DS-SF-34
5/25/17

Titanic Dataset - Classification Lab

Instructions: Today you are going to split up into teams to compete for top accuracy on a real-world dataset. To excel in this challenge, you will need to integrate many skills: pandas, EDA, feature engineering, model selection, regularization, and model validation.

You will be provided two CSV files: “titanic-train.csv” and “titanic-test.csv”. Use the training set for EDA, feature-engineering, and to build and validate your model. When your team is ready to make a submission, use your trained model to make predictions on the test set. Send your predictions to me on Slack, along with one notebook per group. **Your final predictions on the “titanic-test.csv” are how your team will be assessed.**

The Dataset: Real data on the Titanic’s passengers. This is a classification problem; The binary response you are to predict is survival (yes or no).

Variable	**Definition**
survival	Survival
pclass	Ticket class
sex	Sex
Age	Age in years
sibsp	# of siblings / spouses aboard the Titanic
parch	# of parents / children aboard the Titanic
ticket	Ticket number
fare	Passenger fare
cabin	Cabin number
embarked	Port of Embarkation

Guiding questions:

- What would be valuable in an exploratory analysis?
- How should null values be handled?
    - I will count any dropped rows in the test submission as incorrectly classified.
- KNN or Logistic Regression?
- Is there any implicit information in the dataset that you can use to engineer new features?
    - Example: Creating a new binary variable “man_lowerclass” for passengers who were male and lower class. This generated feature might allow the model to cleanly separate passengers who died from those who survived.
    - **IF YOU ARE GENERATING FEATURES ON THE TRAINING SET, ENSURE THE PROCESS IS EXACTLY REPRODUCIBLE FOR THE FINAL TEST SET**
- What tools do you have to balance bias and variance?
- How do you test for overfitting? What do you do to prevent it?

Teams:

- Team 1
    - Sarah
    - Laura
    - Grady
- Team 2
    - Nikhil
    - Peter
    - ~~Eunice~~ (absent)
    - Zen
- Team 3
    - Carol
    - Matthew
    - Jason
    - Zack
- Team 4
    - Fidel
    - Alana
    - Ji
    - Simon