1. How should one approach the Credit Card Users Churn Prediction project?

- Before starting the project, please read the problem statement carefully and go through the criteria and descriptions mentioned in the rubric
- Then you should start with an exploratory analysis of the data.
- This understanding will help you identify the need for pre-processing the data.
- Once the data is ready, you can start with the steps that need to be followed as mentioned in the rubric
    - Build 6 models with original data
    - Build 6 models with oversampled data
    - Build 6 models with undersampled data
    - Choose 3 best models among 18 models built in the previous 3 steps
    - Tune 3 models
    - Choose one best model and productionize it using pipelines
- It is important to close the analysis with key findings and recommendations to the business.

2. I am trying to fit a model and getting this error:

```
ValueError: could not convert string to float: 'M'
```

How to resolve?

Please check if the X_train and X_test consists of strings, and then create dummy variables using **pd.get_dummies**

3. I am trying to find out the model performance (confusion matrix, recall, precision, or accuracy) and getting this error:

```
ValueError: pos_label=1 is not a valid label. It should be one of ['Attrited Customer', 'Existing Customer']
```

How to resolve?

The target variable consists of strings and needs to be encoded in 0s and 1s.

Please refer to FMST - MLS1 Notebook for reference.

4. I am trying to tune the decision tree with the pipeline and getting this error:

```
ValueError: Invalid parameter classifier for estimator Pipeline(steps=[('standardscaler', StandardScaler()),  ('decisiontreeclassi
```

```
Check the list of available parameters with `estimator.get_params().keys()`.
```

How to resolve?

The parameter grid passed for tuning is not defined properly. Please find out the correct names of hyperparameters that need to be passed using **pipe.get_params()**

5. I am getting this error while importing SMOTE even after successful installation of imblearn library:

```
ImportError: cannot import name 'delayed' from 'sklearn.utils.fixes' (C:\Users\anaconda3\lib\site-packages\sklearn\utils\fixes.py)
```

How to resolve?

1. Run **!pip install delayed** in your Jupyter notebook.
2. Restart the kernel and try importing SMOTE again.

6. I am getting this error while trying to tune random forest:

```
NotFittedError: All estimators failed to fit
```

How to resolve?

The Numpy library might not be updated. You can update the Numpy library to the latest version using

!pip install numpy==1.20.3 in your Jupyter notebook

OR

pip install numpy==1.20.3 in Anaconda prompt

7. Do we need to do anything with the income variable, mainly around the signs "K","$", "less"? Should we eliminate these and use a different range?

The category names can be renamed but it is not necessary as it won't affect your model in any way.

8. One column has "abc" values. Can I process by dropping these or replacing them with the most frequent values? Which one is recommended?

Dropping the values is not suggested. You can treat them as missing and replace them using an appropriate method.

9. I Did the capping method during EDA as shown in the supermarket campaign model notebook. But that was then capping the extreme outliers only. what about the remaining outliers. can I treat them after Splitting the data, before model building?

For finding outliers you can use the IQR method. But if you want to remove them based on say IQR method (say 25 and 75 percentiles) then after splitting data your test set might have a different range than your training set for a particular feature for which you want to detect the outliers. In this case, after applying IQR your test set might not represent the training set very well and this can reduce the accuracy of the test set.

It's recommended to remove outliers before splitting the dataset. Treating outliers depends on the business problem we need to analyze whether they are outliers or the values that can be possible.

Outliers need to drop because they can make your model worse. According to the business problem if there are more outliers and cannot be dropped just use any transformations such as log or square root so that we can reduce their effect.

10. Why I am getting a different number of columns after imputation and one-hot encoding?

The extra column you are getting is due to a common mistake while using simple imputer.

Whenever we use simple imputer you should fit only once not multiple times.

It's imputing the value 'married' for the missing value of the education_level column and while one-hot encoding it's creating a column for that.

The attached code is the correct way of doing

```
 reqd_col_for_impute = [NAME OF COLUMNS WITH MISSING VALUES]
imputer = SimpleImputer(missing_values=np.nan, strategy="most_frequent")

# Fit and transform the train data
X_train[reqd_col_for_impute] = imputer.fit_transform(X_train[reqd_col_for_impute])

# Transform the validation data
X_val[reqd_col_for_impute] = imputer.transform(X_val[reqd_col_for_impute])

# Transform the test data
X_test[reqd_col_for_impute] = imputer.transform(X_test[reqd_col_for_impute])
```

11. In the MLS session, we were only explained to build logistic regression with oversampled and undersampled data, how to do it for all the models as asked in the rubric?

Oversampling and undersampling are kind of pre-processing techniques, so once you get X_train_over and y_train_over - you can use them to train any model just as you do with X_train and y_train.