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!*

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

**Ans-**

1**.** If any data is duplicate then drop that duplicated data.

2. check and evaluate with missing data

3. is dataset is balanced or not

4. evaluating outliers

5**. Correlation between the attributes**

6**. Distribution of Skewness, Standard Deviation Values per rows & columns**

7**.** Feature Engineering: - Performing feature engineering by using-

- Permutation Importance

- Partial dependence plots

8. **Handling of imbalanced data**

1. **How are you going to handle missing values? Ideate and list them.**

**Ans-** As it is mentioned that there are quite few missing values so I can handle by the below method according to the data and situation as the dataset is huge (i.e., in millions).

1. Delete Rows:

If the record or row have nan or missing value more than 70% - 75% (means most of features) of features in a particular record is empty then I will delete that particular record as the dataset is huge in millions.

1. Imputation Using (Mean/Median) Values:

This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.

1. Imputation Using (Most Frequent):

**Most Frequent**isanother statistical strategy to impute missing values. It works with categorical features (strings or numerical representations) by replacing missing data with the most frequent values within each column.

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

**Ans-** Yes, I will check whether the dataset is balanced or not. According to the information given I am having a unbalanced data, where 95% of the data is not a default behavior & 5 % of the data is a default behavior.

1. **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?**

**Ans-** Once I get the actual dataset with all the feature then I can think more and analyze and validate whether the labelling strategy is correct or not.

1. **What will be your X and Y?**

**Ans-** Independent features will be X

Dependent lable will be Y

1. **How are you going to handle outliers, numerical columns and categorical columns?**

**Ans- For numerical columns:**

I can use scatterplot, box plot with IQR technique, z-score.

I can drop the idea of **scatterplot** because the scatter plot is the collection of points that shows values for two variables and it is difficult to analyze for 100 features.

I can use **z-score**. The intuition behind Z-score is to describe any data point by finding their relationship with the Standard Deviation and Mean of the group of data points. Z-score is finding the distribution of data where mean is 0 and standard deviation is 1 i.e. normal distribution. while calculating the Z-score I re-scale and center the data and look for data points which are too far from zero. These data points which are way too far from zero will be treated as the outliers. In most of the cases a threshold of 3 or -3 is used i.e if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers.

Another method **Box plot IQR Score:**

The**interquartile range (IQR)**, is a measure of statistical dispersion, being equal to the difference between 75th(Q3) and 25th(Q1) percentiles, or between upper and lower quartiles, IQR = Q3 − Q1.

As I have the IQR scores, Now I calculate the Max and Min of the boxplot.

Maxi = Q3 + 1.5 \* IQR

Mini = Q1 -1.5 \* IQR

The data point which shows greater than Maxi and less than Mini that indicates presence of an outlier.

**For categorical columns:** I guess, If the data values with low frequency - the best way to detect them is frequency distribution and the best way to treat them is by combining them with similar values. Also, I can use that with any continuous variable and do multivariate outlier analysis using box plot.

1. **Do you want to include the entire 1000 features?**

No, I will not include the entire 1000 features.

First of all, I will convert the categorical features into numerical by using Label encoding if category is two, otherwise use of one hot encoding if category is greater than 2.

1)- I will use the **Feature selection by model.**

Some ML models are designed for the feature selection, such as L1-based linear regression and **Ext**remely **Ra**ndomized **Trees** (Extra-trees model). Comparing to L2 regularization, L1 regularization tends to force the parameters of the unimportant features to zero. The extremely randomized trees split the leaf randomly which is not through information gain or entropy. The important features should still be more important than the unimportant features which is measured by the impurity-based feature importance. I also plot the model importance rankings for each model. The model which has the highest accuracy, so I will only select the top 60% features by that model.

Another method **Feature Importance**

Features which were most influential in predicting our target variable. feature importance by **information gain** which measures each feature’s contribution for each tree in XGBoost.

1. **What model do you want to choose and why?**

**Ans-** As the target value is binary class therefore, I will choose the classification models like logistic regression, randomforesrt classifier, extra-tree classifier, **Light GBM, XgBoost classifier.**

1. **What is your validation strategy?**

**Ans-** I will be using stratified cross-validation, because the split preserves the ratio of the categories on both the training and validation datasets. in stratified cross-validation, the split preserves the ratio of the categories on both the training and validation datasets. For example, here I have a dataset with 5% of category 1 and 95% of category 0, and I use stratified cross-validation, I will have the same proportions in training and validation. In contrast, if I use simple cross-validation, in the worst case I may find that there are no samples of category 0 in the validation set.

1. **How are you going to handle class imbalance?**

**Ans-** I am going to use multiple approaches for dealing with imbalanced datasets.

* **Oversample minority class.**

- Adding more copies of minority class.

- It can be a good option we don’t have that much large data to work.

- Drawback of this process is we are adding info. That can lead to overfitting or poor performance on test data.

* **Under sample majority class.**

-Removing some copies of majority class.

-It can be a good option if we have very large amount of data say in millions to work.

-Drawback of this process is we are removing some valuable info. that can lead to underfitting & poor performance on test data.

* **SMOTE (Synthetic Minority Oversampling technique)**

As per the drawbacks of both the model I will use SMOTE (Synthetic Minority Oversampling technique) that is best than the above as compare to above one's.

**SMOTE** is a statistical technique for increasing the number of cases in your dataset in a balanced way. It uses a nearest neighbors’ algorithm to generate new and synthetic data to use for training the model.

1. **What will be your experiments and how are you going to choose the best model?**

**Ans-** GridSearchCV

GridSearchCV combine an estimator with a grid search preamble to tune hyper-parameters. The method picks the optimal parameter from the grid search and uses it with the estimator selected by the user. GridSearchCV inherits the methods from the classifier,I can use the score, predict, etc.. methods directly through the GridSearchCV interface. If I wish to extract the best hyper-parameters identified by the grid search I can use .bestparams and this will return the best hyper-parameter. I can then pass this hyper-parameter to my estimator separately.

Using. predict directly will yield the same results as getting the best hyper-parameter through bestparams  and then using it in my model.

1. **What metrics are important for you in evaluating your best model?**

Ans- These Classification metrics are important for evaluating best model.

1. **Area Under ROC Curve**.

Area Under ROC Curve is a performance metric for binary classification problems.

The AUC represents a model’s ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random. A ROC Curve is a plot of the true positive rate and the false positive rate for a given set of probability predictions at different thresholds used to map the probabilities to class labels. The area under the curve is then the approximate integral under the ROC Curve.

1. **Confusion Matrix.**

The table presents predictions on the x-axis and accuracy outcomes on the y-axis. The cells of the table are the number of predictions made by a machine learning algorithm. For example, a machine learning algorithm can predict 0 or 1 and each prediction may actually have been a 0 or 1. Predictions for 0 that were actually 0 appear in the cell for prediction=0 and actual=0, whereas predictions for 0 that were actually 1 appear in the cell for prediction = 0 and actual=1. And so on.

1. **Classification Report**.

The classification report () function displays the precision, recall, f1-score and support for each class. If f1 score is high then I will consider the best model.

1. “***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?

**Ans**-

I’ve analyzed and pre-processed data, trained and evaluated with models, for their ability to predict defaults and their probability. I used Precision, Recall, F1 and ROCAUC to evaluate the models’ performance at predicting class labels. I used these metrics in particular and discarded Accuracy given that we’re dealing with an imbalanced dataset. I investigated which features were most important to our predictions using feature importance by information gain. With this said, I can round up my strategy to address this concern of how machine learning can be applied to credit risk *strategy*.