay higher fees too.