Predicting survival rates of the Titanic using machine learning techniques (the Kaggle dataset)

Rune Munkholm Jensen

1 Strategy

The strategy taken will be to identify good data for generalization, by visualizing the correlation between attributes and survival. A starting point is to identify those attributes that appear to generalize well in terms of predicting survival directly, i.e. where the value range of the attribute has a clear and obvious correlation with survival outcome.

Once relevant attributes have been identified, these will be prepared for machine learning algorithms by predicting (filling in) missing values, a step generally necessary since most machine learning algorithms do not work well (or at all) with missing data. The methods chosen for modeling missing values for different attributes may differ, depending on the assumed nature of the data specific to each particular attribute.

Next, we will use the training data and an appropriate machine learning algorithm to produce a model for predicting survival and finally evaluate the model's accuracy on the test data set to see how well we fared. The results may come in handy the next time one is to escape a sinking early nineteen-hundreds steamship in icy waters, although they do unfortunately come a bit too late for the good folks aboard the *Titanic*. We will proceed anyway, however, in the name of science.

2 Missing data

An overview of missing data for the train and test sets is given by tables 1 and 2.

Attribute	Count
Age	177
Cabin	687
Embarked	2
Fare	0
Name	0
Parch	0
Pclass	0
Sex	0
SibSp	0
Survived	0
Ticket	0

Table 1: Missing entries count in the training data set. Three attributes have missing values: *Age*, *Cabin* and *Embarked* .

Attribute	Count
Age	86
Cabin	327
Embarked	0
Fare	1
Name	0
Parch	0
Pclass	0
Sex	0
SibSp	0
Survived	(hidden by Kaggle)
Ticket	0

Table 2: Missing entries count in the test data set. Three attributes have missing values: *Age*, *Cabin* and *Fare*.

3 Relevant attributes

Here, the attributes that are identified as relevant to survival are outlined.

Gender

Interestingly, it appears from the attribute *Sex*, shown in figure 1, that females had a slightly better chance of survival. This could possibly indicate that emotional nature/response to the impact was beneficial in abandoning the ship in time.

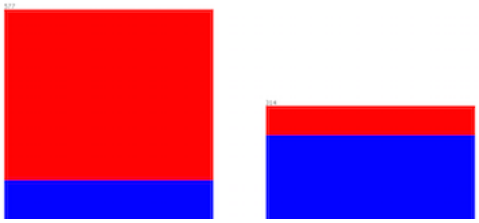

Figure 1: Females (right) fared far better in the crucial moments, compared to men (left).

Age

Age also appears to be a key factor, as seen in figure 2. Clearly, young infants may have benefited from adult and female supervision, but then mortality takes dominance abruptly from working age and remains high thereafter.

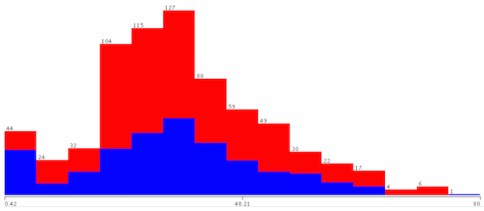

Figure 2: Age versus survival, with opposing trends at each end of the spectrum.

Fare

Also seen is that *Fare* has a strong influence on survival, as shown in figure 3.

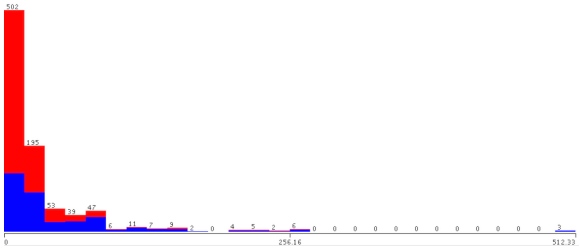

Figure 3: More money, more survival.

Social class

Somewhat related to *Fare* is the social class, as indicated by the *Pclass* attribute. As one might expect, figure 4 similarly shows that the higher the class, the better the chance of survival.

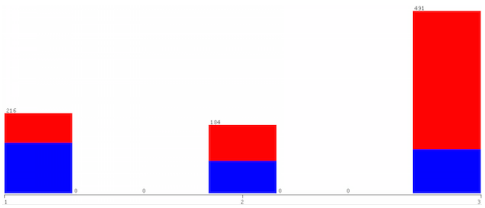

Figure 4: Survival for social classes 1 through 3, left to right. By far most of the people on the lowest of the classes perished.

4 Irrelevant attributes

Some attributes are presumed to be irrelevant by nature for generalizing on the target label, *Survival*. These will therefore be discarded: *PassengerId* (a linearly increasing number), *Name* (presumed to be random characters) and *Ticket* (a string generated independently from any other data).

In addition, attributes that have not been chosen as relevant in the previous will discarded too, due to either being assumed irrelevant, noisy or highly correlated with another relevant attribute.

For an example of the latter, we note that embarkation location also have something to say about survival, as shown in figure 5, with slightly diminished survival for those that embarked at Southampton.

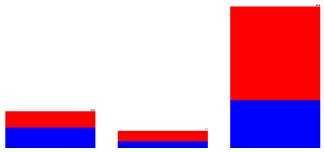

Figure 5: Embarkation and survival, with Southampton on the far right.

However, we note that this is highly similar to the previous figure 4 for social class, indicating that embarkation is highly correlated social class. Comparing the two, as shown in figure 6, we get the intuitive overview: most of the lower class boarded at Southampton, an early 1900s industrial city, and these were also the ones that got the short end of the stick in terms of survival.

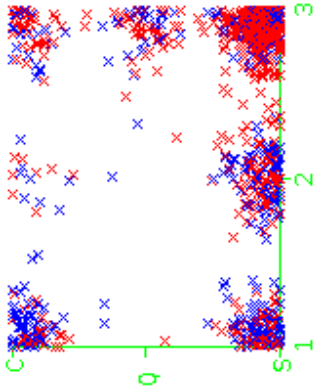

Figure 6: Embarkation versus social class, with Southampton on the far right and lowest class on top.

5 Additional attributes

To play around with potential hidden information, some assumptions will be made that allows for extracting additional attributes, as described in the following.

Name length

We will assume that, in general, fancy people of the early 1900s fancied to carry fancy names and that simple people would be given simple names. For instance, a low-status blacksmith would have been named "John Blacksmith", as opposed to "Distinguished Service Order and Military Pink Cross recipient by the Queen and

Emir of Arabia, Sir Lord John Smithsonian Grandius of All Blacksmiths and Foreign Subjectables".

With this assumption in mind, we may introduce an additional attribute, *NameLength*, that will hopefully provide some indication of class. The assumption made here, is that people of class are more likely to mingle with other people of class, but may carry a ticket that reflects this to a lesser extent (the other way around would be unfathomable, as fancy folks will always try and suck up to those upwards in the hirearchy of fanciness, not downwards). Hence, people with long names may, in general, have been higher above the waterline, because the fancy folks wanted to be on top - and thus may have had better chances of survival.

As the derived figure 7 shows, this correlation does in fact exist and that long names are well worth the effort it in the long run.

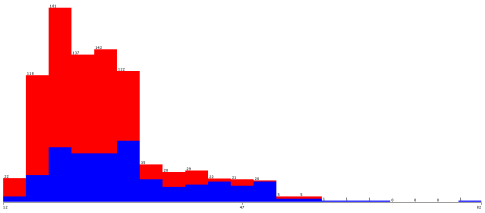

Figure 7: People with short names (small x-values) appear to have had the odds against them in terms of survival.

Cabin deck and number

The *Cabin* attribute is given to us as zero-to-many string tokens, each composed of both the deck letter and cabin number, which is difficult to generalize on. We may instead parse this as two respective attributes, settling on the first string token as the primary (to guard against cases where multiple tokens are given).

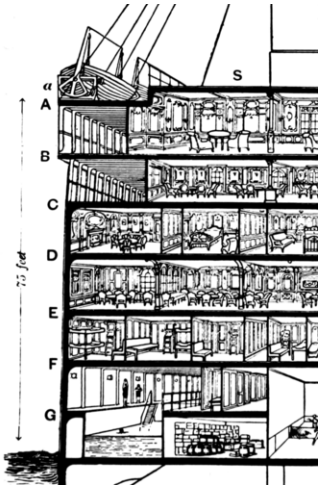

Figure 8: The decks of the Titanic, each indentified by the letters A through G (a T deck also exists in the training data, which was the boat deck on the very top).

The intuition behind how we may use this information, is clear from the prior knowledge we have on how the ship sunk: the crew were sailing above the speed limit in the Atlantic and struck an ice berg almost head-on, after which the ship began to pivot around the gravitational center of its longitudinal axis, slowly emerging the cabins from the bow and lower decks[1] (had the crew instead been trying to parallel park their way to New York, then we would have expected the ship to have sunk from the rear). We thus have the misère[2] shown in figure 9, from which one could guess that passengers at higher deck levels and higher cabin numbers had better chances of survival [3] .

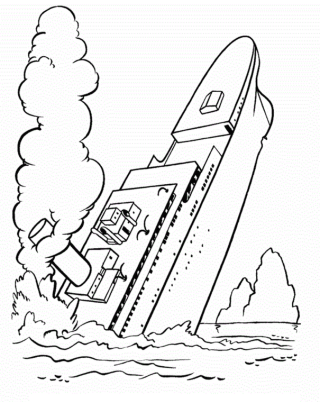

Figure 9: An artist's captivating dramatization of the gradual sinking, which is currently serving as cute child coloring material on onlycoloringpages.com. For passengers genuinely interested in maintaining good health, cabins at the stern (rear) are obviously preferable at this point in time and may thus, to the statistician that came after, be an indicator of post-disaster longevity.

As seen in figure 10, survival does differ between the decks (making it a useful attribute) and surprisingly decks D and E appear to have been optimal.

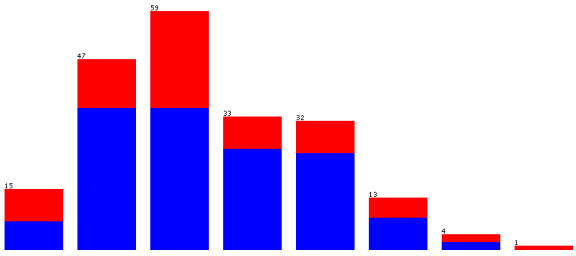

Figure 10: From left to right, survival for decks A through T. Here, fancy folks on the top deck did not have the upper hand in terms of survival.

[1] Anyone that saw the blockbuster movie *Titanic* will be able to confirm that this is in fact how things went down

[2] French for "misery"

[3] Using the deck plans available on www.encyclopedia-titanica.org/titanic-deckplans/

Also, we see from figure 11 that higher cabin numbers do not necessarily imply higher survival, as one could initially assume due to decreasing proximity to the (increasing) waterline. However, there does appear to a noticeable optimum in terms of survival, around cabin numbers 30, making the cabin number an informative attribute.

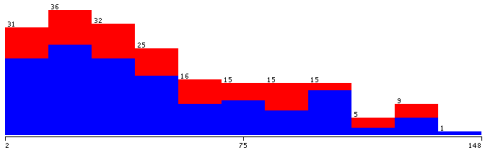

Figure 11: Survival for increasing cabin numbers.

6 Data prep, missing values

Missing values for age

When comparing age to survival, as shown in figure 12, it could appear that age roughly follows a Gaussian distribution that differs slightly between survivors and non-survivors. This simplification obviously does not capture all intricacies perfectly, most notable the sudden drop around adolescence/adult age and the increasing spike towards infancy. One could certainly easily explain the latter quirks by speculating that adolescents were kept in school up up until a specific age regardless of class - and perhaps affirmative action was showing its filthy face when filling up the life boats, i.e. favoring women and children.

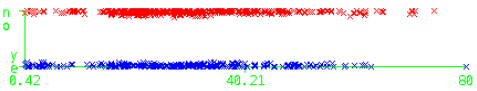

Figure 12: *Age* could appear to follow a Gaussian distribution.

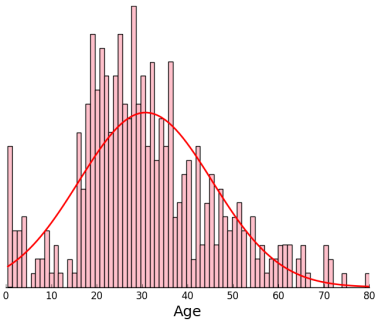

Figure 13: Fitting a Gaussian to age on men (assumed to be non-survivors).

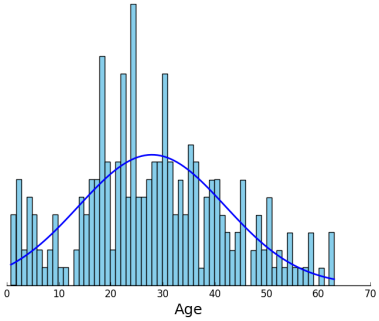

Figure 14: Fitting a Gaussian to age on females (assumed to be survivors).

To deal with these peculiarities we will assume this: the age distribution on board the Titanic followed a Gaussian, but infants introduced an additional Gaussian of their own, making it a multimodal distribution. To simplify a bit, we will model these two distributions individually: we ignore the affect of the overrepresentation of infants on the general age and model the infants separately using some age threshold. When sampling, we will then randomly choose to sample from the infant distribution at the same fraction as the infant group composes of the entire group (infants + non-infants).

Furthermore, we will separate the cases for survivors and non-survivors, since they appear to differ somewhat. But since we do not have the prediction label *Survived* available for the test set, we will rely on the strong correlation between survival and gender: females will be assumed to be survivors in general and men non-survivors. This fitting is shown in detail in 13 and 14.

In numbers, the infant thresholds (as chosen by visual inspection) are $Age < 11$ for females and $age < 14$ for men, yielding the 'infant' models shown in figures 15 and 16.

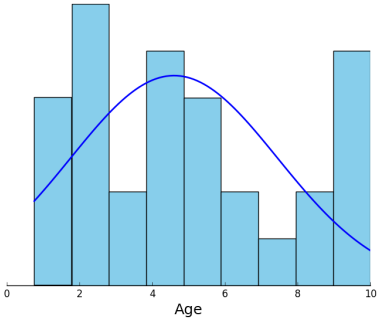

Figure 15: Fitting a Gaussian to females (survivors) under age 11.

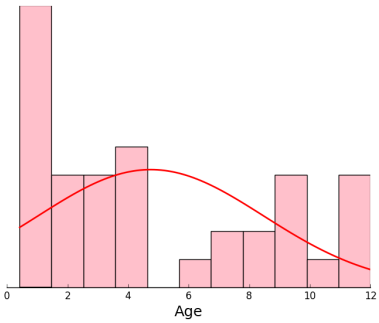

Figure 16: Fitting a Gaussian to men (non-survivors) under age 14.

Clearly, survival predictions appear to become more fluctuant at these lower intervals, but we will justify the choice of model by the assumption that the probabilities of finding any children aboard drops sharply at both ends of the 'infant' intervals, and hence, the Gaussian will certainly be superior to e.g. simply taking mean values.

We could also have chosen to refine the fraction of infant samples by investigating which distribution has the highest likelihood. Technically, this would have been done by finding the fraction of infants up until the age at which the two Gaussians intersect, although this could seem less correct given the abrupt (not smooth) cutoff for adults. This idea is exemplified in figure 17 using the survival attribute from the train set.

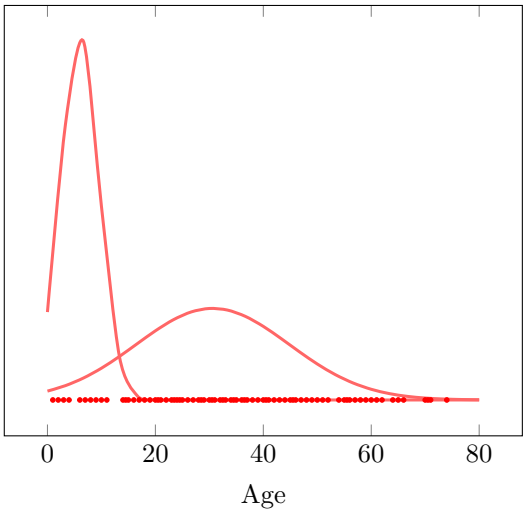

Figure 17: Age for non-survivors modeled by two guassians that have (positive) intersect at $Age = 13.3339$. The right Gaussian is modeled for all non-survivors, and the left for all survivors under age 14.

Missing values for deck(number)

Reinvestigating the previously shown figure 10, it could appear that deck allocations follow a Gaussian distribution that would be almost identical for each survival outcome (hence, we will not branch on survival assumptions for this attribute). A single Gaussian model, as shown in figure 18, will therefore be used to model missing values for decks.

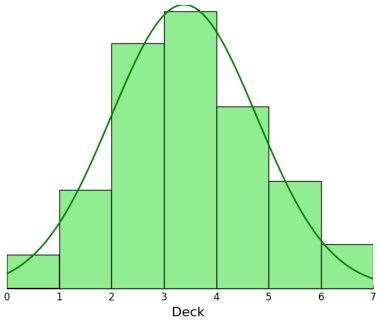

Figure 18: Fitting a Gaussian to deck numbers for all passengers.

Missing values for deck cabin

As previously shown in figure 11, there is a tendency for increased survival for a subrange of the cabin numbers, although survival is generally very intermingled for this attribute, as shown in figure 19.

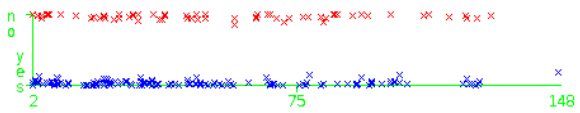

Figure 19: Deducing survival trends based on cabin numbers is messy business.

To complicate things further, the allocation of cabin numbers do not follow the longitudinal axis of the ship in a strict sense and the individual decks vary wildly in layout, with some having sections dedicated to various types of common or functional areas. Figures 20 through 33 shows attempts at cracking this by binning up and fitting a Gaussian to each deck (again branching on gender as a pseudo-survival indicator).

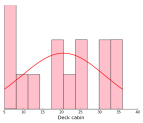

Figure 20: Deck A.

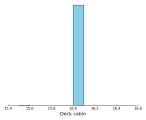

Figure 21: Deck A

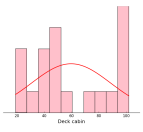

Figure 22: Deck B.

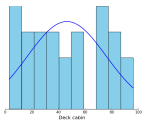

Figure 23: Deck B

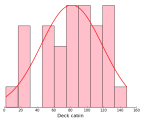

Figure 24: Deck C.

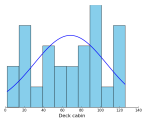

Figure 25: Deck C

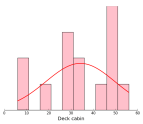

Figure 26: Deck D.

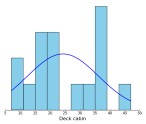

Figure 27: Deck D

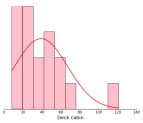

Figure 28: Deck E.

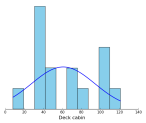

Figure 29: Deck E

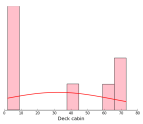

Figure 30: Deck F.

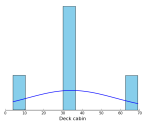

Figure 31: Deck F

Figure 32: Deck G.

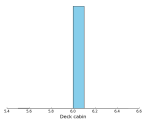

Figure 33: Deck G

However, the best, although not perfect, generalization appears when all decks are combined, as shown in figures 34 and 35.

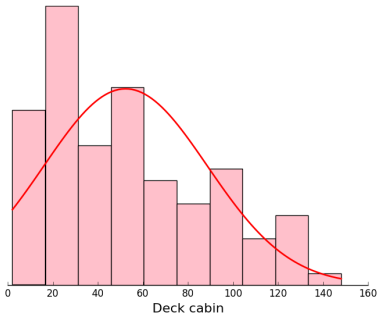

Figure 34: Binning and fitting a Gaussian for man (all decks).

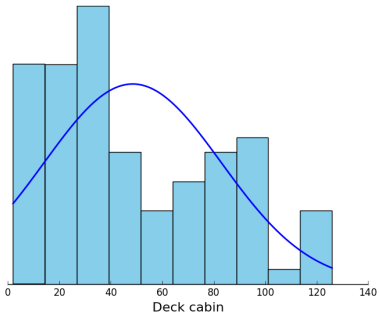

Figure 35: Binning and fitting a Gaussian for females (all decks).

From these we get some resemblance of a bell curve, although the peak could appear to be skewed towards the lower cabin numbers, i.e. the bow (front) of ship. We will explain and accommodate this by assuming that the impact was felt and heard most violently at the general area of impact, the bow, provoking most fear amongst these passengers and higher tendency to flee, compared to those at the stern. A better fitted Gaussian may therefore arise by visually determining some mean and assuming an "imaginary" extension of the bow and passengers, with the left side of the mean symmetric to the right half (of which we have data). Assuming visually estimated means and reflecting the data (as the "left half" of a bell curve), we get the resulting Gaussian fittings shown in figures 36 and 37, both of which appear to capture the nature of the data better.

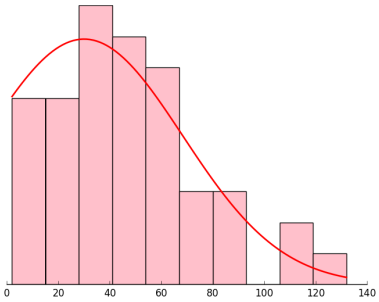

Figure 36: Binning and fitting a Gaussian on men for all decks, assuming a mean of 30.

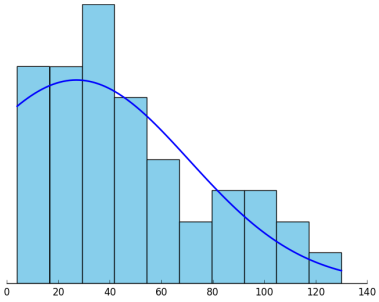

Figure 37: Binning and fitting a Gaussian on females for all decks, assuming a mean of 27.

Our model for filling in missing deck cabin numbers is therefore obtained by sampling positive values from one of these two models, as dictated by gender.

Missing values for fare

A single fare value is missing in the test data. Under the assumption that a single observation has little impact, it will simple be filled in by the mean of the fares in the set.

7 Implementation

Technology Choices

The model was implemented in Python using *Numpy* and *Panda*. A conversion from Panda frames to the Weka ARFF format allowed for using Weka's excellent inspection tools, to get an overview of the data.

For training and classification, a random forest was used from the *scikit-learn* library since, although rendering our search for missing values less relevant, they are fast and was assumed to be of low risk to the desire for undertaking computational heavy parameter sweep and cross-fold validation operations. Furthermore, they are robust to outliers, of which no action was taken to guard against here. In addition, it was assumed that the several hundred observations we have for each data set is adequate for the 'randomization' to take affect.

Optimizing accuracy

An optimal *n_estimators* parameter value (the number of trees in the forest) was determined by sweeping through the range 1-300 and picking the parameter that gave the greatest accuracy. For each sweep, accuracy was determined using the mean average precision of a ten-fold cross-validation: the training set was divided into blocks of ten, each tested by a model obtained from training on the remaning 9 blocks (hence, 10 rounds in total, each yielding $1/10$ contribution to the final accuracy of the sweep).

8 Results

Train accuracy

For the parameter sweeping of the training set, an optimal *n_estimators* parameter value of 25 was obtained along with mean average precisions, as shown in figure 38. It is seen that the solution found is most likely a fluctuation spike around a trend that could appear to still be rising with increasing parameter value. Hence, a slightly better parameter could perhaps be found by investigating minima over averaged data points.

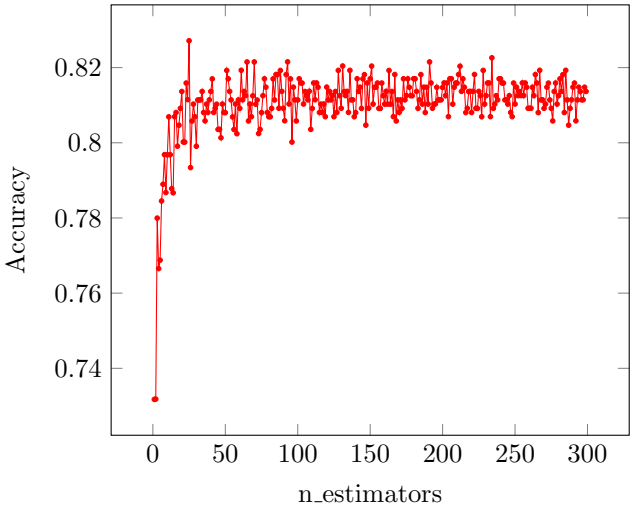

Figure 38: Parameter sweep accuracies.

Test accuracy

For the accuracy of the submission, the Kaggle leaderboard reports 0.75120 (username "RuneJensen"). It it seen that the test accuracy is slightly lower than the train cross-fold accuracy, which was anticipated given that the model was evaluated on unseen data that is likely to have slightly different variations of the general data patterns.

References

[1] Christopher M. Bishop, Pattern Recognition and Machine Learning, 2006, ISBN-10: 0-387-31073-8

[2] Andrew Ngs notes