AIML Online

Frequently Asked Questions in Problem Statement

Course: Featurization, Model Selection and Tuning

** Direct or Self-explanatory questions are not covered in this FAQ.*

Objective: *FMT project's main objective is feature engineering and feature extraction. Finally, only or most of the useful variables must be fitted to the model. Once data is clean and we have all relevant variables, then our model will perform well. This project gives learners wisdom to explore Feature engineering steps. Clean your data as much as possible.*

2. Data cleansing:

2A. Write a for loop which will remove all the features with 20%+ Null values and impute rest with the mean of the feature. [5 Marks]

Query: Should I remove the features with 20%+ null values or remove only those rows?

You have to remove features having 20%+ null values and not the rows.

Query: I know how to remove null values but how to remove variables having 20%+ null values and how to impute the remaining variables with less than 20% null values?

Here learners are expected to check % of null values and remove those features having more than 20% null values present in them. For performing this step, you have to write a 'for' loop that calculates % of null values for that feature.

And for remaining features which have less than 20% null values, impute those features with its mean.

2B. Identify and drop the features which are having the same value for all the rows. [3 Marks]

2C. Drop other features if required using relevant functional knowledge. Clearly justify the same.

Query: I could not understand what to do in project FMT project q2b and q2c?
Can you please elaborate what I have to do in these questions?

Feature engineering consists of creation, transformation, extraction, and selection of features, also known as variables, that are most conducive to creating an accurate ML algorithm.

For Q.2.B and Q.2.C - It is expected from the learners to do all feature engineering steps and extract only those features which are good for building models.
Here you should drop the features having the same kind of information, for this you can choose to do different feature engineering steps like PCA and others like forward selection, backward elimination. You can check whether features have zero standard deviation and drop them, also can check high correlation etc.

Note: Doing PCA is not mandatory, it's just a suggestion, you can choose to do PCA on 5.D. But mention

your observations here.

In Q.2.C please justify why you are again choosing a feature engineering step to drop the features. Your statement should justify your action. Meanwhile, if you feel all your features are good and there is no need to drop any of them, then justify the same.

For Q.2 Data Cleansing--- Clean the data to the best of your knowledge and drop all highly correlated and not so useful columns. Data should be cleaned before building a model.

2E. Make all relevant modifications on the data using both functional/logical reasoning/assumptions. [2 Marks]

Query: Please elaborate or provide the hint as data cleansing is already done in above all questions related to this project.

Here list down all the modifications made to the data (2.a, 2.b, 2.c, 2.d) and your assumptions for choosing these steps in cleaning data. And What can be done further, is there any scope for PCA or any feature engineering steps. You can also express your assumptions on the cleaned data. A brief explanation is needed here.

3. Data analysis & visualisation: [5 Marks]

3A. Perform a detailed univariate Analysis with appropriate detailed comments after each analysis. [2 Marks]

3B. Perform bivariate and multivariate analysis with appropriate detailed comments after each analysis. [3 Marks]

Query: How easy is it to do Univariate, Bivariate, and Multivariate analyses, when I have more than 500+ features?

 Yes, there are huge number of variables which is way more difficult to interpret. But in real life problems you will have still more columns and to make the learners understand the concepts, this project is designed.

Since we don't have variable names here, it is difficult to understand which variable is giving us what information. So, please choose any 3 or 4 variables and perform univariate analysis. Likewise choose any two variables and perform bivariate analysis. Pair plot is a challenge here, so please avoid doing it. Once you perform a correlation plot you can mention your observations there.

For correlation plot or heat map, there is no need to specify any column name; you just have to give your overall interpretation and observations, like if you observe any correlation or not.

4D. Check if the train and test data have similar statistical characteristics when compared with original data. [2 Marks]

For this question please print 5-point summary of original data, train data and test data separately, for which you can use 'describe' function, and note down your observation like do you feel they are still same or any variations between them.

Statistical characteristics are many like Sampling and Errors, Statistical measures of the data etc. From one description function you can know about a 5 points summary like Mean, median, mode, std, Range, IQR, counts etc. These are all describing how your data is distributed, that is what statistical characteristics mean in the question.

5A. Use any Supervised Learning technique to train a model. [2 Marks]

Query: For questions 5A-5C, can we just use "raw" data (i.e. data that is not balanced or standardised)? The reason is because 5D already asks for the same. Can we build any Supervised model of our choice?

For Question 5.A to 5.C, you can continue with the same data which use used in Question 4, follow all the steps as asked in problem statement.
In 5D, it's just a hint to improvise your model performance, you are free to explore. E.g.: you can choose to do PCA.

Yes, you can build any Supervised model of your choice.

****************************HAPPY LEARNING********************************