

## Assignment 3

### **Data preparation, exploration, visualization**

Training data and test datasets are loaded. There are 12 columns in the training dataset.

Then I used isnull function to review how many values are missing in both datasets. There are a lot of null entries in age, and way too many in Cabin column (687/891 for training set). For that reason, the age data will need to be imputed later to fillna, and the Cabin column will be dropped for training.

I also grouped the data by gender and calculate mean survival, it is obvious that gender has large impact on whether one survived. Two histograms were made to display number of people in each Pclass and their age distribution as well as survival status (color coded). Three line-plots were also generated to display survival numbers for different passenger classes and embarked places. These plots show that there is a correlation between passenger class and survival, and embarked places and age are also related to survival.

A few unrelated or severely incomplete features were dropped: 'Ticket', 'Cabin', 'Name', 'PassengerId'. The Sex feature is transformed to numerical data by mapping Female:0 and Male:1. The null values in Age column are replaced by mean age and the empty Embarked entries are filled using the mode. The embarked feature is also mapped to integers, 'S': 0, 'C': 1, 'Q':2.

Finally, I used standard scaler to transform both training and testing datasets.

### **Review research design and modeling methods**

The cleaned and transformed training dataset is used to train two different classification models:

Logistic regression and Naïve Bayes classification. Sklearn module is imported and predictions were made using the trained models for the test dataset.

## **Review results, evaluate models**

The cross-validation is applied, with 5-fold, and the scoring method is “roc\_auc” – area under the ROC curve. The mean score is calculated for both models.

The logistic regression model's cross-validation mean AUC is 0.851 and the Bayes classification yield AUC 0.831, lower than the regression results.

I also evaluated the coefficients in the logistic regression model for each feature. The coefficients show Pclass, Gender and Age are most important factors affecting survival. Gender is the most important factor.

## **Implementation and programming as evidenced by Kaggle submission**

The code is in the appendix. Two models were submitted to Kaggle and both yield score over 0.75, meaning that over 75% of the survivals were correctly predicted.


## **Exposition, problem description, and management recommendations**

Regarding the management problem, imagine that you are providing evidence regarding characteristics associated with survival on this ill-fated voyage to a historian writing a book. Which of the two modeling methods would you recommend and why?

I will recommend the logistic regression model, as it is straightforward to understand, and yield better prediction accuracies than the Bayes model. The coefficients can be directly associated with each evaluated factor (scaled) - the characteristics with larger abs(coefficient) have larger impact on chance of survival.

## Appendix


### KAGGLE:


145...	Jordan Zhang #2		0.76555	2	now
--------	-----------------	---	---------	---	-----

**Your Best Entry** ↑

You advanced 1,543 places on the leaderboard!

Your submission scored 0.76555, which is an improvement of your previous score of 0.75119. Great job!

 Tweet this!

160...	Jordan Zhang #2		0.75119	1	~10s
--------	-----------------	---	---------	---	------

**Your First Entry** ↑

Welcome to the leaderboard!

Your score represents your submission's accuracy. For example, a score of 0.7 in this competition indicates you predicted Titanic survival correctly for 70% of people.

What next? You've got a few options:

- 👉 Learn skills that can improve your score in [our Intro to Machine Learning course by Dan Becker](#).
- 🔍 Check out [the discussion forum](#) to find lots of tutorials and insights from other competitors.
- 🏆 Find a new challenge by entering one of our [open, active competitions](#) or searching our [public datasets](#).