Lead scoring case study report

Objective:

To build a model such that a lead score is assigned to each of the leads in such a way that the customers with higher lead score have a higher conversion chance, and the customers with lower lead score have a lower conversion chance. And the target lead conversion rate is around 80%.

Analysis methodology:

1. **Loading the data**
2. **Understanding the data:** In the given assignment, we have leads dataset from the past with around 9240 data points. The shape of the dataset is (9240, 37). The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.
3. **Data cleaning:**
    a. On inspecting, we saw that several columns have select as a value, which is a result of the prospect is not selecting any value in that field, which means it is as good as having Null. Therefore replacing all select values with Null.
    b. Null check value: Dropped 10 columns which had null percentage of more than 30%.
    c. On inspecting further, saw that most of the columns have 1 or 2 unique values and the majority of then have almost 100% of rows containing the same value. Hence dropped all those columns -around 16 columns.
    d. Filtered few of the nulls from the rows.
    e. Removed outliers from 'total visit' and 'total time spent on website' columns.
    f. At the end of data cleaning, we were left with 9 columns and 8901 data points.
4. **Univariate analysis:** Visualized the original data variables to look for any pattern or correlation. Few of the observations were - 'Olark chat conversation' has the highest successful conversion count compared to other last notable activities. Similarly, SMS sent has a higher successful conversion count compared to other last activities.It was also observed that the individuals who are working tend to take up the course more than the un employed individuals. When coming to the Lead Profile, Dual specialization and Lateral Students tend to have far higher conversion rate. Hence emphasis should be made to reach such individuals
5. **Dummy creation** – Created the dummy variables of all the categorical variables.
6. Converted 'A free copy of Mastering The Interview' – binary variables (Yes/No) to 0/1
7. **Test- train split:** Split the data with 70% of the data as the train set.
8. **Scaling the data:** Standardised the columns- 'TotalVisits', 'Total Time Spent on Website' and 'Page Views Per Visit'.
9. **Model building:**
    a. As an initial step, ran the stats GLM model on train dataset. Saw that many features are not significant as it had large p- values.
    b. We chose top 20 features using RFE and then ended up with 14 features after manual elimination.
    c. ROC curve was plotted and the cutoff point was chosen which is 0.34 in this case

d. The model parameters for test and train sets were consistent and hence the same model was fitted with the entire data to obtain final result
e. It can be seen that Lead Origin_Lead Add Form, Last Activity_Had a Phone Conversation and Lead Origin Lead Import are the variables which have the highest positive effect on conversion