**Case Study of the Titanic Data Set:**

**A Classification Problem Statement using Stratified Cross-Validation**

dataset:

https://s3.amazonaws.com/acadgildsite/wordpress\_images/datasets/titanic/test.csv

https://s3.amazonaws.com/acadgildsite/wordpress\_images/datasets/titanic/train.csv

1. **Prepare Datasets**
2. read the data sets
3. 'X' data frame will store train data
4. 'y' series will store the 'Survived' data
5. 'Z' will store test data
6. Check for missing values in 'Age', 'Fare', 'Embarked' and fill them with the median value; do it for both test and train sets
7. Perform dummy treatment on 'Pclass', 'Sex', 'SibSp', 'Parch', 'Embarked' for both test and train sets and drop the original columns
8. Check data structure

## Feature Engineering

1. Perform Binning for Age and Fare
2. Perform median groups and drop the original columns
3. Cabin usage: add feature 'has\_cabin' which will yield cell values of 'false' or 'true'
4. Extract the Titles of passengers (use lambda function) and do dummy treatment
5. Concatenate the dummies with the parent data frame

## Decision Tree

1. Set stratified cross validation parameters
2. Set the tree parameters
3. Set the tree grid
4. print out the best tree parameters
5. print out the best tree cross validation score

## Logit

1. Set the logit parameters
2. perform the grid search algo
3. print out the best tuning parameter
4. Predict on the test data set