STAT 412 - FINAL REPORT

(Deadline: June 2, 2024 - Sunday, 11:59 p.m)

STEPS IN DATA ANALYSIS PROJECT

1. ~~Clearly define the aim of the study. Introduce your data.~~
2. ~~To understand the data, define research questions. There is no lower or upper bound for the number of questions. This is up to your understanding of the data.~~
3. If you need to clean or tidy data, apply the tools that we cover in the course, or you can try different methods to have tidy and clean data. Do some feature engineering, if possible. In this step, please keep the missing data as it is and try to understand the missingness mechanism. If the missingness mechanism is MCAR or MAR, impute missing data after EDA. If it is MNAR, your job will be difficult. You may try to find previous studies, and try to handle this type of missing data by using the prior information.
4. For each of your research, choose the most appropriate and fancy numerical and/or graphical methods to get some answers. Please examine the summary statistics and interpret crucial ones. Add also a scatter plot matrix to discuss the linearity of relationships between variables. Then, start the confirmatory data analysis part after imputing missing data. In this section, you will conduct statistical tests (simple hypothesis tests, ANOVA, tests for independence of categorical by checking whether the assumptions are satisfied or not. If assumptions are not satisfied, you may conduct nonparametric tests, etc.).
5. For categorical variables, apply one-hot encoding. I recommend you use vtreat but not necessary.
6. For a regression problem, check the distribution of the response variable. If it is not normally distributed, apply the proper transformation. Then, examine the matrix scatter plot to see the linearity of the relationship between response and explanatory variables. If there are non-linear relationships, apply a transformation on explanatory variables. You may need to consider the interaction effect too. Show the statistical skills that you have learned so far.
7. If you have many variables and want to use PCA for dimension reduction, Please examine your principle components and try to understand their nature. Try to give good names/descriptions to them. After PCR, you need to interpret your model. If each component has some meaning, it will be easier for everyone to understand your findings.
8. For a prediction or a classification problem, choose your cross-validation method. Please set the seed to a specific value at the beginning so that one can get the same results that you obtain. Use the same train and test sets on different methodologies like regression, logistic, ANN, SVM, etc.
9. For the classification problem, use the same proportion of 1 – 0 in your cross-validation setup. In addition, after predicting the success probabilities you may need to choose the optimum cut-off point for your data set to calculate the model performance in case of imbalanced data. (for unbalanced classification problems, you can try some methods like SMOTE that are given in the class)
10. Apply
    - regression (multiple, logistic, Poisson, ordinal logistic etc. and interpret the model coefficients. Discuss whether they are logical or not whether you observe the similar results that you saw in your EDA step),
    - ANN,
    - SVM,

- RF,
- XGBoost (if you can, add lightgbm, catboost, stacking)

on the data (All these 5 models should be applied. You will get credit from each application).

11. Tune your parameters to get the best performance from each model. Give information about tuning process. Then, obtain the model performance scores like RMSE, MAPE or Accuracy, AUC, Sensitivity, F1 score, Kappa etc. (based on the problem) both for train and test sets. Your final decision on the best-performed model will be based on the test performance.
12. Give the variable importance that you get from the best-performed model. Make a comment on them.
13. Write a nice conclusion. What you learn from the data, what your comments to the decision makers are so on.

- **YOU CANNOT CHANGE YOUR DATASET AFTER THE INTERIM REPORT SUBMISSION!!!**

STYLE AND FORMAT INFORMATION

- Write the final project in **the IEEE conference article format** given in ODTUCLASS, including all the sections that should be in the article (such as Abstract, Introduction, etc.). Include only the most important findings. An example article format is provided.
- Your article should be **at least 5 A4 pages and at most 7 A4 pages in the given IEEE format, including graphs & tables (if any). Please write comments on each line of your R code and upload it separately into ODTUCLASS**. Note that, you should use a considerably large amount of space for your text (not figures, tables, R code, etc.). Do not just copy-paste your output; you should include interpretations, findings, etc. in words.
- **You can find one sample project at the ODTUCLASS. Please do not use the same wordings and exactly the same format. This is a sample for you to process easily.**
- Format of the tables: Table headings should be numbered (such as Table 1. Summary statistics of the data) and given on top of the table.
  Example:

Table 1. Importance of recruitment and retention strategies compared with provision of rideshare services (N= 32).

Variables	Recruitment and retention strategies		
	Mean (SD)	*t* test	*P* value
Reasons for study completion			
Rideshare service was provided[a]	5.75 (1.70)	—[b]	—
The study visits were in the evening	5.47 (1.90)	0.64	.53

- Format the figures and give the names of the figures at the end of the figure and numbered them as Figure 1…..
  Example:

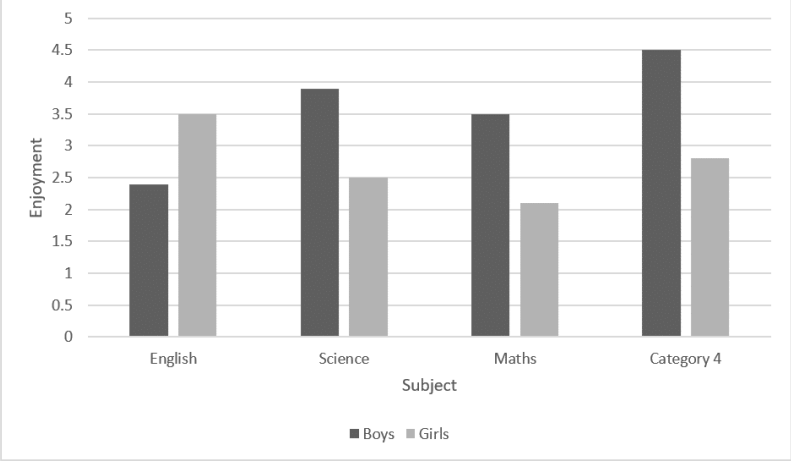

Figure 2. Boys' and girls' self-rated enjoyment of core subjects.

- Submit your report to **ODTUCLASS by June 2, 2024 - Sunday, 11:59 p.m. Do not include R codes in the TURNITIN version.**