**DISCUSSION QUESTIONS**

***CONCEPTUAL APPROACH***

* Split the data into train and validation sets. Looked at the event rate in the training data (i.e. % of records with target = 1) ~ 4%
* Did a univariate EDA (exploratory data analysis) to understand the behavior of each variable with target variable –

1. For continuous variable: Plotted boxplots and conditional density plots
2. For categorical variable: Plotted bar graphs

* Imputed the missing values in the training data –

1. For categorical variable: calculated the proportion of each level and then distributed the missing values in the same proportion.
2. For continuous variable: imputed by median

* Since there were large number of variables, applied VIF (for continuous variables) and set a threshold of 5. Continued to removed variables till all variables have VIF < 5
* Applied PCA but the results were not significant.
* Calculated WOE and IV (for categorical variables) and removed variables with insignificant IV value.
* Applied Logistic Regression (ran a few iterations wherein variables with p-value > 0.05 were removed) to get the predictions. Used Cross validation technique as well.

***MODEL PERFORMANCE***

* ROC curve and AUC: Plotted the ROC curve to evaluate the performance of logit model compared to a random model.
* Decile coverage for target =1: Arranged the predicted probabilities in descending order and then made decile group. Ideally, the # of cases where target =1 should be maximum in top (first) decile and should decrease continuously, leading to minimum number of such cases in last decile (where probabilities are small)
* RESULTS – The AUC was 60% on both the training and validation set.

***IMPROVEMENTS***

* Considering there was some memory issue, to evaluate the model performance –

1. I would have calculated concordance and discordance, and
2. Used optimal cut points criteria (identifying point where sensitivity and specificity is maximum) to extract the confusion matrix.

* Tried building different models like random forest/XGBoost to obtain better model accuracy.