The X Education company which is trying to solve their business problem of getting the lead and converting that is addressed in this case study

1. Read and understand the data given in two spread sheets
2. Understand the data structure, information about data like data types, missing values
3. Perform EDA:
    a. Investigate the missing values and impute that with suitable method viz, mean, mode, median etcc
    b. Check outliers in data and take suitable action
    c. Check data imbalance and treat that
    d. Cleaned data to investigate the relationship of variables with that of TARGET variable in this case it is "CONVERTED or NOT CONVERTED"
4. Get the train and test set by splitting the cleaned data set
5. Using sklearn library performed RFE and selected 15 features.
6. Using stats module perform Logistic Regression.
7. The different models are built based on VIF, coefficients, p-value and VIF
8. This process was done iteratively so as to get low VIF (i.e., less than 5) and low p-value (i.e., less than 0.05)
9. So, 2 iterations were performed
10. With this model the confusion matrix is created for train data and then Accuracy, precision, recall, sensitivity and specificity is calculated
11. This process is done for different probability cut off (0.1-0.9 range) and then optimum value of that is calculated based on RoC curve.
12. RoC curve:
    a. It shows the trade-off between sensitivity and specificity
    b. The higher the area of curve the more accurate is the prediction.
    c. If the curve is closer to 45 Degree line, then it is less accurate.
13. Finally, optimum value of probability cut off is achieved at 0.25.
14. Some of the key criteria based on which decision were made are sensitivity, specificity, precision, recall. But there always a trade-off for these two quantities. Hence according to business problem at hand the criteria need to be chosen. For this creating confusion matrix is key.
15. Then similar approach has been taken care for test data set. This is done by transform the test data
16. Then lead score is calculated and what are variable which could affect for conversion is analysed
17. The lead score is evaluated as lead conversion score = (conversion probability * 100). This gives value between 0 to 100. If the lead score is higher then there is possibility of converting this led to a potential customer and other can be revisited by understanding the business scenarios
18. Also, few variables which need more attention from business owner has also been enumerated.

This case study helps to learns:

1. How to formulate problem statement.
2. How to understand data and understand its nature
3. How to perform EDA, Cleaning data, data imputations, checking outlier and treating them
4. Skleran and statsmodule usage.

5. How to perform regression analysis and get the best models based on certain critical criteria viz. sensitivity and specificity or precision and recall
6. This case study also helps to understand the perspective of other team member and his inputs were valuable for solving business problems