Introduction:

Titanic was a disaster which struck during 1912 from England to New York. There were lost of people who survived and the challenege is to find out based on the data given in trai.csv to build a Machine Learning Model to predict whether a person would hav esurvied or died during the Titanic Crash. The data is provided.

Background: The data is providedby Kaggle team and explaination of each of the field is provided. Please see below for more details on the data.

VARIABLE DESCRIPTIONS: survival Survival (0 = No; 1 = Yes) pclass Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd) name Name sex Sex age Age sibsp Number of Siblings/Spouses Aboard parch Number of Parents/Children Aboard ticket Ticket Number fare Passenger Fare cabin Cabin embarked Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton) SPECIAL NOTES: Pclass is a proxy for socio-economic status (SES) 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower Age is in Years; Fractional if Age less than One (1) If the Age is Estimated, it is in the form xx.5 With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored. The following are the definitions used for sibsp and parch. Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored) Parent: Mother or Father of Passenger Aboard Titanic Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbors in a village, however, the definitions do not support such relations.

Data Analysis: This is one of the most important part of the Building Machine leanring models. This provides basic information about the data and data distribution as well as it provides whether there are some data missing.

Cabin data is not missing but it is blank and I will not use this column for further analysis. Age data is missing and around 177 rows are missing. I wont be ignoring these rows but I will try to fill the data with the Median age. Then there is ticket information and it doesn’t seems to be that much important and I will not be looking into it deeply. I had a look at the data by slicing and dicing and by using the table command in R to see if I could see some pattern in the data especially which of the variable seems to have some sort of relationship with respect to the survival. Here Age and Class are one

Data Visualization: There are various ways of analysing the data and Data Visualization is one of the most import way to find the relationships and patterns. I have plotted lots of graphs and I will be presenting only the graphs which are most important. As mentioned above, PClass, Sex and Age were most important factors and Below are the graphs for these columns for survival and non survival.

There are couple of takeawys here.

1. If you were female your chances of Survival were higher.
2. If you were travelling in 1st class then your chances were higher.
3. If your age group was less than 20 or more than 30 -50 than chances of survival are higher.

If you plot some of other graphs e.g. Fare and You will see that higher the Fare more chances of your survival are. This is valid case but it has been covered by PClass variable and thus we do not need to keep this column. The graph for this is shown below.

Similarly, you will see that there are very high chances that if you embarked from Q your survival chances are very less but it is contributed due to the fact that most of these people were in 3rd Class and that is already explained by PClass role. Similarly, there are chances that some of the Pclass data is explained.

These graphs were mailnyl created in R. However, I have generated couple of these graphs in D3. Code for this can be found in repository.

Model:

Now I created a model and I wanted to create the model based on the Age, PClass,Sex, Sibsp and parch columns only.

I will be using a random Forest to achive the same

Resuts: I will use 70% of data as train data and 30% data as test data.

Conclusion:

Feature selection and finding pattern are very important tasks and these taks can be accomplished either by using vizualition( for smaller data set and with limited number of varibales) or by using some sort of Machine Learning Algorutthm to find out which are important features based on the data set. Then you can vzualise only say top 10 featurs and see hows things looks like. The vizaulation makes it easier to present and it is very important part of the final model.

References: