Team YSI

Titanic: Machine Learning from Disaster

Development plan proposal

# General Info

**Train data**

﻿PassengerId 1

Survived 0

Pclass 3

Name Braund, Mr. Owen Harris

Sex male

Age 22

SibSp 1

Parch 0

Ticket A/5 21171

Fare 7.25

Cabin NaN

Embarked S

Number of rows: **891**

**Test data**

PassengerId 892

Pclass 3

Name Kelly, Mr. James

Sex male

Age 34.5

SibSp 0

Parch 0

Ticket 330911

Fare 7.8292

Cabin NaN

Embarked Q

Number of rows: **418**

# Data Preparation

1. PassengerId – unique ID, no missing values, not used as feature.
2. Survived – integer number, 1 – survived, 0 – not survived, use as target.
3. Pclass – integer number of the class, no missing values, use as feature.
4. Name – string, no missing values, extract title and family, might check ethnicity.
5. Sex – string, no missing values, male or female, transform to 0, 1
6. Age – integer number, lots of missing values, use regression to determine missing age.
7. SibSp – integer number, some missing values, use to determine family.
8. Parch – integer number, some missing values, use to determine family.
9. Ticket – string, number of the ticket, some missing values, might be useful – check Titanic history.
10. Fare – double, price of the ticket/trip, some missing values, useful to determine Embarked.
11. Cabin – string, number of the cabin, lots of missing values, might be useful – check Titanic cabin map.
12. Embarked – char, letter of the city where the passenger embarked, lots of missing values.

# Create Model

## Random Forest

# Test model

## Adjust features

## Check accuracy

## Submit to kaggle.com

# Phases

## I Phase – Initial submit

The idea is that we should start with some primary model and submit it to the kaggle.com. It should not be perfect one, but also it should have some bits of data preparation, but not a lot. In this way, we’ll ensure that there is something submitted on the site and also use it as comparison when we are doing the next versions.

### Data Preparation

Features used as they are:

* Pclass – 1,2,3

Newly generated features (formatted existing features):

- TitleId – extract the title from the Name and form groups with ID numbers.

- GenderId – 0 – male, 1 – female

- Age – calculate the average age per gender and apply it for missing values.

- SibSp – fill the missing values with 0.

- Parch – fill the missing values with 0.

- Fare – calculate average price based on Pclass and Embarked.

- Cabin – extract just the first letter and form IDs. Put 0 for missing.

- Embarked – transform the letters to IDs. For the missing ones, use Fare and Pclass to determine possible embarked place.

### Modelling

* Execute the Random forest method.
* Find the accuracy – cross validation. If very low try to do some tweeks with the data.