Titanic: Machine Learning from Disaster
EECS 510

Xiaoyang Tan
Zhaoyang Liu

June 5, 2014

1 Current work

1.1 Defining the problem

Information is given on a training set of passengers of the Titanic, for which the survival outcome is known. Given the training set information, the goal remains to predict each passenger?s survival outcome from a test set of passengers. The details of the challenge are given on the Kaggle site.

We will apply machine learning tools to solve the problem. In this case, we may consider approaches based on random forests. And will use Python package to plot the result. The python package used includes:

- NumPy
- Pandas
- SciKit-Learn
- SciPy
- StatsModels
- Patsy
- Matplotlib

1.2 Analyze the data

In the training dataset, there are 891 passengers. Each passenger has 12 attributes. Except the "passangerId" and "Survived" attribute, we need to consider 10 other attributes and predict the survival possibility.

1.2.1 Take care of missing values

The features *Ticket* and *Cabin* have many missing values and so can?t add much value to our analysis. To handle this we will drop them from the data frame to preserve the integrity of our dataset. What's more, we will remove NaN values from every remaining column. Using **drop()** and **dropna()** function in could easily achieve goal. Now we have a clean and tidy dataset that is ready for analysis, we cut the dataset from 891 to 712, and get 8 effective attributes to do prediction.

1.2.2 Graphical view of data

The point of this competition is to predict if an individual will survive based on the features in the data like:

- Traveling Class (called pclass in the data)
- Sex
- Age

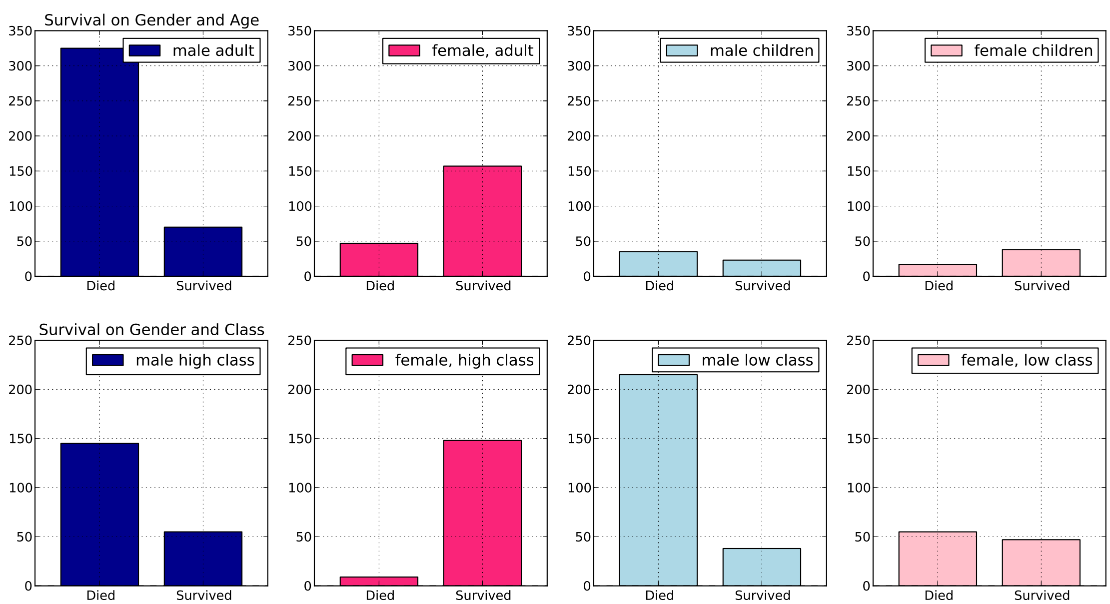

Figure 1: Survival on gender, age and class

Figure 1 shows the distribution of survival on gender age and class. ******

1.3 Simple trying

In the film, they tried to rescue women and children first, so a good guess would be on gender. From the graphical view of the data, its clear that although more men died and survived in raw value counts, females had a greater survival rate proportionally(25%), than men (20%).

After applying that all the female would survive, the survival rate become 36.36%, which comes close to the original rate 38.38%. However, we can refine our results by considering other variables.

We ruled out the age factor, as the survival rate doesn't change much with or without the involvement of age. Thus we split the fare class into four payments range and assumes that any group with more than half survivors will always be modeled to survive, while the rest will be mapped to death. The result is different from the previous one in that women in 3rd class who paid more than $20 will not survive, which bring the survival rate closer to training set.

1.4 Supervised machine learning

In paticular, we are going to try **Logistic Regression**. In statistics, logistic regression or logit regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable. We use python package **statsmodels** to build the model and use predictor variable including "Plass", "Sex", "Embarked", "Age", "SibSp".

First, we define our formula for our Logit regression. In the next cell we create a regression friendly dataframe that sets up boolean values for the categorical variables in our formula and lets our regression model know the types of inputs we're giving it. The model is then instantiated and fitted before a summary of the model's performance is printed. In the last cell we graphically compare the predictions of our model to the actual values we are trying to predict, as well as the residual errors from our model to check for any structure we may have missed. The result as Table 1 shows.

Table 1: Logit Regression Results

	coef	std err	z	P >z	[95.0% Conf.	Int.]
C(Pclass)[T.2]	-1.2673	0.299	-4.245	0.000	-1.852	-0.682
C(Pclass)[T.3]	-2.4966	0.296	-8.422	0.000	-3.078	-1.916
C(Sex)[T.male]	-2.6239	0.218	-12.060	0.000	-3.050	-2.197
C(Embarked)[T.Q]	-0.8351	0.597	-1.398	0.162	-2.006	0.335
C(Embarked)[T.S]	-0.4254	0.271	-1.572	0.116	-0.956	0.105
Age	-0.0436	0.008	-5.264	0.000	-0.060	-0.027
SibSp	-0.3697	0.123	-3.004	0.003	-0.611	-0.129

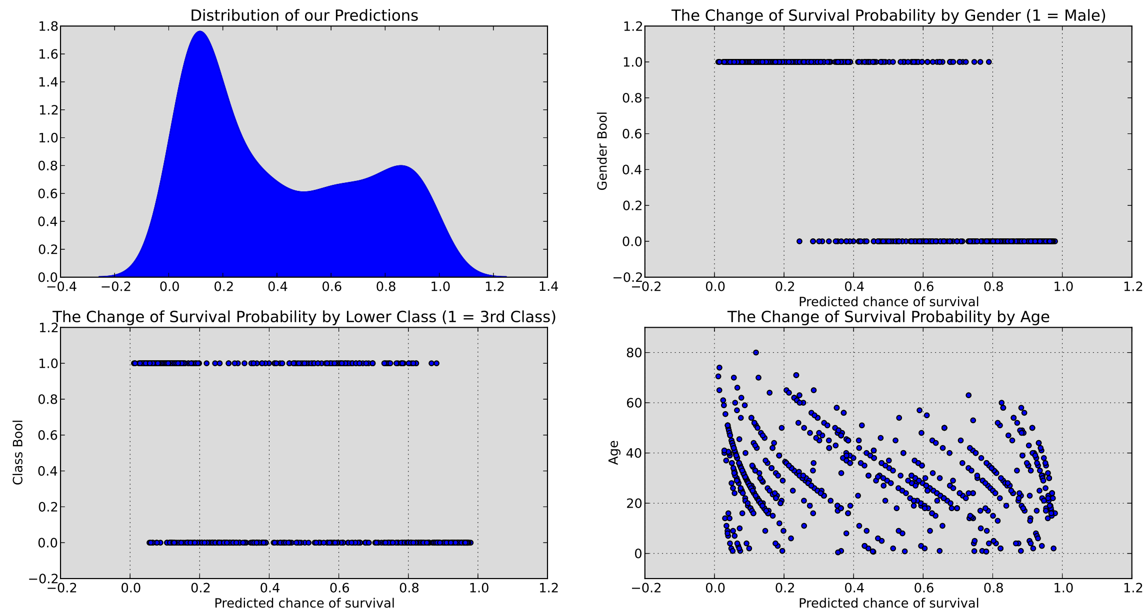

Figure 2: Prediction from Logit Regression

The Figure 2 shows the distribution of our predictions. The second graph also indicates that male has a higher possibility of died while female prefer to survive. This is consistent to our previous observations. Also, We can see from the third graph that ship class has slight effect on the prediction result.

After applying the model on the test data, we get an output file. Submitting this file shows that we got 77.03%.

2 Next step

Next step is too try use support vector machine and random forest methods to make prediction.

2.1 SVM

Logit model we just implemented was great in that it showed exactly where to draw our decision boundary. A linear line is okay, but we can do better with a more complex decision boundary like a wave, circle, or maybe some sort of strange polygon would describe the variance observed in our sample better than a line.

2.2 Random Forest

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. This technique is a form of non-parametric modeling that does away with all those equations we created above, and uses raw computing power and a clever statistical observation to tease the structure out of the data.

3 Possible results

SVM and Random Forest are supposed to perform better than Logit model, maybe could get an accuracy more than 0.8