**Kaggle Training – Titanic**

**Portland Data Science Group  
Created by Andrew Ferlitsch  
July 2017**

****

**Session I**

* Introduction What is Kaggle.
  + Show Kaggle’s YouTube video (4mins) - https://www.youtube.com/watch?v=PoD84TVdD-4
  + Show YouTube video on how to submit from Data Science Dojo –

<https://www.youtube.com/watch?v=68l47Zu4Yxg>

* + Sign into Kaggle using Google+ account (or Facebook)
* Discussion on why you can’t predict 100% accuracy.
  + Ex. Not all men died, not all women and children lived.
  + Ex. Not all 3rd class died, not all 1st class lived.
  + Some fields are missing (blank) for some passengers (e.g., age, class, etc).
* How to install Python and Jupyter
  + Download installer .exe - <https://www.python.org/downloads/>
  + Install Jupyter using pip3 - <http://jupyter.org/install.html>
  + pip3 install --upgrade pip
  + pip3 install jupyter
  + from cmd line: jupyter notebook
* Review Dataset
  + View Data: https://www.kaggle.com/c/titanic/data
  + Highlight types of features
  + Point out the ‘label’ (that’s what you are predicting).
  + Point out where data is missing.
* Do everybody dies submission.
  + Show submission process and ranking.
* (Brief) Introduction to Preparing a Dataset
  + Split into training and test data.
    - ML Presentation on Splitting Datasets - [https://github.com/andrewferlitsch/Portland-Data-Science-Group/blob/master/Presentations/ML Regression - Splitting Datasets.ppt](https://github.com/andrewferlitsch/Portland-Data-Science-Group/blob/master/Presentations/ML%20Regression%20-%20Splitting%20Datasets.ppt)
  + Methods for handling blank (missing) entries.
    - ML Presentation on Handling Missing Data - <https://github.com/andrewferlitsch/Portland-Data-Science-Group/blob/master/Presentations/ML%20Dataset%20Preparation.pptx>
    - Ex. Mean average for real-values.
    - Ex. Random selection for categorical.
    - Ex. Add a new value ‘Unknown’ for categorical.
  + Dummy variable conversion for categorical variables
    - ML Presentation on Categorical Variables
  + Feature Scaling for real variables.
    - ML Presentation on Feature Scaling
* Introduction and Explanation of Approaches covered.
  + (Classic) Linear Regression
    - ML Presentation on Simple Linear Regression - <https://github.com/andrewferlitsch/Portland-Data-Science-Group/blob/master/Presentations/ML%20Dataset%20Preparation.pptx>
    - ML Presentation on Multiple Linear Regression - <https://github.com/andrewferlitsch/Portland-Data-Science-Group/blob/master/Presentations/ML%20Multiple%20Linear%20Regression.pptx>
  + (Classic) Tree Classification
    - ML Presentation on Tree Classification
  + Ensemble: Random Forest.
    - ML Presentation on Ensemble Methods
* Introduction to Predicting
  + How test data is used to measure accuracy.
  + Accuracy Measurement.
    - ML Presentation on Measuring Accuracy - <https://github.com/andrewferlitsch/Portland-Data-Science-Group/blob/master/Presentations/ML%20Multiple%20Linear%20Regression.pptx>
  + Confusion matrix.
    - ML Presentation on Confusion Matrix
* First Example Prediction.
  + Prepare dataset by simply removing passengers with missing data.
  + Do dummy variable conversion
  + Do feature scaling
  + Do simple linear regression.
  + Show Accuracy and Confusion Matrix.
  + Do submission.

**Session II**

* Recap on What is Kaggle (for newcomers)
  + Show YouTube video on how to submit.
* Recap Review of Dataset
  + Highlight types of features
  + Point out the ‘label’ (that’s what you are predicting).
  + Point out where data is missing.
* Review two submissions
  + Everybody Dies
  + Linear Regression
* Recap Introduction to Preparing a Dataset
  + Split into training and test data.
  + Methods for handling blank (missing) entries.
    - Ex. Mean average for real-values.
    - Ex. Random selection for categorical.
    - Ex. Add a new value ‘Unknown’ for categorical.
  + Dummy variable conversion for categorical variables
  + Feature Scaling for real variables.
* Recap explanation of linear regression
* Second Example Prediction (Linear Regression).
  + Prepare dataset
    - mean average for missing real-value fields.
    - unknown class value for missing categorical fields.
  + Do dummy variable conversion
  + Do feature scaling
  + Do linear regression.
  + Show Accuracy and Confusion Matrix.
  + Do submission.
* Recap explanation of tree classification
* Third Example Prediction (Tree Classification).
  + Prepare dataset
    - mean average for missing real-value fields.
    - unknown class value for missing categorical fields.
  + Do dummy variable conversion
  + Do feature scaling
  + Do tree classification.
  + Show Accuracy and Confusion Matrix.
  + Do submission.
* Group Discussion and Suggestion of other methods for handling missing fields
  + Try group suggestion.
  + Make prediction/suggestion (4th) with Linear Regression
  + Make prediction/suggestion (5th) with Tree Classification.

**Session III**

* Recap Session II submissions.
* Recap Explanation of Random Forrest
  + K-Fold Sampling
  + Majority Voting
* Sixth Example Prediction (Random Forrest).
  + Prepare dataset
    - Use Group suggestion for handling missing data.
  + Do dummy variable conversion
  + Do feature scaling
  + Do simple linear regression.
  + Show Accuracy and Confusion Matrix.
  + Do submission.
* Group Discussion on Alternative/Creative Approaches.
  + Make List.
  + Feedback.
* Select another example Kaggle dataset.
  + TBD based on dataset.