Case Study Summary

Lead scoring case study using logistic regression in python

Our approach to Case Study comprises of 4 parts Data Cleaning, EDA, Model building and testing, Metrics evaluation.

Data Cleaning:

- Column renaming:
    a. We noticed columns that were too lengthy to read, therefore renamed it based on the context.

- Constant Features:
    a. The dataset had many single labelled categorical features, since these features contribute less to the model's performance, we decided to drop all those features.

- Missing value imputation
    a. Initially when looked at the dataset, there were a lot of unknowns, also it had a lot of features, understanding the dataset took some time,
    b. Additionally, there were a lot missing values and labels like 'select' in some features,
    c. We came to conclusion that they are of no meaning to a particular feature,
    d. We therefore had to devise a plan to either impute or drop those values, making this call took some time, in the end we dropped most of those values.

- Outlier treatment
    a. Very few features were numerical in nature, such features had few outliers. we decided to cap those outliers.

- Label Encoding/Dummy variable creation for categorical variables
    a. Since there were lot of categorical features, we thought using label encoding will help reduce the features,
    b. whereas dummy encoding increases the features count multifold, so we tried out both the approaches, the outcome results differed significantly

- Test-train split of the data
    a. For Model Building purpose we split the data into train and test set at a ratio of 70:30

- Standardization of the scales of continuous variables
    a. we planned to two try different types of scaling, namely StandardScaler and MinMaxScaler.

Overall, the data cleaning took a lot of time. As we progressed towards the solution, we were able to solve most of the unknowns. therefore, it is necessary to allocate more time to understand the data and chalk out steps required, before developing the solution. Also, a little bit of flexibility is necessary to overcome any unforeseen challenges.

Exploratory Data Analysis:

- Univariate
- Bivariate
- Multivariate analysis

Model Building Approach:

we tried to find out that for this particular requirement there were two outcomes one without dropping rows we found out that tags were more important in terms of business outcome but in other approach we found out that after dropping columns the other variable had higher priority so overall it depends upon our business requirement. more the features we have to find out a way to reduce the features therefore feature selection plays an important role in the final outcome of the model matrix

Model Building:

a. logistic regression model was built in Python using the function GLM() under statsmodel library.
b. This model contained all the variables, some of which had insignificant coefficients.
c. Hence, some of these variables were removed first based on an automated approach, i.e. RFE and then a manual approach based on the VIFs and p-values

Metrics Evaluation:

a. For Model building purposes we had few approaches planned beforehand.
b. Similarly, after building the model, we try to find out the sensitivity and specificity as well as precision and recall
c. it is very important to understand the business requirement and based on which we have to recalibrate our final outcome model tuning needs to be based on the business requirement,

Overall, the case study was a great experience and challenging, it has helped us understand various challenges that one face during building a Machine Learning model.