Brian McMillan

Data Mining

Dr. Wachsmuth

April 29th, 2022

**Titanic Project Paper**

**The sinking of the Titanic occurred in the early morning hours of April 15th, 1912. Approximately 1,500 people died out of the 2,200 passengers and crew members on board. In this paper, I explore data regarding the Titanic, and attempt to discover how variables such as Age, Gender, and Social Class, could be used to predict survival.**

**Data Preparation**

**Specified the dataset (train) by renaming it as data frame (df).**

****

I began the data preparation phase by studying the train set and figuring out what data types are associated with each of the variables involved. From this, a few issues arose. Firstly, “Survived” and “PClass” ought to be changed to categorical variables. I went a step further and specified “PClass” as an ordinal categorical variable, since it entails a ranking of social-class.





Second, upon studying the data, I noticed “Age” has an alarming number of null values. I hypothesize age will be an important indicator of survival, so I replaced all null “Age” values with the mean.



I also removed “Cabin”, “PassengerID”, “Fare”, “Ticket”, and “Embarked”, from the set. “Cabin” had far too many null values. The other variables are not of great importance.



Next, I normalized the data. Text

Description automatically generated

I feel fairly confident that my data is clean. However, I have to consider application of algorithms later down the road. I plan on trying a decision tree as well as K-Nearest Neighbor. K-Nearest Neighbor will require some data preparation. Specifically, I need to make a new dataset where the null values are omitted, and the factor variables included are cast as numeric variables.



This concludes the data preparation phase. I have also downloaded and imported some possible functions I might use.

Text

Description automatically generated

**Data Exploration**

It is unfortunate that there is so little data in the “Cabin” category. Each “Cabin” is associated with a letter, and the closer that letter is to “A”, the closer it is to the Promenade Deck. This could’ve been interesting to explore since safety equipment is usually held on the Promenade deck. Perhaps proximity to this deck could affect survival.

It is likely class influences survival rate too. Here I have created some data visualizations.

Chart, bar chart

Description automatically generated

Here we can see that of those who did not survive, over 200 of the deaths were from the third social class. But the third class also leads in survival. This is likely because the third class is the largest class. Let’s take a look at a table to get a greater understanding.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 3rd class | 2nd class | 1st class | TOTAL |
| Did not Survive | 231 | 59 | 58 | 348 |
| Survived | 84 | 59 | 78 | 221 |
| TOTAL | 315 | 118 | 136 | 569 |

While the third class leads in survival, they are also the most populated class. As a percentage only 27% of the 3rd class survived. Compare this to the first class, which had a 57% survival rate. The first class, interestingly enough, is the only class that had more survivors than deaths. It seems class has some impact on survival.

Next, I have created a similar graph for age:

Chart, histogram

Description automatically generated

Most of the passengers were in the age group 20-40, so I am not surprised that there is a spike there. What this graph really shows is that kids had a high likelihood of survival as compared to the older age groups.

Let’s see if we can make this a bit more evident by grouping the ages together before graphing.

\*sample R code

A picture containing chart

Description automatically generated

Chart, bar chart

Description automatically generated

Age group 0 to 10 is the only group that had more survivors than deaths. We know that children and women were saved first, so this isn’t surprising. However, here we have shown graphically how age can be a great predictor of survival. Next, I will display a similar theme regarding women.

Since we have established that class is an indicator of survival, I will add this to the next graph and display it next to sex. The following graph shows Survival by “Sex” **and** “Pclass”.

Chart, bar chart

Description automatically generated

These results are interesting. The amount of women who survived in the third class is alarmingly low as compared to the number of women who survived in the second and third class. If you were a woman and a part of the second or third class, it seems you had a great chance of surviving as compared to other variable combos.

Chart, bar chart

Description automatically generated

A picture containing text, clock

Description automatically generated

\*Table format of same data\*

As suspected, having a parent or a kid also seems to be a decent indicator of survival. If you had two kids, or two parents on board, your likelihood of survival was greater than your likelihood of death. This makes sense logically. If you were a kid with two parents on board, you had a father to assist you to safety, as well as a mother to help you get on a lifeboat. Likewise, if you had a kid on board, your likelihood of survival was bolstered because mothers were grouped with children and helped first.

**Data Modeling**

Data models are visual representations of data elements and the connections between them. In the next few paragraphs, I will be discerning which classification method to apply and eventually choosing one.

Neural Networks:

A neural network is a machine learning technique that is used to identify the category of new observations by using the training set. If applied to the Titanic set, a neural network could classify the category “Survived” based on the training set. A neural network takes inspiration from the learning process of a human brain (thus, its importance in the field of cognitive psychology). Like a brain, the network learns over time by analyzing new data. I will refrain from using a neural network as their success can be dependent on large amounts of data. I would much rather use a neural network to classify something like type of cancer, which would have millions of entries in the dataset. The titanic dataset is quite small and is an event that only happened once, so there is only so much data that can be obtained.

Nearest Neighbor:

K Nearest Neighbor is an algorithm that stores all available cases and classifies new data based on how its neighbors are classified. I will try KNN because it is quite simple to implement and fine tune. I have also already created a dataset that will work with KNN.

Graphical user interface, text, application

Description automatically generated

KNN classified correctly 76% of the time.

Decision Tree:

A decision tree is a graph used in machine learning and datamining that represents different variables in the form of a tree. I think a decision tree will be my best course of action. They are known for being able to classify data with many variables extremely quickly. I also think the titanic dataset will be easier to classify if it’s broken down into smaller and smaller subsets—and--as shown in the exploratory data analysis, it seems there are very specific combinations or subsets of variables that predict survival.

Text, letter

Description automatically generated

Diagram, schematic

Description automatically generated

With this we have achieved 88.5% accuracy.

A picture containing text, receipt

Description automatically generated

**Ethical Issues:**

*To broaden the subject matter, consider a cruise line company. The company may want to understand more about its customers, such as demographic information. What classifications of its customers may serve its business goals? What ethical issues may arise from these applications? Discuss ethical issues arising from automatically classifying the cruise line customers and customizing services according to the classifications.*

There would undoubtedly be ethical issues that arise from doing so. Perhaps the company has an email list where they offer previous customers deals and discounts. Let’s also assume the data has brought the company to the conclusion that race, gender, and class has an impact on how much money someone will spend on the ship. Specifically targeting, emailing (offering deals/discounts) , and attempting to recruit people based on these factors could be considered unethical. Even if the motivating factor is only money, the company would still be guilty of prejudice. This could be considered even worse if the customers never consented to having their information be used in research or ad campaigns. No doubt, if this was exposed to the public the company would face many legal battles as well as a tarnished reputation.