The Amazing Bank (AB) is one of the leading financial institutions in the world. You have been recruited as a freelance data science consultant by AB to help the bank design their credit risk strategy, enabling data driven decisions. The CEO of the bank writes the following email to you:

*“Welcome aboard and we are very proud to have you. Our existing credit strategy need some serious fine tuning as it has completely failed to identify potential default behaviors in the post covid world. We need your expertise and help to support us with redesigning our credit risk strategy in this new covid world. We have shared a sample data for you to get started which has data from Mar 2020 till current date; please let us know if you might need more data or any other requirements in specific for you to get started.* ***More than identifying defaulters, we also want to understand why they would default; that’s the key.*** *Very excited to look forward to what you can bring to the table!”*

- Bravo, CEO, Amazing Bank

You skimmed through the data and learnt that there are 1 million customers, 1000 features, with 700 numerical and 300 categorical and 5% defaulters, and there are quite a few missing values as well in different levels. Think aloud and help us understand your approach towards solving this problem!

a) What would be your first step? List different EDA you would like to do with the data before you get started.

**Solution**: Once we have the general idea about the dataset, we can perform basic EDA’s (e.g. handling missing values and outliers, understanding the correlation among features, and treating them):

1. Drop the duplicates.
2. Perform descriptive or summary stats.
3. Identify the missing values, it’s percentages across each column.
4. Handle Missing values and Bad data inside columns.
5. The univariate and bivariate analysis can help for the above scenario’s.
6. Try to convert the categorical columns to Numerics using Label Encoding or One Hot Encoding to gain significant computation power. Eg. Gender to numerics, Timestamp to Month, date, year, hour format.
7. The data type of columns can be corrected.
8. Remove Multicollinearity to reduce the dimensions. Statistical and Domain Knowledge can be used to find the best feature vectors.
9. Detect and Handle outliers.

Generally, while working on a huge dataset, automated EDA libraries (e.g. Pandas\_Profiling, D-Tale, Autoviz, Sweetviz) can be used to expedite the process. Samples can be created to analyze this data efficiently. Jotting down the notes after every step works to enable quick revision.

b) How are you going to handle missing values? Ideate and list them.

**Solution**: There are several treatments for missing values. Some of them are:

1. Drop the missing values: If 80% of the data is missing, drop the variable.
2. If there are few categorical data, the default value is assigned.
3. Impute with Mean, Median or Mode: For the numerical columns, we can replace the missing values with mean, median or mode values. Prob. Density Function graphs can be created to verify that the imputed values should define the same relationship as the original data.
4. Apply unsupervised learning techniques using predictor variables.
5. Use supervised learning for imputation.

c) Before getting into modeling, apart from points a. and b., do you want to do anything else with the data to understand default behavior?

**Solution**: We need to store the cleaned data in a different file to reduce the storage space and increase computation power. Visualize the transformed data and interpret the relationship among multiple feature vectors. Normalization and Standardization can be performed.

d) The default labeling is based on customers who did not pay 3 installments continuously. Do you want to rethink about this labelling strategy for the target? How will you validate the labelling strategy is correct?

**Solution**: As per the domain and past knowledge, I assume that the labeling strategy seems sufficient. This can be validated after we perform our model analysis and analyze the best features.

e) What will be your X and Y?

**Solution**: All the features except the Defaulter class will be the X or the independent feature. The defaulter class will be the Dependent/predictor variable.

f) How are you going to handle outliers, numerical columns and categorical columns?

**Solution**: Handling outliers for Numerical columns:

1. Deleting or Trimming the outlier records.
2. Quantile based flooring and capping (IQR)
3. Using Z-score calculation.
4. When the count is less, assess individually. Box or whisker plots can be created.

Handling outliers for Categorical columns:

1. We can use univariate or bivariate analysis to visualize the categorical columns.
2. List the unique values in the columns and their sum.

g) Do you want to include the entire 1000 features?

**Solution**: Including 1000 features to extract the information can be computationally inefficient, we can reduce the dimensions of the dataset by using Dimensionality reduction techniques, removing multicollinearity, using chi-squared test, using more label encoding instead of One hot encoding will be preferred.

h) What model do you want to choose and why?

**Solution**: As this is the case of class imbalance, we can opt for the following models:

1. IsolationForest
2. LocalOutlierFactor
3. XGBoost
4. RandomForestClassifer
5. SVM
6. ExtraTreesClassifier
7. AdaBoostClassifier

Note: We can go with the automated python libraries (e.g. LazyPredict, EvalML, etc.) to generate various ML algorithms and compare the outputs. This will help us to find the best fit models on the fly.

i) What is your validation strategy?

**Solution**: After computing the ROC-AUC curve and the error/confusion matrix to validate the results, perform the KfoldCrossvalidation techniques and perform hyper tuning of the model to achieve a high level of recall.

j) How are you going to handle class imbalance?

**Solution**: Class imbalance can be handled by the below techniques:

1. RandomOversampling
2. RandomUnderStampling
3. Synthetic minority oversampling technique (SMOTE)
4. Modified synthetic minority oversampling technique (MSMOTE)
5. Isolation Forest Classifier
6. Local Outlier Classifier
7. Bagging and Boosting Algorithms (Examples: XGBoost, LightGBM, Adaboost, RF).

k) What will be your experiments and how are you going to choose the best model?

**Solution**: Using cross Validation techniques and the Hypertuning of the model parameters will be performed to find the best estimators of the model and predict it.

l) What metrics are important for you in evaluating your best model?

**Solution**: Recall seems to be the best metric to evaluate the model.

Recall gives a measure of how accurately our model can identify the relevant data. Thus, for all the customers who actually are defaulters, recall tells us how many we correctly identified as a potential defaulters.

m) “More than identifying defaulters, we also want to understand why they would default; that’s the key” – The CEO specifically mentions this in his email. What is your strategy to address this concern?

**Solution**: Once the model is developed to predict a defaulter with high Recall, we can find the 10-15 best features with the highest importance which describe the reason for the defaulters. Apply clustering algorithms to gain in-depth knowledge. We can use Shapash or Graphviz library to decode the best fit ML model while keeping domain knowledge in mind.