Titanic - Machine Learning from Disaster

Adrian Jelenici
CS Undergraduate Student
Simon Fraser University
aja60@sfu.ca

Alfonso Ocampo
CS Undergraduate Student
Simon Fraser University
alo6@sfu.ca

Dalveer Dosanjh
CS Undergraduate Student
Simon Fraser University
dsd6@sfu.ca

Daven Chohan
CS Undergraduate Student
Simon Fraser University
dca120@sfu.ca

Ricky Xian
CS Undergraduate Student
Simon Fraser University
rxa13@sfu.ca

Abstract

In this report based on the Titanic Kaggle Competition, we built a machine learning model to predict who would survive the Titanic Shipwreck. We first reviewed and cleaned the data, and found that age, gender, and class are the most important variables. Then, we used the K-Nearest Neighbor machine learning model along with the Random Forest Classifier model to analyze and predict the data. This resulted in the Random Forest Classifier being more accurate in predicting the titanic mortality.

1 Introduction

This report will tackle the problem of predicting the survivability of passengers on the titanic, a historical tragedy. Given specific information about each passenger that conclusively did or did not survive, a predictive model can be made. This predictive model can be used to accurately determine the survivability of passengers if similar data about the individuals is known. This is important in allowing safer ships to be constructed in the future.

1.1 Data Analysis

The data given will be trained to help assist in building the predictive models. Initially information will be analyzed to find out if certain statistics relate to the survivability of the passengers. This can be done by carefully checking the survival rates of passengers with certain characteristics compared to passengers with the opposite characteristic. Building graphs and/or charts of this analysis of the data will extensively help in finding out how important a passenger trait is in determining survivability. Once the analysis is completed, statistics found to be irrelevant will be removed to allow the predictive model to be more efficient.

1.2 Predictive Models

Although multiple predictive models exist that can be utilized to create an algorithm for this problem, k-nearest neighbors and random forest will be outlined in this report. First, k-nearest neighbors (KNN) will be used as a sufficient machine learning model after an acceptable metric is created. KNN will be able to accurately classify if each passenger in the test data survived or not by calculating the distance between the specific test data and the training data.

The second predictive model used is random forest, a machine learning algorithm that makes use of multiple decision trees to classify testing data. This can be used to predict whether a passenger survived with high accuracy, while being resistant to potential outliers. Random forest will be more accurate than a single decision tree, thus being more favorable of a choice. It is an easy and flexible model that can be properly utilized to solve this problem.

2 Data

Kaggle's original competition data consists of 891 entries of train data and 418 entries of test data. This data includes fields for name, age, gender, socio-economic class, and more. We found, however, that some data fields may prove to be more useful than others when predicting which passengers will survive. Our first step in this project was therefore deciding which data fields to keep and which to ignore in our analysis. We predicted that age, gender, and class would be the most impactful (and thus the ones we will be focusing on). We predicted that men of enlistment age would be the least likely to survive, while children and seniors of any gender would be the most likely to survive. We also predicted that those of higher class would be more likely to survive than those of lower class since they would have been occupying the upper deck - the furthest point away from the iceberg impact. On the other hand, we predicted that fields such as name, ticket number, and port of embarkment would not have any noticeable effect on a passenger's survivability, so we did not take those into account. In the figure below, we can see how each piece of data correlates to each other.

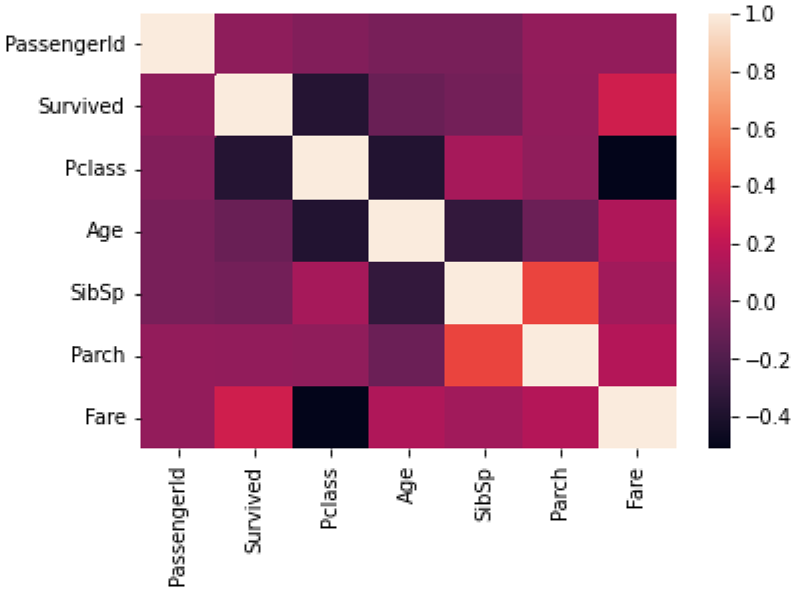

Figure 1: Heatmap visualization of the data correlation

3 Analysis

3.1 K Nearest Neighbors

The first machine learning model we used to approach the problem was K Nearest Neighbours. The KNN model can classify and take into account multiple categories of data. This helped us find relations among different features of data, since each point is plotted around similar points, we were able to find consistencies between them. Although KNN has good pattern recognition with the data points it uses, it can be influenced by outliers so it was important to use the correct amount of neighbors to compare with and features that misrepresented the predictions needed to be cleaned and removed. With cleaned data, KNN model was able to more accurately plot data points in the graph, creating a consistent survivability rate prediction based on the data around the k nearest neighbors.

3.2 Random Forest

The second machine learning model we used to approach the problem was the Random Forest Machine Learning Model. We decided to use this model for a variety of different reasons. The first reason is due to its low risk of overfitting (thanks to bagging and feature selection). By using this model we can therefore succeed in fitting additional data and/or predicting future observations reliably, something that is incredibly important with this specific project. The second reason we decided to use this model is because of its often high accuracy. By binding the variables, Random Forest ensures that it's not influenced by outliers to a heavy degree. Furthermore, by building multiple decision trees, Random Forest also ensures that it's more accurate than a single decision tree. With all that being said, however, there are a few downsides to this model as well. Namely: Random Forests can be computationally intensive for large datasets such as the ones that are being dealt with in this project. It's for this reason that the Random Forest Machine Learning Model was used as our second choice rather than our first.

4 Results

After cleaning and implementing our machine learning models, we are able to score our results and accuracy. Firstly, our KNN model seemed to have little variation in score with our reduced data size. After testing K values, we found that the K value of 3 was the most optimal. With our K value of 3, age, sex, and fare were the more important values of the data. Secondly, our RF model had a lot more variation in score compared to the KNN model. After a few runs, it fluctuated around the score of 0.80. The highest score obtained for Random Forest was 82.08% in our limited testing. The more important features of RF were the family size and sex. After reducing our data, the sex and family size were a lot more variant. Finally, after scoring both our KNN and RF models, we found that our RF model was approximately 2% more accurate than our KNN model. However, even though our RF model seemed to be more accurate than KNN, it also had more variation, which means that occasionally it can be less accurate than KNN.

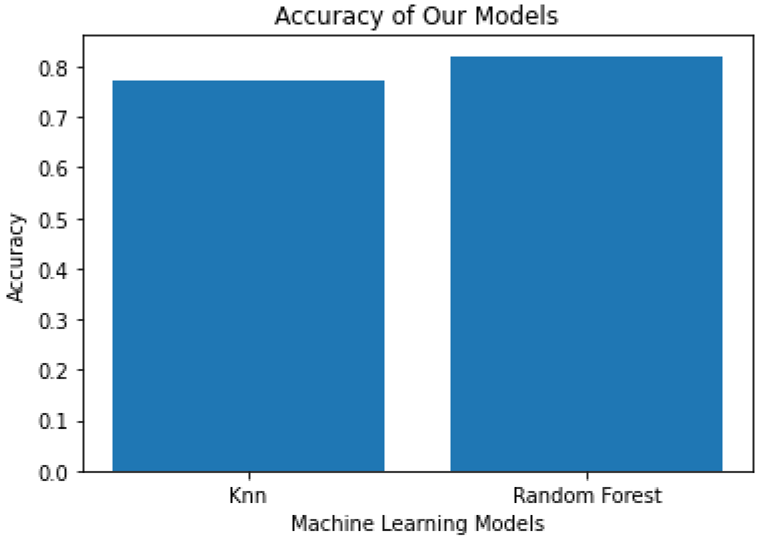

Figure 2: Accuracy Visualization of Our Machine Learning Models

5 Conclusion

In conclusion, the goal of the project was to build a predictive model for the survivability of passengers on the Titanic. The Titanic survivability predictive model is important because it will give us vital information that is needed to create safer ships. For data processing, we removed all null values and reduced the dataset to 400 entries. In addition to this, we also removed columns that we believed were irrelevant like name, ticket, and cabin. The algorithm chosen for this project was KNN and Random Forest. For KNN, we got an accuracy score of 77.05% and for Random Forest the accuracy score was roughly 80%.

Contributions

Adrian Jelenici - Equal Split (Focused on: Data & Random Forest)

Alfonso Ocampo - Equal Split (Focused on: Results & K Nearest Neighbors)

Dalveer Dosanjh - Equal Split (Focused on: Data & Results)

Daven Chohan - Equal Split (Focused on: Abstract, Introduction, & Data)

Ricky Xian - Equal Split (Focused on: Conclusion & Data)

References

Balakrishnan, P. (2014, August 29). *Tutorial: Titanic dataset machine learning for Kaggle - Corpocrat Magazine*. Corpocrat Magazine | Design, Programming and Finance Blog. https://corpocrat.com/2014/08/29/tutorial-titanic-dataset-machine-learning-for-kaggle/

Sehgal, M. (2019, February 11). *Titanic Data Science Solutions*. Kaggle. https://www.kaggle.com/code/startupsci/titanic-data-science-solutions

Titanic - *Machine Learning from Disaster* | Kaggle. (n.d.). Kaggle. https://www.kaggle.com/c/titanic