Project Report : Survival in Titanic

Shashank Agarwal **Afsar Equebal** **Sindhu Somasundaram**

Abstract

There are multiple classification algorithms available. We wanted to pick a few of them and compare them to how each algorithms performs for a particular dataset. We chose the passengers in the titanic ship dataset. And we will be predicting the survival of the passengers. By using this Titanic passengers dataset, we will be training multiple classifiers and find the best ones by plotting the ROC curves.

1 Dataset and Data processing

1.1 Dataset

We obtained the Titanic dataset from the Kaggle link listed below.

`https://www.kaggle.com/c/titanic/data`

Titanic dataset consists of 11 input features and one output label. The features are passengerid, name of the passenger, ticket class, sex, age, siblings or spouse on board, parents or children on board, ticket number, passenger fare, cabin number and port of embarkation. The output label is survival rate of the passenger which is a binary value 0 0r 1. There are 891 entries in train.csv.

1.2 Data observations

We did the following visualizations to see how each feature is significant to the target label, and how can we group or combine each other. (see appendix C figures to see the visualizations charts)

1. Female passengers survived more than males.
2. Number of siblings and parents travelled with the passenger does not seem to have any correlations. So 'SibSp' and 'Parch' seems insignificant.
3. A lot of young people have survived in total.
4. Upper class people survived a lot.
5. Children and young people have survived a lot in class 2 and 3.
6. More female survived for embarked S and Q and more male survived for embarked C.
7. People who have given higher fare survived more.

1.3 Data Cleaning

Feature Removal
Ticket number, cabin number, name of the passenger, passenger id, parent/children, sibling/spouse and family size might not be correlated with the survival, hence removed it.

Converting data types
Sex: Male and Female are converted to binary type 0 and 1 respectively.
Title: Grouped into Master, Miss, Mr, Mrs and Rare. These groups are represented by integers 0, 1, 2, 3 and 4 respectively. 'Rare' group represents all other title that constitute a very small percentage

of data.
Age: Younger people have a better survival rate (see Appendix), so divided the age into four groups of size 16 years. Replaced age groups with numbers 0, 1, 2, 3.
Embarked: Converted embarked values S, C and Q to numeric values 0, 1 and 2 respectively.
Fare: Converted into the fare groups into 0, 1, 2 and 3.

Missing values
Age: Used random numbers between the mean and standard deviation for each class and gender combination to fill in missing age values.
Embarked: Filled missing values with the most command value. **Fare:** Filled missing values with the median of fare.

Combining features
Based on data observation, people traveling alone have a lesser chance of survival. Hence, combined sibling/spouse and parent/children feature to create a new feature family size. Using family size created a new feature isAlone.
Created a new feature using the product of age and class.

2 Models

Below are the following models that we used to classify the survival of the passengers.

2.1 Logistic Regression

Since the target feature is binary, logistic regression can be used to determine whether a person on board will survive. After training the model, the model accurately predicted survival chances 96 % of the time.

2.2 Naive Bayes

It is a supervised learning algorithm, It uses a classification technique based on the Bayes theorem. So, it can be used for predicting survival chances. When we ran the model created using Naive bases, we got an accuracy of 78 %.

2.3 K-Nearest Neighbor

It is one of the simplest model to predict survival rate. It uses the mean value of k nearest neighbor to predict. We used different K values ranging from 1-50. We found our model gave 87 % accuracy on test data when the number of neighbors is 5.

2.4 Random Forest

It is used when the feature set is categorical. Our dataset has categorical features including age, fare brand, survived, embarked. Hence, we trained our data on random forest algorithms using the sklearn library. Increasing the number of trees in the forest decreases the variance of the overall model, and doesn't contribute to over-fitting. On the other hand, tuning tree size can improve performance by balancing between over- and under-fitting. We found an accuracy of 98 % on testing data.

2.5 Decision Tree

A decision tree is simply a series of sequential decisions made to reach a specific result. Titanic data set consists of categorical features. So it is easy to run a decision tree algorithm on the training data. Parameters for this model were chosen in a similar fashion to that of Random Forest. The model trained on the decision tree gave an accuracy of 94 % on test data.

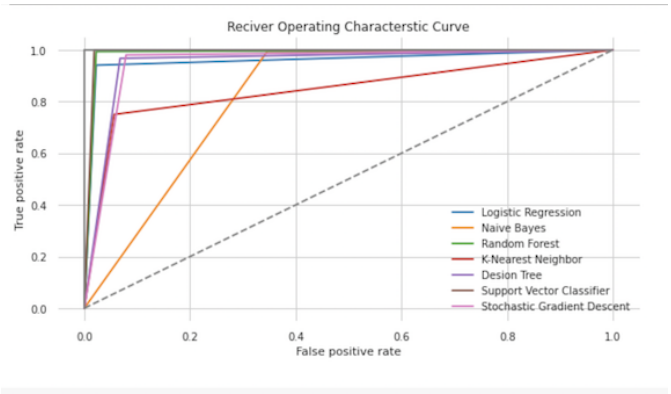

Figure 1: ROC of models

	Model	Accuracy
0	Logistic Regression	96.411483
1	Naive Bayes	77.751196
2	Random Forest	98.325359
3	K-Nearest Neighbour	87.320574
4	Decision Tree	94.497608
5	Support Vector Machine	98.803828
6	SGD	94.258373

Figure 2: Accuracy of models

2.6 Support Vector Classifier

Since logistic regression is likely to over-fit data, we have used Support Vector Classifier that tries to maximize the margin between the closest support vectors. We used a similar fashion as KNN in order to choose the value of the regularization parameter 'C'. Different values of C were tested out in order to get the value which maximizes accuracy. The model trained on the SVC gave an accuracy of 98 % on test data.

2.7 Stochastic Gradient Descent

To reduce the parameter learning time we have trained our data using SGD algorithm. It randomly picks data and computes gradient. It converges faster but jumps off after reaching optima. We can find global optima by reducing the learning rate. The model trained on the SGD gave an accuracy of 94 % on test data.

3 Evaluation method

ROC curve: ROC curve is plotted against True positive rate vs False positive rate. The farther the ROC curve and larger the AOC is, the better the model. So, we access our models using ROC curve.

Metrics: Out of the accuracy, precision, recall and F1 metrics, we accessed the model using **'accuracy'**, since this is the past event, we are expecting the correctly predicted observation to the total number of observations. (see appendix B, for metrics definitions)

4 Results

ROC Curve
In the Figure 1, the farther ROC curves belongs to the models Random Forest, support vector machine classifier. For both of these models, ROC curve has higher AUC to the 0.5 model curve.

Accuracy
In the Figure 2, you can see Random Forest and Support vector machine has the highest accuracies.

Our Colab notebook

`https://colab.research.google.com/drive/1nxuE0lM7SnnTt5H3KYNP8ZeDeYcFZVVG?usp=sharing`

Table 1: Features descriptions

Variable name	Description	Key
PassengerId	Id for each row data of passengers	number
Name	First and last name of the passenger	string
PClass	Ticket class	1=Upper class, 2=Middle class, 3=Lower class
Sex	gender of the passenger travelled	male female
Age	age of the passenger	integer value
Ticket	Ticket number	string
Fare	Fare each passenger paid to be on board	decimal numbers
Cabin	Cabin number	string
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton
SibSp	Number of siblings and spouses travelled. Sibling = sisters, brothers, step sister and step brother. Spouse = husband or wife.	number
Parch	The number of parents or children travelled with a passenger. Parent = Father, mother. Children = daughter, son, step-daughter, step-son. Some children travels with nanny, hence parch is 0 for them.	number

References

[1] Kaggle Competition: `https://www.kaggle.com/c/titanic`

Appendix A

Refer Table 1: Feature descriptions to know about the feature variables, their keys and their data types.

Appendix B

Metrics:

Accuracy is the ratio of correctly predicted observation to the total number of observations. Here, accuracy means, how many passengers' survival rate did the model predict correctly out of all the dataset?

Precision is the ratio of correctly predicted observation to the total number of predicted positive observations. Here, precision means, out of all the passengers that are predicted as survived, how many passengers actually survived the wreck?

Recall is the ratio of correctly predicted observation to the total number of predicted observations. Here, recall means, out of all the survived passengers, how many were predicted correctly?

F1 score is the weighted average of precision and recall.

Appendix C

Figure 3: Sample of Kaggle's titanic dataset.

	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

Figure 4: Female survived more in this shipwreck

	Sex	Survived
0	female	0.742038
1	male	0.188908

Figure 5: Parents and siblings travelled with passengers have no correlation with survival

	Parch	Survived
3	3	0.600000
1	1	0.550847
2	2	0.500000
0	0	0.343658
5	5	0.200000
4	4	0.000000
6	6	0.000000

	SibSp	Survived
1	1	0.535885
2	2	0.464286
0	0	0.345395
3	3	0.250000
4	4	0.166667
5	5	0.000000
6	8	0.000000

Figure 6: Younger people survived more

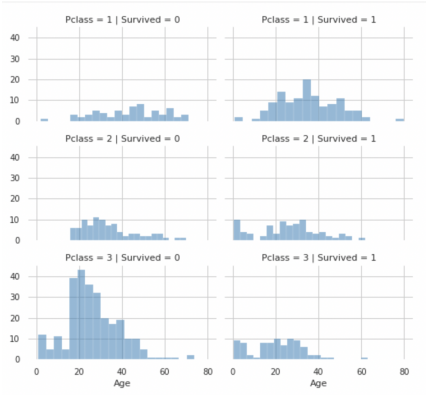

Figure 7: Upper class people survived more

	Pclass	Survived
0	1	0.629630
1	2	0.472826
2	3	0.242363

Figure 8: Higher the fare, higher the survival rate

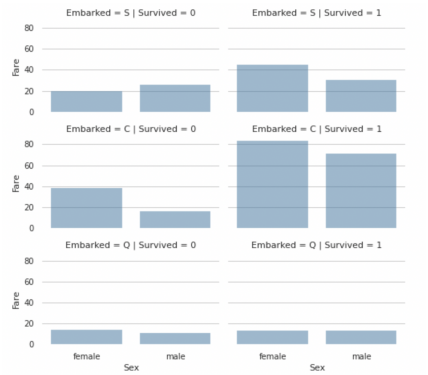

Figure 9: Female and embarked with survival rate

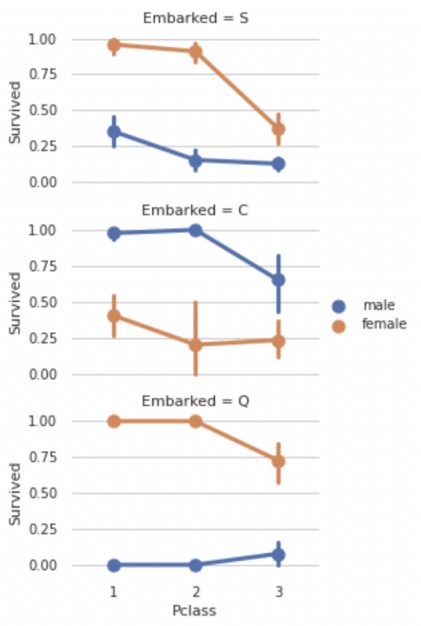

Figure 10: Title of a person vs survival rate

	Title	Survived
0	Master	0.575000
1	Miss	0.702703
2	Mr	0.156673
3	Mrs	0.793651
4	Rare	0.347826