Lead Scoring Case Study Summary

Problem Understanding

We tried to understand what are the various factors how a lead gets converted. We googled and found blogs, medium articles etc. on this topic. We also tried to visualize how we were leads in a few cases for e.g. when querying for upgrad course and what were the factors that impacted whether we got converted or not. Now the columns started making more sense to us.

Data Understanding And Cleaning

On looking at column names we found out many column names were not appropriate. We renamed them. We removed **Prospect Id** and **Lead Number** as they were just unique identifiers. We converted columns having Yes/No to 1/0. **Do not Email** and **Do not Call** were renamed to **Email** and **Call** with their values negated i.e. Yes became no and vice versa as it is sometimes difficult to understand if the attribute is in not format. We treated **Select** as null values. We calculated the null value percentage and **dropped the column with null value more than 10%**. Rest was having less than 2% null values. They were filled using median or mode.

EDA

We performed Univariate, Bivariate and multivariate analysis on the cleaned data. Few columns were filled with the same values more than 95%. Those columns were dropped. Also few columns were having a lot of levels in a feature with very less percentage. They were combined under other levels in that column. Few numerical columns were having outliers. So there we took up to 95th percentile. Then we compared all other columns against the target variable converted. We had some interesting results which are shared in ppt as well as documented in Notebook.

Model Building

We divided the data into a train and test set with 0.7 and 0.3 ratio. We checked the data, created a **dummy** for all the categorical variables and then scale and fit the numerical data using **Standard Scaler**.We used Scikit RFE to arrive at the top 15 features. Then we built the model using statsmodel taking these top 15 features into account. Then we checked for the VIF of features. We removed features having more than **5 VIF**. We took a look at the model stats and removed the features having more than 0.02 p-value. Finally we created the model with the final 9 features. We built the ROC curve. Now we calculated sensitivity, specificity and accuracy at different cutoff values between 0 and 1 with step size 0.1. We plotted its graph and 0.3 was found out as the ideal cutoff point where all the three graphs intersected. **So in ideal condition 0.3 was decided to be an ideal cutoff point. Following were values at 0.3:-**

accuracy = 0.7990819209039548
sensitivity = 0.8112566715186803
specificity = 0.7921176797113516

Lead Scoring Case Study Summary

Model Evaluation on Test Data

We predict the y value using the same model on the test data. Here also it gives the almost same performance.
accuracy = 0.7969522240527183
sensitivity = 0.8060538116591929
specificity = 0.7916666666666666

Finally we create a new column lead-score and assign it the value by multiplying probability by 100.

Case Study Partners-

1. Keshav Raj
2. Karishma Sahay
3. Swati Mehta