**CMP 414 Mid-Term Project**

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although where was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

For the mid-term project, we will tackle the Titanic dataset on Kaggle.com. In this project, you will get to apply machine learning tools to analyze the passenger information from the training set, and try to predict which passengers in the test set survived the tragedy. You are expected to submit a complete Jupyter notebook with python code (with detailed comments), numerical results, and graphs (with proper labeling) to Blackboard.

**Stage I: Data Exploration**

1. Download the training data and test data from <https://www.kaggle.com/c/titanic/data> as .csv files

2. Load the two datasets as pandas DataFrames. Combine them into a single DataFrame. For the purpose of this stage, we will only use the combined dataset.

3. Obtain basic information of variables. For each variable, find out:

1) What does this variable represent

2) The meaning of values

3) Numerical summary

4) Graphical distributions of values

4. Prediction Power

A good model should only include variables that are indeed related to the task. Use correlation coefficients and scatter plots to decide which variables have strong relationship to passengers’ survival, and which don’t.

5. Data Cleaning

Are there missing values in each column? If so, is it possible to impute the missing data in a reasonable way? Discard all passengers with incomplete information may significantly reduce the number of data example. We should use different methods to handle the missing data according to the nature of the variable:

1) If a variable only has a tiny fraction of values, a common practice is imputing with mean or median value among passengers belonging to the same group.

2) If most values for a variable are missing, it is usually better to simply discard this feature.

Please go through all variables and handle the missing values.

5. (optional) You may think that the name feature is irrelevant. However, a closer look at the names will show that some passengers have rare titles such as Dr., Mme, or Master., which indicates their high social status. Please create a column to indicate whether a passenger has a rare title.

**Stage II: Model Construction**

1. Data preparation

1) Sex in the original dataset is categorial, and thus cannot be directly fed into Machine Learning models. Please convert the categories into 0’s and 1’s.

2) Embarked is also categorical. However, it is not ideal to simply convert the three categories into 0, 1, and 2, since it creates a bias by imposing an order to the values. Instead, we can add an indicator variable (also known as a dummy variable) for each category separately. Please replace the Embarked variable with three variables – Embark\_S, Embark\_C, and Embark\_Q – to indicate the place a passenger boarded the ship. For example, if a passenger boarded from Southamptons, Embark\_S should be set to 1, and Embark\_C and Embark\_Q should be zero. (pandas.get\_dummies() can help you with this)

3) Feature Scaling: Machine learning models usually work best on datasets whose variables have similar ranges. Please scale each variable so that it has zero mean and unit variance (sklearn.preprocessing.StandardScaler() can help you with this).

4) Because the test dataset has no information on survival, it cannot be used to evaluate the performance of the model. As a result, we need to further split the data in train.csv into training set (80%) and validation set (20%). We will use training set to build the models, and use validation set to evaluate them. Split the dataset into training set, validation set, and test set. (sklearn.model\_selection.train\_test\_split() can help you with this)

5) (Optional: Feature Engineering) Previously we noticed that some passengers’ names have rare titles, which may indicate their unusual social status. Extracting a new feature from existing data is called feature engineering. Another interesting thing you may observed is that some passenger got on board for free! This could also be a potential indicator of the social status of the passenger. Please create an indicator variable on whether a person paid for the ticket.

2. Building models

So far, we have learned five models: linear regression, polynomial regression, logistic regression, and k-nearest neighbors method, and support vector machine. Choose models that are suitable for this problem and use corresponding sklearn classes to fit the data. Note that each model class has several parameters, you are welcome to change the default parameter values and try to improve the performance of the model.

**Stage III: Model Selection and Interpretation**

1. Model Selection. For each model you have trained, perform the following evaluations:

1) cross validation

2) confusion matrix, precision, and recall on the validation set

3) ROC curve and show AUC (area under curve)

Based on these results, select a model that has the best performance.

2. Model Interpretation.

Based on the model that you have selected, can you identify a few key variables that are important in predicting the survival of the passengers? Try to find out which characteristics can improve passengers’ chance of survival.

3. (optional) Submit Results to Kaggle.com

Make predictions on the test set, submit your results to Kaggle.com.