Data Mining – Assignment 1

Andrew Bedard – Artagan Malsagov – Shabaz Sultan

April 15, 2015

1 Own Dataset Initiative

2 Titanic Survivors
or: How I learned to Stop Worrying and Love the Data

2.1 Introduction

On the 14th of April, 1912 the RMS Titanic hit an iceberg and sank a few hours later. Of the 2207 people on that ship 1501 lost their lives on that night. These tragic events offer the opportunity to study exactly how people behave during such a life-and-death event.
From an economic perspective one can wonder if the model of humans as 'homo economicus', that of humans as rational, self-interested actors is the best way to model behaviour of people on that night. An econometric analysis shows that such a 'homo economicus' model is overly simplistic because females and children were more likely to survive than physically stronger males. An analysis using said model is still valuable however because people who were closer to their prime age, were of higher social class or had access to more information were more likely to survive[1][2].
The reason why the deathrate was so high starts with the fact that there were not enough lifeboats. On top of that fewer passengers survived than there was space in lifeboats because of supposed reluctance to leave the ship (e.g. because of disbelief that the ship would actually sink or wives that did not want to be separated from their husbands). The difference in survival between males and females can be explained as a result of policy. The official explanation from the Mersey inquiry to explain the difference in survival rates between classes was that lower-class people were less willing to be parted from their belongings and that their English was poorer making them less able to follow orders from the crew. Statistical analysis based on nationality as a proxy for language ability refutes this second claim and suggest that explanations rejected by the inquiry (layout of the ship disadvantaging lower class people and outright discrimination when letting people on lifeboards) are more likely [3].
We can use publicly available data with personal details of Titanic passengers to build our own models to see if we are able to predict survival of a passenger based on things like sex and class.

2.2 Analysis of the Data

The data used in this section is provided through the Kaggle website, which host machine learning competitions. The Titanic dataset is intended as one that users can use to practice and get up to speed with machine learning. The data is provided as a list of 891 passengers in a training set and 418 passengers in a test set. Both sets have a number of attributes marked, mentioned in table 1.

Table 1: Attributes proved for each passenger in the dataset.

attribute	possible values
passenger class	1,2 or 3 (1 is upper class, 3 is lower)
name	first and last name with title
	(i.e. mr., miss. etc) and possible initials
sex	male or female
age	number in years
siblings & spouses	number of siblings and spouses on board
parents & children	number of parents and children on board
ticket	ticket number
fare	ticket fare
cabin	cabin code (not available for some, multiple for some)
port of embarkation	Cherbourg, Queenstown or Southampton

In addition to the attributes mentioned in table 1 the passengers in the training set also have an attribute denoting if they survived. This allows us to use the training set as the input for a supervised learning model and use said model to create predictions for survival on the test set. The Kaggle system allows said predictions to submitted and percentage of correct predictions on the test set get displayed in a leaderboard. It should be noted that because this is a dataset available fully marked up (i.e. with survival marked for all passengers) it is trivial to cheat and get a 100% score. As such the leaderboard is less reliable than other Kaggle competitions and is only intended for practice by Kaggle.

2.2.1 Statistical Exploration of Data

We can start by gathering some basic statistics on the training set to get a sense of the dataset and the distributions for certain attributes. In the trainingset 38.4% survived. There are 216 first class passengers, 184 second class passengers and 491 third class passengers in the set. The passengers are 64.8% male and 35.2% female.
Based on both intuition and the literature sex and class seem like the most likely candidates for predicting the survival of a passenger. Because these are also the most straightforward attributes to analyse it makes sense to start with them when first exploring correlations in the dataset. We can start by looking at the correlation coefficient between sex and survival, which is 0.54, meaning there is a positive correlation between being female and surviving (male and non-survival are encoded as 0, female and survival as 1).
Next we look at class and survival. There is a positive correlation between being a first class passenger and surviving, with a correlation coefficient of 0.29. There

is a much lower positive correlation for second class passengers of 0.09 and a negative correlation of -0.32 for being a third class passenger and surviving.

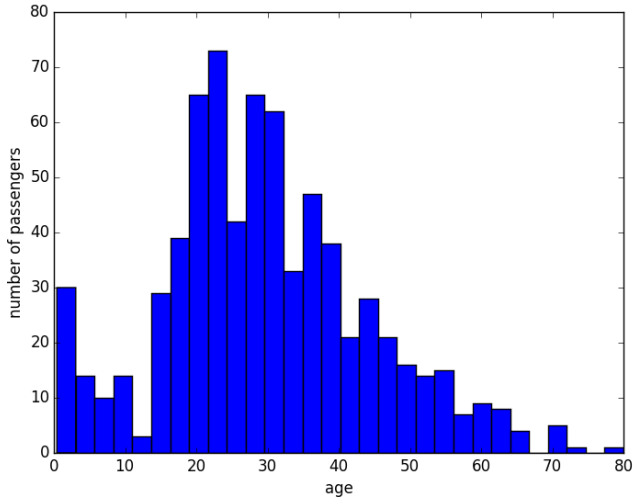
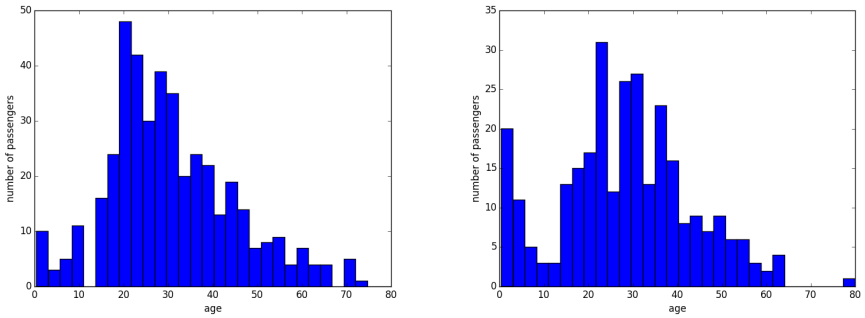


Figure 1: Distribution of ages in training set.

.

Only 714 of the 891 passengers in the training set have their age marked. The average age of these 714 passengers is 29.7 years, with a standard deviation of 14.5. The distribution of ages is plotted in a histogram in figure 1.
Next we can look if there is a difference between the age distribution between survivors and non-survivors. Judging by the naked eye the two histograms in figure 2 look pretty similar and indeed the mean and standard deviation are pretty close as well. There is perhaps a slight difference in the shape of the two distributions, which is expressed as a difference in the kurtosis.
Based on this analysis it seems like age would not have great predictive value for survival. It might however still be true that age is more predictive after the population has been split up based on other criteria first.



(a) Age distribution of non-survivors. Mean 30.6, std.dev. 14.2, kurtosis 0.28.

(b) Age distribution of survivors. Mean 28.3, std.dev. 15.0, kurtosis −0.06.

Figure 2: Age distributions of survivors and non-survivors in the training set.

Finally we look at embarkation ports. Based on intuition it seems less likely that this attribute is particularly predictive of survival. Looking at the data however, passengers who embarked from Southampton had a negative correlation with survival at −0.16. Passengers from Cherbourg had a positive survival correlation at 0.17 and those from Queenstown had almost no correlation at 0.00. One can wonder if these differences can be explained by a difference in e.g. sex and class between passengers from different ports. One way to explore this is building a decision tree with embarkation port, sex and class as attributes and see how far in the tree embarkation port ends up at, which we will try in section 2.3.

2.2.2 Data Augmentation

REPLACE in particular building up family trees

2.3 Machine Learning Models

probably decision trees and logistic regression at the very least

3 Research and Theory

Describing someone else's work is certainly a good way to advance further until you begin to feel your own initiatives coming. Let's try that by starting this section by describing the winning entry of a Kaggle competition.

3.1 IJCNN Social Network Challenge

The Social Network Challenge was hosted by IJCNN and posted on the Kaggle platform in the period of Nov 2010 - Jan 2011. The challenge was to predict edges/connections between nodes/people in an online social network based on an edge dataset obtainted by crawling said network. This dataset was partioned into a training set and a test set, where the test set was expanded by an equal number of fake edges. Predicting then meant the trained algorithm had to classify the $8,960$ test edges as either true or false, after having trained on the $7,237,983$ training edges. As an evalutation measure the area under the

ROC-curve was used (AUC), meaning the closer the AUC is to 1 the better the evaluation of the algorithm. Needless to say user identities of the nodes were obfuscated by assigning random IDs to the nodes in the provided dataset, otherwise a group might cheat its way through.

And the winner was a team going by the name "IND CCA". Its members wrote their winning approach in an article [4] to which the interested reader is referred. The details of their approach are quite intricate, so here's the gist of it. While de-anonymization was forbidden, that is exactly what the winning team did. After finding out the data had been obtained by crawling Flickr, the group decided to crawl Flickr themselves and in so doing managed to de-anonymize 64.7% of the test edge-set. On the remainder of the test set they used a Random Forest Classifier, training this algorithm on standard link prediction features of both the training and the de-anonymized test set, thus achieving a whopping and winning AUC of 0.981.

It was this piece of bravado to de-anonymize the test set that made the method stand out. It wasn't cheating, since once the group got a lead in the competition, they contacted the organizers and explained their method, offering to resign all together from the competition. Their method was approved though, and they went on to win the competition. The aim of the slight cheat in their method was of a different kind though, for they wished to raise awareness of the lasting possibility of de-anonymization in machine learning contests and hoped this would provide food for thought on how contest should be run in the future.

4 MSE vs. MAE

MSE and MAE are used as evaluation statistics and preference of one above the other is heatedly debated. However, even before delving into the proposed arguments, one can naturally assume a position of temperance: both metrics have their proper use and should be used accordingly. Understanding the properties of these metrics to do exactly that is crucial and shall be the subject of this section.

4.1 Discussion

A defintion of both would be in order: suppose that $\hat{y}$ is a vector of n predicted values and y is the same vector of true values. MSE is measured as:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

and MAE measures as:

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

References

[1] Bruno S Frey, David a Savage, and Benno Torgler. Interaction of natural survival instincts and internalized social norms exploring the Titanic and

Lusitania disasters. *Proceedings of the National Academy of Sciences of the United States of America*, 107(11):4862–4865, 2010.

[2] Bruno S. Frey, David A. Savage, and Benno Torgler. Behavior under extreme conditions: The "titanic" disaster. *The Journal of Economic Perspectives*, 25(1):pp. 209–221, 2011.

[3] W Hall. Social class and survival on the S.S. Titanic. *Social science & medicine*, 22(6):687–690, 1986.

[4] A. Narayanan, E. Shi, and B.I.P. Rubinstein. Link prediction by de-anonymization: How we won the kaggle social network challenge. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1825–1834, July 2011.