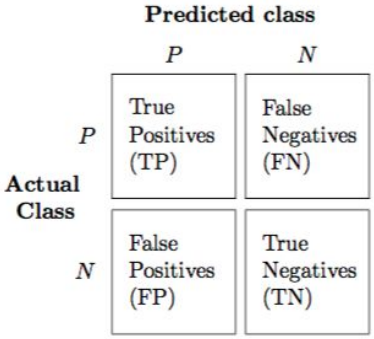
Milestone 2 - ML Algorithms & Evaluation

Team ID: 21

Team Members: Phuong Pham, Jing Lin, Shang-Yun (Maggie) Wu

- ML Algorithms
  - Random Forest
  - Decision Tree
  - KNN
  - Support Vector Machines
  - Logistic Regression
  - Linear SVC
  - Perceptron
  - Stochastic Gradient Descent
  - Naive Bayes
  - Neural Network [Implement from scratch via Keras]
- Model Selection
  - Cross validation
  - Regularization
- Metrics
  - Confusion Matrix



  - Main Metrics
    - Accuracy [Overall performance of model]

$$(TP + TN) / (TP + TN + FP + FN)$$

    - Precision [Accuracy on positive prediction]

$$TP / (TP + FP)$$

    - Recall Sensitivity [Coverage of positive samples]

$$TP / (TP + FN)$$

    - Specificity [Coverage of negative samples]

$$TN / (TN + FP)$$

    - F1 Score [Hybrid metric]

$$2TP / (2TP + FP + FN)$$

  - ROC [Recall sensitivity vs. Specificity for various threshold]
  - AUC [Area under ROC]

- Pre-processing Data
    - Convert time/date-related feature [range from 2009 - 2014] to year feature & month feature & duration feature
    - Sentiment analysis on project description & name
    - Natural language processing to find correlation between words used & final result
- Additional Note on Data
    - The data is pretty much complete, and the only missing pieces are some description data and name data which will not be a problem since our group might or might not use them when we train our data.