**Project Flow**

**Item1: Data Visualization**

* Plot attributes against Survived results, determine which features are likely to be influential
* For categorical features, compare the probability of survival between different values

**Item2: Data Preprocessing**

* Fill in ‘Cabin’ attribute by “Y/N” according to whether the value is null.
* Apply Random Forest Regressor to fill in missing ‘Age’s
* Feature factorization of categorical features: ‘Cabin’, ‘Embarked’, ‘Sex’, ‘Pclass’
* Feature scaling of ‘Age’ and ‘Fare’, restrict the range into [-1, 1]

**Item3: Model Training and Prediction**

* Use Logistic Regression Model to fit data
* Apply the same preprocessing to test data, predict based on the model

**Item4: Model Analysis**

* Draw the learning curve, ensure there is no overfitting
* See the coefficient of each attribute, determine which ones are significant in the model
* Apply Cross Validation to help feature engineering
* Split data into train and cv data (70% and 30%, respectively)
* Use train data to fit the model
* Run prediction on the cv data
* Compare prediction and original ‘answer’ of the cv data, try to figure characteristics that can be analyzed further

**Item5: Model Optimization**

* Adding extra features
* Combine Sex and Pclass together as a new feature
* Categorize age into different groups(Young, Adult, Senior)
* Other potential features: Family Size, Title, etc.
* Model ensemble: Bagging
* Using logistic regressor only, but each time use only a subset of train data to fit
* Use sklearn.BaggingRegressor function, combine the results by voting