Machine Learning FAQs

Problem 1

- **The data set is in xslx format so is it necessary to convert it into CSV?**

when you load data to notebook only difference would be read_csv or read_excel depending on the file type you are reading. Post this read, it is a data frame to work upon and it doesn't matter which type of file data is read.

- **In the first problem, the participants have assessed several parameters on a scale of 1 to 5. Here, is that 5 stands for a high score (for e.g., a participant who gives a score of 5 for the current national economic translates to him/her perceiving the economic conditions to be very good). Is this true, or is it the other way round? Kindly clarify?**

Yes, it is right. 1 represents low and 5 represents a high score.

- **What exactly do we need to consider for model tuning? Do we need to do tuning for all the models?**

you can use GridSearchCV with different values of hyperparameters for tuning and overfitting issues.

- **The initial check reveals that there are 8 duplicate records. Should we really drop them, because the question says that they surveyed 1525 persons and there are only 1525 records? Is there are a possibility that this is a mere coincidence of people having similar attributes (including age) rather than actual duplicates?**

These duplicates need to be dropped because they do not add any value to the study, be it associated with different people.

- **What is expected from problem1- 1.4 and 1.5? How it is different from 1.7?**

Q 1.4, 1.5:- you need to make the models and write the basic information like recall, accuracy, etc. Comment for overfitting and underfitting issues. Explain the hyper-parameters you used in your basic model and the reason behind the selected values or default values.

Q1.7:- You need to write a detailed explanation like AUC/ROC, classification report, confusion matrics, model performance parameters like recall, accuracy, precision for both basic and tuned models using GridSearchCV.

Model Tuning :

This applies to all models including bagging and boosting models. For Example

1. KNN. Tuning is to choose the best k value, distance, and weights

2. Logistic Regression: Different Solvers, C values, and regularization parameters

- **What exactly is expected from model tuning?. Has SMOTE considered a technique of model tuning? Is every model supposed to be tuned? Random Forest itself uses bootstrap bagging, thus it itself represents bagging. Is it correct?**

GridsearchCV is what we expect as model tuning. And every model needs to be tuned. You need to use RF and Bagging as separate models. You can use SMOTE if you think there is a class imbalance problem and prepare a different model and compare it with the original one.

- **What are we expected to do in "Model Tuning" (Carries 2 marks)? Just explain the concept? As we are any which ways checking on the performance of different models to suggest the best.**

you are expected to explain the steps that you adopt for model tuning and the reasons behind choosing those parameters to tune the models. You may also use hyperparameters only for certain models, so in that case, the reason behind this would also be expected.

Model tuning means finding better parameters for the model and not just use the default values. For what parameters can be changed refer to the algorithm documentation.

Each ML algorithm has its own parameters to tune the model. Please check the grid search implementation in mentoring session notebooks. Ex: In random forest, two important hyperparameters are n_estimators,max_features

- **The independent variables apart from age and gender are all ordered. If I do label encoding, then the original order gets changed. In this scenario should encoding be done? if we do one hot encoding then it will result in more than 30 columns as each, is that ok, since the dataset is small? Please advice.**

You can do one-hot encoding or get_dummies for the independent variable gender. For age, you do not need to do any encoding as it should be considered as a continuous variable rather.

While doing the dummies section, you can explicitly mention which variable you want to create a dummy. by doing this, it will affect the other variables.

- **Should we consider converting the ratings on economic condition national, household, Hague, Blair into categorical types?**

It would be better to have the variables in ordinal data types.

- **If both the Classes are equally important, then Which Class can be given label 1 and which one is 0? Which Performance Metrics are critical for model evaluation?**

The class to be given as label 1 is based on the business problem that you want to prove/predict the model for. Both Recall and Precision can be equally important in this case.

Selecting the performance metrics depends on whether the model is part of supervised learning or unsupervised learning. Accordingly, the performance models can be chosen based on all methods that you have learned so far.

- **The features such as 'economic.cond.national', 'economic.cond.household', 'Blair', 'Hague', 'Europe', 'political.knowledge' has dtypes is int64. Is it necessary to convert int64 into an object or we can proceed with int64?**

The machine will pull and load data and define the data type that is most suitable for the values in that column. It is our job to make sure that the correct data type is set for processing, especially when it comes to the point when categorical values having numbers are imported as integers by default. So yes it would be logical to check the data type by looking at the unique values of the column and see if they should actually be numeric or should be treated and worked as categorical.

- **Problem 1 has ordinal values ranging from 1-11,1-5, and even 0-3. Should we encode them or treat them as continuous values?**

Treat the values as it is and build the model for prediction, no need for a change.

- **Problem 1 has features that are ordinal but the rating scales are different. Since Logistic regression is sensitive to scaling, is it necessary to normalize all such features to bring to a similar rating scale (or) do we need to encode such features before various models are applied?**

Scaling is a necessity when using Distance-based models such as KNN etc. Scaling can be done on continuous and ordinal variables. *When it comes to LDA, a suggestion would be to do the encoding.*

- **If by using GridSearchCV we get poorer performance values for recall, F1-score, etc from the tuned model than the base model which has no hyperparameters, should we**

go ahead and choose the base model as our final model? But In GridSearchCV we also use the cross-validation to get the best optimal model, so can we interpret this in the following way: Although the performance is poorer than the base model, it is still an optimized model and will perform better on the unseen data since cross-validation was performed on it along with probable hyperparameter values. So we should go ahead with the model formed using gridsearchCV although its performance is slightly less than the base model. Kindly tell me if this is correct?

Cross-validation can also be performed on the base model. See here we want to see your call while selecting a model, if you think the base model is good go for it.

- **For questions 1.4, 1.5, & 1.6 models are built in jupyter notebooks, and what needs to be shown in the project report?**

All workings will be done as part of the jupyter notebook, in the project report, you can explain the steps involved in building the model, attributes(Hyper-parameters) that you selected, and the reason behind it. Incase of tuning parameters being used, you can explain why you have thought to use it and how it has helped in improving the model. Comment on the values that you took for these parameters, overfitting-underfitting, etc.

- **Do we need to create data points in case of an under-sampling? Not mentioned in the steps so not sure whether the same has to be done.**

If you see the dataset is imbalanced and thus want to do a better tuning check based on balanced data then you should try SMOTE and see the results on the SMOTE dataset. Make sure you apply SMOTE only to train data and not to test data. Then you compare the test results between using the original data Vs SMOTE train data and see which way the model is more generalized.

- **Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. (7 marks) for this, we have to use the random forest as a base estimator for bagging. For boosting we use the ADA boost and gradient boost should we use any base estimator or without any base estimator.**

You are correct, base estimator would be required for bagging, bagging classifier should also be used here along with a random forest. For boosting you can use ADA and gradient boosting without a specific base estimator.

- **In problem set 1, we have few outliers in only two features and the feature set is ordinal, so should we treat them.**

Yes, outliers are to be treated for all analyses.

- **Rather than doing a complete EDA code separately, can we do a panda's profile and infer the results from there in the submission report?**

Pandas profile report is one simple to get all EDA done directly. But as part of the assignment requirement, you need to give each line of code and analysis separately. So for your own analysis, you can use pandas profile report, but for assignment submission, detailed EDA has to be done.

- **Can we scale only one feature in the data set and leave remaining as it is, as they are already encoded kind of values? Is this the correct approach or not, please do a guide on this.**

Scaling as an option is done if you have continuous nature of variables in the data set with different measures, but when you want to model, data with different measures would give us incorrect answers. So in case, the data points for independent variables are binary or ordinal in nature already, you can skip the step of scaling, else the process has to be adopted.

- **Should I convert the target variable to (0,1) as this is a classification problem and I'm getting the same results for labels Labour and Conservative without converting?**

There won't be any difference in encoding your target variable. You can keep it as it is or you can encode the target variable.

- **If and when we do Scaling, we surely scale continuous variables like Salary, age, and so on 2. But what about coded Categorical variables. Especially if one variable has 2 categories and some other has 15. Should we scale these too?**

Scaling is only for the continuous variables and ordinal variables. If we are applying Min-Max scaling even binary values won't change:
Min-max = (X−Xmin)/(Xmax−Xmin)
0 will remain >> (0-0)/(1-0) = 0
1 will remain >> (1-0)/(1-0) = 1

- ***So, only 'Age' variable should be scaled? Or How about transforming it into categories of Age bracket?***

You are most welcome to try binning. Please try and see if any changes to model performance.

- **All the variables in the election dataset are categorical except age. Though they are categorical, they are varying in scales. As LDA assumes that each input variable has the same variance, do we have to scale the variables?**

LDA needs the data with the same variance but it is very difficult to make categorical data as normally distributed data. So please do not scale the categorical columns and build LDA.

- **Can you explain more about the 'Blair' and 'Hague' variables of the 1st Project? I am unable to understand what "Assessment of a leader" means here. Are voters evaluating the leaders here 1 being the least evaluation score n 5 the highest? Same for economic conditions?**

Blair is Tony Blair (Name of the person Contesting from Labour Party) and Hague is the name of the person from the conservative party. Yes. Assessment scores. Lower 1 and higher is 5.

- **Regarding model tuning for the KNN model, I am confused about which value to be used as an input parameter for weights. as if we use uniform, the results are entirely different from if we distance as weights. KNN is a distance-based algorithm, so should we consider distance as the correct option to be used as an input parameter. pls, let us know?**

Weights can take two values. Uniform and distance. The Distance input for weights is different from a metric distance (Euclidian , Manhattan). Ex : You had given k value as 5 For nearest neighbor calculation, the algorithm uses the metric distance, After identifying the 5 nearest neighbors , do you want to give the equal weights for all the neighbors or any distance based weightage for generating the predictions.

- **In Problem1: Elections, the question is like " Model Tuning, Bagging (Random Forest should be applied for Bagging) and Boosting."Model tuning means it is using SMOTE or only hyperparameters, please be specific.**

SMOTE is not a tuning algorithm. SMOTE is used to treat class imbalance. If your dataset is a class imbalanced dataset and it is affecting your performance metrics, please use SMOTE Tuning parameters is model related and will be used to tune the performance metrics. Ex : In random forest, the model tuning parameters are n_estimators and max_features.

Problem - 2

- **Do we have to remove only stopwords or should we remove punctuations and do stemming as well?**

Please do the preprocessing that include as much cleaning as you think is required. Removing punctuations and stemming will surely be helpful.

- **The number of characters is with/without spaces in between?**

With space will also suffice. But it would be good if you count it without space.

- **Find the number of characters, words, and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts() inaugural. words('1941-Roosevelt.txt') are not giving me a count of the words in this speech. Please could u give me a hint on how to frame the code for getting the count of words or count of characters or count of sentences? As this is not a CSV file I am facing this problem. Please could you help me with this?**

Read the text file separately with a different file name and try to create a new variable as a list of all text files. With this new file name, index the list as a Dataframe for all your analysis. This will then be like a data set that you have dealt with in your sessions.

- **Using "inaugural.words()" gives me a different count than when I use ".split()" method. The count arrived at from the ".split()" method matches with the physical word count from MS Word. Which one should I consider and why is there a difference? Please advise.**

inaugural. words(), takes into consideration the total count of all words, but the split functions take the spaces count also into consideration, so the count may differ. It would be suggested to go with the first option.

- **Sent() is given as a hint to use extracting sentences. But is the syntax is correct? B'coz it's not working.**

It is .sents(). .sent() is not right.

- **Remove all the stopwords from the three speeches. What is expected in the business report for this question?**

In the report, you can say, what were the stopwords that were a part of the text and after cleaning the data, you can show how the data looks like.

- **Calculating the top 3 frequently used word and word cloud, do we need to remove words like 'us', 'know', and 'let' or consider them in the count?**

As part of EDA for text mining, once all punctuations and stopwords are removed, ideally the data is ready for modeling. Ideally, you do not need to remove them and can go ahead with modeling.

- **While predicting the top 3 words if some output displays "--" shall it be considered in output as a word or we need to drop it? Also for model tuning, shall we need to present the output of the various model performance metrics in the report, or only the best parameters which we obtained after grid search CV is sufficient enough.**

"--" this is a garbage value which needs to be removed... For model tuning, all model output needs to be mentioned in the business report.

- **In the case of a word cloud, do we need to use stemming?**

Stemming is a process of pre-preparation of data for text analysis and would be required for any text modeling. Would suggest including the method before building the model.

- **Using techniques other than the techniques mentioned in the question for preprocessing is valid? Confirming that all the three speeches as mentioned in the**

questions have their own independent analysis and hence independent preprocessing right?

Yes, preprocessing techniques are valid either mentioned specifically or not. Your judgment of pre-processing would be useful to mention when you apply something you believe is better than provide a short justification note too for the same.

Yes, It is confirmed that these speeches would require independent analysis.

- **Is there any specific stop word dictionary to be used?**

No, not a specific one, if you are not sure of any then refer to the mentor session or learning material to find the stop word dictionary. NLTK package provides a useful dictionary for stop words.

- **Do we remove special characters like '--' before doing the frequency count for Top 3 words as asked in Q2.3? Otherwise, the most repeated word comes out as a special character. FYI - None of the questions has asked us to do basic pre-processing steps, only stop-word removal is asked to be done as part of Q2.2 and Q2.3**

Even though no questions have clearly mentioned doing pre-processing of data, this shall form part of data science techniques before the real analysis of the data. So would suggest applying all EDA techniques to the data and also clean all special characters before writing inferences.

- **Need Confirmation, For Q 2.3, 2.4 Do we need to create Word Cloud & Most occurring Words Frequency for each Speech Separately? OR they Need to be clubbed in one Word Cloud?**

For each President's address, one separate word cloud needs to be framed and analyzed.

- **How can we check the list of stopwords present in the word cloud Package? When we Pass text through stopwords from nltk and stopwords from word cloud, the output is different. If there are some symbols like "…", ".", "----" etc are present in the text, should we remove them or add them to the stopword list?**

Please remove all the special characters except spaces, Also, you can define your own stop words if necessary. Ex: In apple tweets, case study, there is a tweet "LOVE U APPLE". The actual stopword is you but as we know "U" means you. So you can add "U" as one of your stop words.