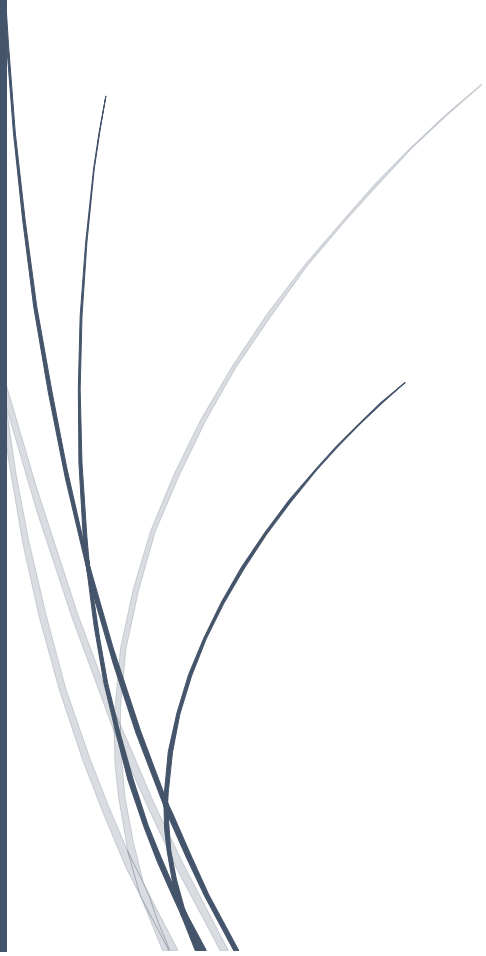
3/6/2020

Titanic: Predict survival in Crisis

a model that predicts which passengers survived the Titanic shipwreck.



Part 1: Problem Description

We are going to explore the competition **Titanic: Machine Learning from Disaster**

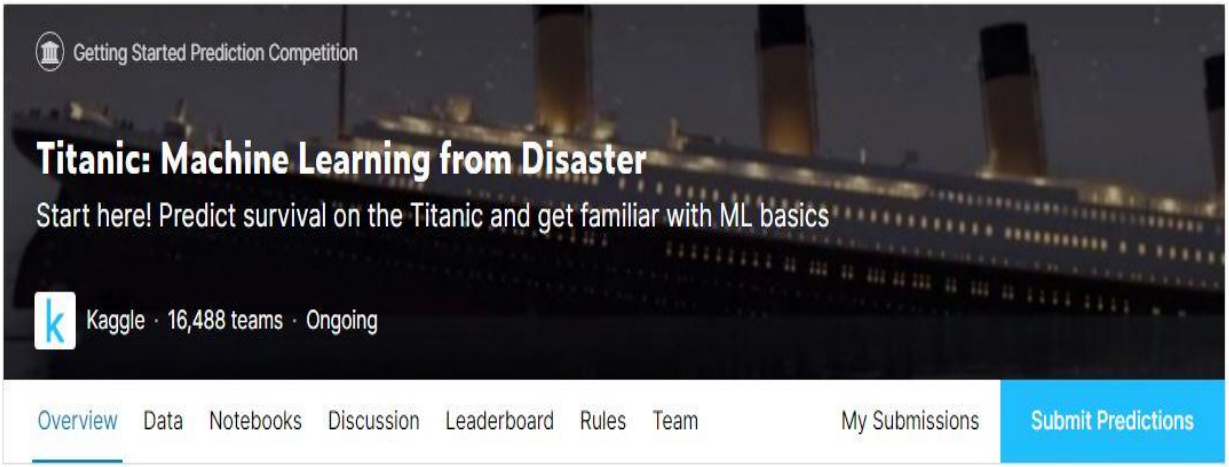
The task is to predict whether a given passenger survived the sinking of the Titanic based on various attributes including age, location of the passenger's cabin on the ship, family members, the fare they paid, and other information. Solutions are evaluated by comparing the percentage of correct answers on a test dataset.

we have gain access to two similar datasets that include passenger information ['PassengerId' 'Survived' 'Pclass' 'Name' 'Sex' 'Age' 'SibSp' 'Parch','Ticket' 'Fare' 'Cabin' 'Embarked']

Train.csv contain the details of a subset of the passengers on board (891 to be exact) and importantly, will reveal whether they survived or not, also known as the “ground truth”.

The `test.csv` dataset contains similar information but does not disclose the “ground truth” for each passenger

Using the patterns, we will find in the train.csv data, predict whether the other 418 passengers on board (found in test.csv) survived.



Part 2: Analysis Approach

a-exploring the data

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

b-Wrangle, prepare and cleanse the data.

Correcting by dropping features

Based on my assumptions and decisions I will drop the Cabin and Ticket features.

Creating new feature extracting from existing

I want to analyze if Name feature can be engineered to extract titles and test correlation between titles and survival, before dropping Name and PassengerId features.

When we plot Title, Age, and Survived, we note the following observations.

Most titles band Age groups accurately. For example: Master title has Age mean of 5 years.

Survival among Title Age bands varies slightly.

Certain titles mostly survived (Mme, Lady, Sir) or did not (Don, Rev, Jonkheer).

➔ So I decide to retain the new Title feature for model training.

Converting a categorical feature

I converted Sex feature to a new feature called Gender where female=1 and male=0.

Create new feature combining existing features

I created a new feature for FamilySize which combines Parch and SibSp. This will enable us to drop Parch and SibSp from our datasets.

Completing a categorical feature

Embarked feature takes S, Q, C values based on port of embarkation. Our training dataset has two missing values. I fill these with the most common occurance.

Quick completing and converting a numeric feature

I now complete the Fare feature for single missing value in test dataset using mode to get the value that occurs most frequently for this feature. And round off the fare to two decimals as it represents currency.

Part 3: Initial Solution

After analyzing the data I decided to use Logistic Regression is a useful model to run early in the workflow. Logistic regression measures the relationship between the categorical dependent variable (feature) and one or more independent variables (features) by estimating probabilities using a logistic function, which is the cumulative logistic distribution.

Part 4:Initial solution analysis

- ➔ the confidence score generated by the model based on our training dataset using Logistic regression was 80.36

```
Out[40] 80.36
```

- ➔ We can use Logistic Regression to validate our assumptions and decisions for feature creating and completing goals. This can be done by calculating the coefficient of the features in the decision function. Positive coefficients increase the log-odds of the response (and thus increase the probability), and negative coefficients decrease the log-odds of the response (and thus decrease the probability).
- ➔ **Sex is highest positivie coefficient**, implying as the Sex value increases (male: 0 to female: 1), the probability of Survived=1 increases the most.
- ➔ Inversely as Pclass increases, probability of Survived=1 decreases the most. This way Age*Class is a good artificial feature to model as it has second highest negative correlation with Survived.
  So is Title as second highest positive correlation.

Out[41]:

	Feature	Correlation
1	Sex	2.201619
5	Title	0.397888
2	Age	0.287011
4	Embarked	0.261473
6	IsAlone	0.126553
3	Fare	-0.086655
7	Age*Class	-0.311069
0	Pclass	-0.750700

Part 5: Revised Solution and Analysis

This model uses a decision tree as a predictive model which maps features (tree branches) to conclusions about the target value (tree leaves). Tree models where the target variable can take a finite set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

- ➔ The model confidence score is 86.76%

The next model Random Forests. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees (n_estimators=100) at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

- ➔ The model confidence score is 86.76%
- ➔ The model confidence score is the highest among models evaluated so far. We decide to use this model's output (Y_pred) for creating our competition submission of results.

.While both Decision Tree and Random Forest score the same, **I choose to use Random Forest** as they correct **for decision trees' habit of overfitting** to their training set.