Summary and Learnings

Steps involved:

- Data cleaning, manipulation, and transformation:
    - Handling missing values (dropping columns with too many missing values and imputing some columns to not lose out on too much information).
    - Handling outliers in the numerical columns (dropping values less than 1st percentile and greater than 99th percentile if the column has outliers).
    - Correcting spelling mistakes across columns and making the data consistent.
    - The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not provided' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', Outside India' and *not provided'.
- EDA
    - Univariate analysis:
        - On numerical features (identified outliers and dropped them).
        - On categorical variables (identified columns with little to no variance and dropping them because they don’t add any value to our analysis).
    - Bivariate analysis:
        - Identifying correlations between different variables.
    - It was found that a lot of elements in the categorical variables were irrelevant.
- Feature Scaling using Min/max scaling.
- Applying OHE (one hot encoding) to categorical variables and converting them to dummy variables.
- Model Building:
    - Selecting initial feature set using RFE (top 15 features).
    - Using statsmodel to identify statistically insignificant variables and dropping them, thereby creating a better model.
    - Validation of model using metrics such as accuracy, specificity, and sensitivity.
    - Fine-tuning the model by selecting an optimal cutoff using the ROC curve and maintaining a balance between accuracy, specificity and sensitivity across different cutoffs and selecting the best one.

Learnings:

- How to approach a business problem and build end-to-end model, keeping interpretability in mind.
- Handling outliers in numerical features.
- Determining best set of features based on p-values provided by statsmodel.
- How to find optimal cutoff using ROC curve and metrics such as accuracy, specificity and sensitivity.