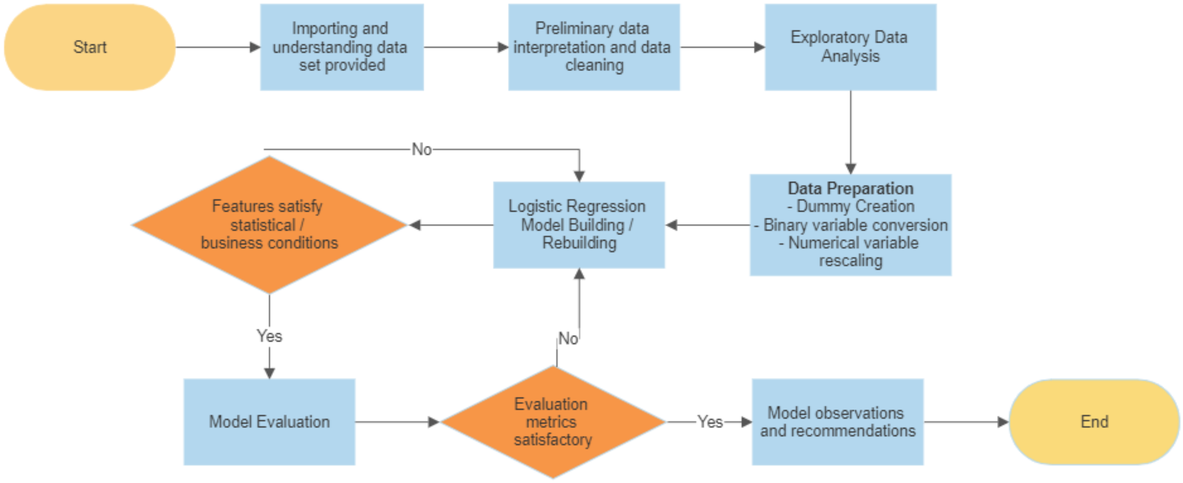
Lead Scoring Case Study

By Soumyajit Mitra, Kalaivannan S and Sonal Verma

The main aim was to have a conversion rate among hot leads identified by a machine learning model to be 80% as demanded by X Education. This was a classification problem and hence a logistic regression model was used to predict whether a particular lead is a hot lead or not.

The approach strategy was as indicated in the following flow chart and as described henceforth:


Start
Importing and understanding data set provided
Preliminary data interpretation and data cleaning
Exploratory Data Analysis
No
Features satisfy statistical / business conditions
Logistic Regression Model Building / Rebuilding
Data Preparation
- Dummy Creation
- Binary variable conversion
- Numerical variable rescaling
Yes
No
Model Evaluation
Evaluation metrics satisfactory
Yes
Model observations and recommendations
End

1. **Data reading and understanding** – Data was provided in excel format with ~9000 entries. After importing in python, preliminary understanding of the basic structure of data frame was performed using python features and functions like shape, describe() and info()
2. **Data cleaning** - From above step it was found out that the dataset had missing values, outliers and some wrong entries which had to be imputed or modified. Columns containing missing values above 30% were dropped, some categorical columns had 'Select' as an entry which meant that the lead might have not selected any category while filling the form, which was replace with null and another missing value check was done, further removing columns with missing value above 30%. Some had repeated entries like "google" and "Google" which conveyed the same meaning which was converted to single column. Numerical outliers were treated using either medians or modes. Also, columns with very high data imbalance were dropped (~99%+) due to very low sample size of one category for the algorithm to learn
3. **Exploratory data analysis** – Was performed to see the distribution of converted and non-converted leads within various categorical variables. Boxplots were plotted to see the ranges of numerical valued variables in the converted and non-converted leads. Target variable data imbalance was checked to make sure that we have a good sample size of either kinds of leads
4. **Data preparation** – Numerical data was rescaled, binary categories converted to 0's and 1's and multicategory variables were converted to dummy variables

5. **Modelling** – Started with 20 top features using Recursive feature selection and iterated till model 6 where 16 variables with acceptable p-values and VIFs were obtained
6. **Model evaluation** – The model was evaluated against an arbitrary cut-off probability score of 50/100, above 50 to be predicted converted and below it as non-converted. Post preliminary checks using confusion matrix and related metrics, optimum score/probability cut-off was found using sensitivity, specificity and accuracy. Using optimum cut-off, model was tested but the precision and positive prediction value was found to be less than 80%, hence, cut-off had to be optimised as per business requirements. Further iterations lead to cut-off score of 50.5 yielding precision of ~80% in train set
7. **Test set application** – Model applied on test set revealed ~81% precision, which satisfied business needs. Model was validated.
8. **Conclusion** – Three most important variables were ‘Total Time Spent on Website’,’ Last Notable Activity Having A Phone Conversation’ and ‘Lead Origin Lead Add Form’ that increased conversion probability by the most as per their coefficients