Date

FINAL PROJECT PROPOSAL

1. TITLE:

Titanic: Machine Learning from Disaster.

2. DESCRIPTION:

The sinking of Titanic is one of the most infamous shipwrecks in history. During Titanic's maid voyage on April 15,1912, she sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crews. This tragedy shocked the international community and led to better safety regulations for ships.

One of the obvious reasons for this shipwreck was that there were not enough lifeboats. Although there was some element of luck involved surviving the sinking, some groups of people were more likely to survive than others, such as women, children and the upper-class.

Our project aims to analysis of what sorts of people were likely to survive. In particular, we are planning to apply the tools of machine learning to predict which passengers survived the tragedy.

3. METHOD:

a. Learning Plan:

We plan to learn the training set using **Logistic regression**, **SVM**, **K-neighbors**, **Decision tree**, **Random forest**, **Gradient Boosting Decision Tree & xgbGBDT model**. And then ensemble the better ones to do prediction on the test set to try **Ensemble Learning.**

b. Reference:

https://www.kaggle.com/c/titanic#tutorials

4. DATA:

The data has been split into two groups: training set/test set.

The training dataset has 891 entries, 12 columns, and there are some data missing.

The test set has 418 entries, 12 columns, also containing missing data.

Data Dictionary:

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st,2 = 2nd, 3 = 3rd
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	C = Cherbourg, Q = Queenstown, S = Southampton

5. TEAM COOPERATION

Yusi Liu : Data set analysis & data scrubbing.

Feature extraction & try LR/SVM/K-neighbors/DT model

Familiar and try XGBoost Algorithm & Ensembling/Stacking.

Yunhao Wu : Try Random forest/GBDT/XGBoot-GBDT & Ensembling model.

Fine tuning and prepare project report.