Summary

In summary, the purpose of this analysis is to help X Education attract more industry professionals to their courses. We learned a lot about the potential consumers' visitation patterns, length of stay, method of access, locations, job information, conversion rate, and other facts from the data that was provided. A machine learning model was developed to identify the potential consumers who are most likely to become leads for X Education based on an analysis of their demographics and behavioral patterns. A likelihood for each consumer to become a lead was also provided by the model. We followed the below steps to generate the score for each lead:

1. We Loaded the data i.e. the CSV file into Jupyter Notebook.
2. We Performed missing value treatment for all the columns and outlier treatment on all the important numeric columns and we dropped the columns which had more than 40% missing values.
3. After that we performed EDA on each categorical and numerical columns to visualize the data for data imbalance and provided the Inferences of it.
4. We then created the Dummy Variables of necessary features using “get_dummies” function in pandas library.
5. Then we did the dataset splitting into train and test dataset using “train_test_split” in sklearn.model_selection library with 70% data in training dataset and 30% data in test dataset.
6. Then we have Scaled the train data using MinMaxScaler. We did this to bring all the features in the dataset to the same magnitude and also to avoid overfitting the model.
7. We started building the model and we used GLM method in statsmodels library to perform Logistic Regression and build a classification model. We initially built the model on the whole prepared dataset.
8. Then we applied RFE (Recursive Feature Elimination) on scaled data for coarse tuning and to find the Top 15 features for the model.
9. Then we built the model on the top 15 features outputted by RFE in an iteration and checked the coefficients, p-value and VIF of the selected features.
10. We Performed 5 iterations after dropping high p-value features one by one. We also checked the VIF score on each iteration and in each iteration, the VIF scores for the features were very low, which indicated that there was no multicollinearity between the independent features. The 5th model was the final and the most efficient model.
11. Then we started evaluating the final model through which we predicted on the training dataset based on the selected features.
12. This model predicted in terms of probability. Therefore, we assumed the initial probability cut-off of greater than 0.5 to be converted as a lead and based on that we created predicted conversion. If the predicted probability is > 0.5, the lead will be converted (1) or else it will not convert (0).
13. After predictions on the training dataset, we created the confusion matrix and checked for Accuracy (92.3%), Specificity (91.8%) and Sensitivity (86.9%).

14. Then, based on this model, we plotted the Receiver Operating Characteristic (ROC) curve to check the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) and we noticed that the model covered 97% under the curve.
15. To find the optimal cut-off probability for the model for balancing between the Sensitivity and Specificity, we created confusion matrix for cut-off of the range between 0.0 to 0.9 probability for every 0.1 increase.
16. Then we plotted Accuracy, Sensitivity and Specificity of each probability. The intersection point of these 3 curves is the optimal cut-off probability value where Sensitivity and Specificity will be balanced. We got the optimal cut-off as 0.3.
17. Then we re-evaluated the model based on 0.3 probability cut-off and found similar values for Accuracy (90.9%), Specificity (95.6%) and Sensitivity (86.9%). This was because, after finding and plotting the Accuracy, Specificity, and Sensitivity for each cut-off between 0.0 and 0.9, we found that the range between 0.3 and 0.5 was the optimal range of cut-off values.
18. After that, to make predictions on the test dataset, we Performed Scaling on the features of the test data.
19. Then we performed prediction of lead conversion on test data.
20. Then, the model was evaluated for the prediction on test data by Accuracy, Specificity and Sensitivity. And we found Accuracy as 90.7%, Specificity as 90.2%, and Sensitivity as 91.4%.
21. We determined that our model is working equally well on unseen data.
22. Then we created lead conversion score, which is (conversion probability * 100) to give a score between 0 to 100 where higher the value means the lead is “hot” and there is high possibility that the lead can be converted. The lower lead score would imply that it is not a hot lead.
23. Then, we determined the importance of each feature by their coefficient values and the relative coefficient value with the highest coefficient value and plotted it as well. Finally, we saw the top 3 features of the model that have the highest contribution in predicting the probability of a lead getting converted.

There are many learning we gathered from this assignment, these are as follows:

1. To predict models, we learned how to deal with outliers and missing values.

2. We learned understanding of the creation of dummy variables and how to use them to our model.

3. Next, we learned how to use Python and its tools to create a Logistic Regression model based on selected features.

4. We have learned how to use statistical data, such as the p-value and VIF, among others, to evaluate the multicollinearity in the model.

5. We have learned how to evaluate and make our model accurate using evaluation matrix like accuracy, sensitivity, specificity, ROC curve and etc

6. We've learned how to tackle a problem statement as a team to get the most out of it and how to work as an integrated team.