X Education - Lead Score Case Study

Summary

We have completed this case study, we applied our best learning experience to cover all the aspects of Python code & libraries, EDA, Logistic Regression Model, Model Evaluation. We did some research to gain domain knowledge of dataset, to understand the impact of variables in this industry and Our approach to this case study is as following;

1. Problem Statement

We have understood the problem of X Education company. The target of the company is to increase lead conversion rate and wants to predict the “Hot Leads” which have higher probability of getting converted, hence the company can focus on such leads more.

2. Data Cleaning

1. We checked our data size, number of columns, null values, missing values.
2. We have properly handled missing values, imputation of missing values in the dataset.
3. Data Imbalance is a major concern while building the Model, so we have taken appropriate action to handle the data imbalance.
4. We found outliers in some variables and treated these appropriately.
5. Data types were checked and corrected.

3. EDA (Exploratory Data Analysis)

After data cleaning, we performed EDA to identify patterns, key insights, correlations which ultimately helped us to sense the important features which needed to be included while building the model.

We could identify some important variables through EDA.

4. Data Preparation

After EDA we created dummy variables for categorical features, and scaled up the numeric variables using standard scaler. Again we dropped some variables which we found not useful for model building.

We did a train & test split of data in a 70:30 ratio.

5. Model Building

We tried some Models and checked the result parameters. We used RFE (Recursive Feature Elimination) to eliminate statistical less important features. We checked P value to determine if any feature is insignificant, we also checked VIF (Variance Inflation Factor) to determine if multicollinearity exists.

Finally after checking all the statistical parameters, having sense of important features to include, we arrived at our final model.

6. Model Evaluation

We checked our model accuracy, also we evaluated our model through a confusion matrix, ROC curve to determine model efficiency, then we performed trade off between precision and recall to get the cut off point i.e. 0.42

7. Conclusion

Our final model is able to predict the test data with 80% accuracy, and we have suggested the company to focus on leads which have higher probability then cut off point and as predicted as potential lead by our model to increase the lead conversion ratio.

Submitted by (batch DS C52 23):

1. Rathina Rajeswari
2. Ranith Sardar
3. Praveen Rathi