
Executive Summary

As very well mentioned in the problem statement we need to build a model to assign a lead score by analyzing the provided data using Logistic Regression Method and find out the variables which directly affects the lead conversion. The method used is a type of Machine Learning method which conclude the desired result by applying the statistics concepts on the data.
The conclusion comes out to be that Dependent Variable, Conversion Score (y) is defined in terms of independent variables (x)

Introduction

The problem statement comprises of an online education agency who is gathering the Data of leads by collecting the online information they fill while visiting the site or clicking the ads on google/any other site like no. of clicks, time spent on the web-page, age, gender etc. This information and analyzing the same to classify the leads based on their seriousness and interest. The genuine leads shall be contacted and will be tried to convert them into a potential prospect

Methodology

We have been provided with a lead's dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

To do this we have applied the Logistic Regression on the Dataset provided, The steps are as under:

1. **Data Cleaning:** The data was cleaned by dropping the irrelevant variables, checking the null values and replacing them if required.
2. **Data Preparation**: Mapping the YES/NO to 1's & 0's , Creating the Dummy variables for the applicable columns, checking for outliers using Quantiles function.
3. **Train-test Split**: The data split in Train and Test Set to perform the regression.
4. **Feature Scaling**: Feature Scaling is performed on the Dataset.
5. **Correlation Matrix & RFE**: Correlation matrix is plotted to see the correlation among the variables but due to high number of variables RFE is applied then correlation was checked again.
6. **Linear Logistic Regression Model**: Regression model is applied on the Train Set and after 7 iterations we got the desired values and accuracy.
7. **Confusion Matrix**: To check the overall accuracy confusion matrix technique is applied and precision score was checked which comes out to be .78 and accuracy score was .81. Therefore, lead conversion rate is 78% and model's overall accuracy is 80%
8. **ROC Curve**: To find out the optimum cut-off ROC curve was plotted, and it showed 60 is the optimum lead score at which lead conversion rate is more than 80%.

Analysis

The Analysis depicts that the final model has a lead score 60 and overall accuracy 80% & lead conversion rate is 83%

Conclusions and Recommendations

The conclusion is simple, the leads are now classified based on their lead score into two segments of 1 and 0.

The concerns can now follow-up with the leads having convert=1 with 80% chances of conversion.

Learnings

We have learned how a product marketing & Sales can optimized using a Regression Model by just using the very basic website visitor's data.