Term project progress report

San José State University
Computer Science Department
CS156, Introduction to Artificial Intelligence, Spring 2022
Term project progress report
Name: Andy Luong
SID: 014222676

1. **AI problem and background:**
    a. The problem is the Titanic was a historical ship that sank with most of its passengers not surviving.
    b. The prediction problem is to build a predictive model that answers the question "what sorts of people were more likely to survive the Titanic sinking?"
    c. It is interesting because it will help practice applying my skills to real world problems, familiarize myself with the Kaggle platform, and launch my Kaggle journey.
    d. It is important to solve because it will help predict and give feedback on what factors are more important than others when considering health and safety precautions.
    e. There have been other previous solutions like custom pipeline, logistic regression, SVM, Random Forest, SVC, and more
    f. Other individuals attempt to arrive at a solution for this problem because this dataset is from a machine learning competition. By applying to my own, it will help Improve my score by learning more about the data, experiment, and learn from other's code and ideas
    g. I agree mostly with the approach since there are multiple ways to design/create some new features, try different preprocessing, try different types of ML models, and combine multiple models (ensemble).
2. **Dataset of choice:**
    a. The Titanic dataset consists of the training set and test set. The training set will provide the result truth for each passenger including the features. The test set compares how well the model performs when the data is not yet known. Since it doesn't have the outcome, the model being trained will help predict if the passengers are likely to survive or not survive the Titanic sinking.
    b. Source: https://www.kaggle.com/competitions/titanic/data?select=train.csv
    c. Was able to download this dataset. The process is accept the rules and join the competition, download the data (from the data tab of the competition page), understand the problem, EDA (Exploratory Data Analysis), train, tune and ensemble machine learning models, and upload my prediction as a submission on Kaggle and receive an accuracy score.
3. **Describe the independent variables**
    a. pclass is the ticket class or a proxy for socio-economic status (SES) where 1 is 1st/Upper, 2 is 2nd/Middle, and 3 is 3rd/Lower. It is numeric.

b. sex is either male or female. It is categorical.
c. age is in years where it will be fractional if less than 1 and be in .5 form if estimated. It is numeric with a chance a passenger's age is unknown.
d. sibsp is the number of siblings (brother, sister, stepbrother, stepsister) and spouses (specifically husband and wife) aboard the Titanic. It is numeric.
e. parch is the number of parents (mother and father, excluding nanny), and children (daughter, son, stepdaughter, stepson) aboard the Titanic. It is numeric.
f. ticket is the ticket number. It is actually categorical since some numeric numbers will need to have letters next to it to indicate the type of ticket.
g. fare is the passenger fare. It is numeric with decimal values.
h. cabin is the cabin number It is categorical either containing the letter(s) followed by the 0 or more numbers or is left unknown.
i. embarked is the port of embarkation where the letter C is Cherbourg Q is Queenstown, and S is Southampton. It is categorical and can be both capitalized or non capitalized and can be unknown.
j. It will train my solution model by deciding that the higher ticket class, male, and younger age have a higher chance of survival while the rest have moderate to low impact with some undeciding factors My independent variables are a mix of both numeric and categorical. The ticket class and port of embarkation will be categorical because there are 3 levels of status in both variables while the rest can still remain numerical. The category does not have a free text (NLP problem). The training data represents . There will be 9 variables (feature space) being used.

4. **Describe the dependent variable:**
    a. survival is simply 0 for no and 1 for yes.It can be binary categorical.
    b. The dependent variable is whether or not the person could survive. It is the variable in my solution model that I will be predicting. The dependent variable is categorical as it will be a binary answer of "no" and "yes" mapped to 0 or 1. Logistic regression is a classification algorithm. There will be two classes.
5. **Describe the data splitis:**
   Since there is no validation dataset, I will be using a cross-validation approach. I will stratify my split because there is a gender column and the sample results given are only for females. I will be using the logistic regression model. The hyperparameters will be passenger ID and name (First name, then comma, then Mr. or Mrs. Last Name and other names in parentheses), and because it is just used for identification and doesn't affect the survival rates. I will be using the DecisionTreeClassifier and RandomForestClassifier. Pandas, sklearn, training test split, crossing val score, gaussian, and confusion matrix will be used throughout the code. Other tools in consideration are seaborn, WordCloud, aggregation, missing values, data imputation, encoding, and dropping columns.
6. **Number of training observations:** 891 (ID 1-891, L892)
7. **Number of validation observations:** N/A
8. **Number of test observations**: 417 (ID 892-1309, K419)