Titanic: Machine Learning from Disaster(ELEN4009A INDIVIDUAL PROJECT)

Tristan Lilford (1843691)

School of Electrical & Information Engineering, University of the Witwatersrand, Private Bag 3, 2050, Johannesburg, South Africa

Abstract: The design and implementation of machine learning to create a model that predicts which passengers survived the Titanic shipwreck. The final solution used feature engineering, model tuning and ensembling to result in a voting clssifier which used a random forest model in conjunction with logistic regression to obtain an accuracy of 77.5%

Key words: Feature Engineering, Random Forest, Log Regression, voting classifier

1. INTRODUCTION

Machine learning is an application of Ai where a system attempts to learn and predict outcomes[1]. These systems do this by using a training data set to effectively train a model which is then able to use test/input data to produce a result. In order to achieve higher accuracy with these results there are methodologies which can be applied, namely, Feature engineering, model selection, model tuning and ensembling[2]. This report covers the design of a machine learning solution to the well known Titanic ML competition provided by Kaggle[3]. Firstly the Background of the task will be provided in Section 2. Then Section 3 will cover the specific design of the final solution. Section 4 will provide a discussion of the results. Section 5 will provide possible recommendations for a more efficient approach. Lastly a conclusion will be provided to summarize the entirety of the project. The project workload was divided amongst 6 team members and as such this report will focus on the data analysis (section 3.2), feature engineering (Section 3.3) and model tuning (Section 3.5) as the delegated areas of work. Firstly the background to task must be understood.

2. BACKGROUND

2.1 *Kaggle Titanic ML Competition*

The task was to successfully determine whether and individual, with predetermined data, would of survived the Titanic Shipwreck through the use of machine learning. To do so Kaggle provided 2 important data sets, the training data set and the testing data set. The training data was to be used to effectively train the model selection so that it would be able to make a prediction. The layout of the data can be seen in Figure 9 in Appendix A, with Figure 10 acting as a descriptor of the data. The test data set was to be used by the trained model to make a prediction on whether those individuals survived. This data set structure can be seen in Figure 11 in Appendix A and it is noted that there is no survived column as this the prediction that has been tasked. Once a system is designed the predicted results can then be uploaded to Kaggle where the overall accuracy of the system is determined by comparing the predicted results to the known survival results. The Kaggle competition aims to attempt to see who can achieve the greatest accuracy rates using their designed ML systems. In order to understand this process a brief overview of machine learning is provided.

2.2 *Machine Learning*

Machine learning falls within the realm of data analytics and uses computational methods to "learn" information directly from data without relying on a predetermined equation as a model[4]. It does this by finding natural patterns within its training data These patterns provide the system insight when making predictions. The way these systems generate these insights is based on the model used. Three models which provided the greatest accuracy for this task were the Random Forest Model, KNN and Logistics Regression. We must first understand the workings of a decision tree algorithm before understanding the Random Forest Model.

2.2.1 Decision Tree: The Decision Tree algorithm falls within the family of supervised learning algorithms. It is often used to solve regression and classification problems. It is able to effectively predict a value by learning simple decision rules obtained from training data. Decision trees determine outcomes by effectively branching whenever a decision is made [5]. This begins at the root decision where a comparison is made with the root attribute and the recorded input. Based on the result of the comparison the corresponding branch for the result will be followed to the next decision node until a final prediction is determined [6].

2.2.2 Random Forest: Random forest is also a supervised learning algorithm. Effectively it creates a forest using an ensmeble of decision trees and is often trained using a "bagging" method. Effectively it creates multiple decision treas which work with one another in order to achieve a greater accuracy even with minimal hyperparameter tuning[7].

2.2.3 Logistics Regression: Logistic Regression is also used for regression and classification problems. It is used to predict the outcome of a class which only has two possible values [8], such as whether someone lived or died on the Titanic. It operates by using a sigmoid function to map the predicted values to probabilities[9]. The Logistic regression equation can be obtained from the Linear Regression equation and results in Equation 1 below.

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_m X_m \tag{1}$$

2.2.4 k- Nearest Neighbors Classifier: KNN is a classification algorithm that plots training data points and determines the labels of an input data point by plotting it and comparing it to its k nearest neighbours. The majority of a data points k nearest neighbours is then used to make a classification [10].

With this basic information it is clear to see that there are two important factors which will effect the accuracy of the system. Namely the quality and quantity of the training data set, as well as how effective the chosen model operates with the specified training data set. Section 3.1 clarifies the methods used in order to generate the best data set and chose the most effective model.

2.3 Team Structure and Delegation

In order to tackle this task a team of 6 was formed. This included Tristan Lilford, Tristan de Groot, Tristan Baesal, Robin Jonker, Van Niekerk Ferreira and Jesse Iyuke. They were delegated sections as follows.

- Tristan Lilford: Data analysis, feature engineering and model tuning
- Tristan de Groot: Data analysis, structure and layout and Ensembling
- Tristan Basel: Model building, model tuning and graphical analysis
- Robin Jonker: Team lead, project oversight, involved in all sections
- Van Niekerk Ferreira: Feature engineering and model building
- Jesse Iyuke: Feature engineering, Cross-validation and model building

The team worked effectively with one another to develop the system as described in section 3 below

3. DESIGN AND IMPLEMENTATION

The most relevant figures and values have been include in text, any less relevant data has been attached within Appendix B

3.1 General Approach

In order to achieve the highest accuracy we would want to have the highest quality of of data and the most suitable model. To do this an understanding of the provided training data must be acquired. To develop this system the following process is followed:

- Data Analysis
- Feature Engineering
- Data Preprocessing
- Model Building
- Model Tuning
- Ensembling

3.2 Data Analysis

Data analysis is all about processing the raw input data into useful information which can be understood and interpreted. With regards to the training data seen in Figure 9 in Appendix A we wish to determine correlations between age, sex, number of siblings... and survival rates. By developing graphical models the data can be best understood. First we assess the quality of data to determine if there are any missing values which may hinder the quality of the data. To do this we determine the number of NaN values in the data set. Table 1 below shows the results of this process.

Table 1: Number of Missing data values per column

	Feature	Missing Values
1	Cabin	687
2	Age	177
3	Embarked	2
4	PassengerId	0
5	Survived	0
6	Pclass	0
7	Name	0
8	Sex	0
9	SibSp	0
10	Parch	0
11	Ticket	0
12	Fare	0

We can see here that in the total 891 records 77% of cabin data and roughly 20% of age data is missing. These issues will be dealt within feature engineering. Outliers also need to be determined to see if any possible data is skewing the information. These outliers can be seen in Figure 1 below.

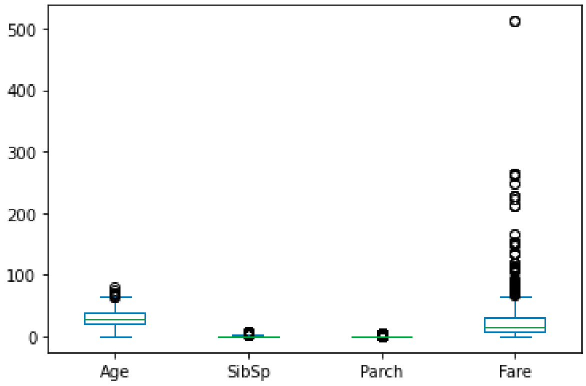


Figure 1: Outliers

There are 2 kinds of data within the specified training data, categorical data and numerical data.

3.2.1 Categorical Data: Categorical data includes survived, sex, pclass and embarked. The most note worthy of theses data sets is the correlation between a persons sex and survival seen inf Figure 2. Another important correlation is the level of the class they were in and their survival.

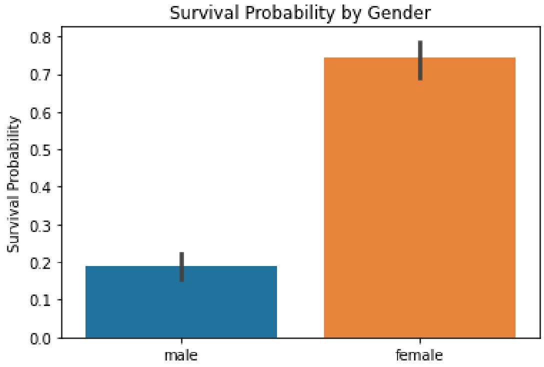


Figure 2: Gender vs Survival

It is clear to sea that there is a high correlation between being female and surviving. This is expected as females would be prioritized with regards to lifeboats. Next class vs survival is studied in Figure 3.

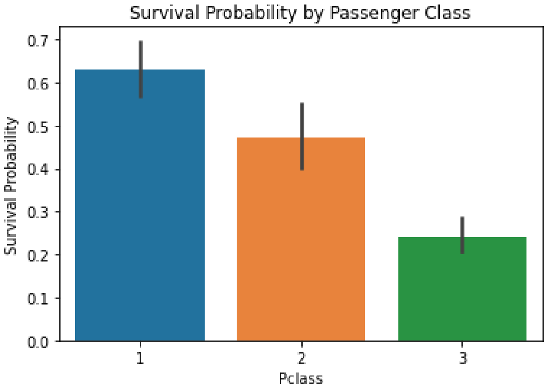


Figure 3: Class vs Survival

It is clear to see that the higher your class was, such as 1st class, the better your chances of survival were.

3.2.2 Numerical Data: First a study of the age correlation with survival rates is seen below in figure 4.

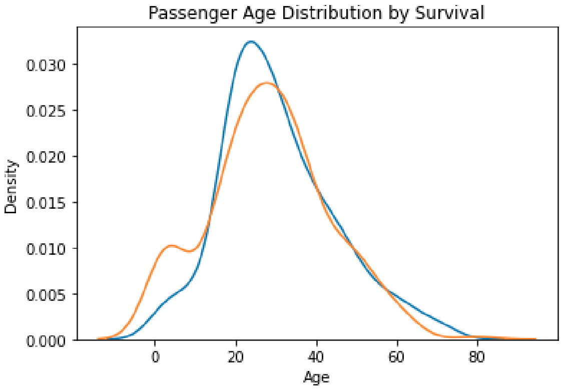


Figure 4: Age and survival

It is seen that survival is not heavily dependant on age except noticing a slightly higher survival rate for children between the ages 0 and 10. This kind of data is represented with numerical values. This means correlations and covariance between features can be determined. Covariance, is a measure of how two random assets jointly vary [11]. It measures how these features change with regards to one another and can be seen in Figure 5 below.

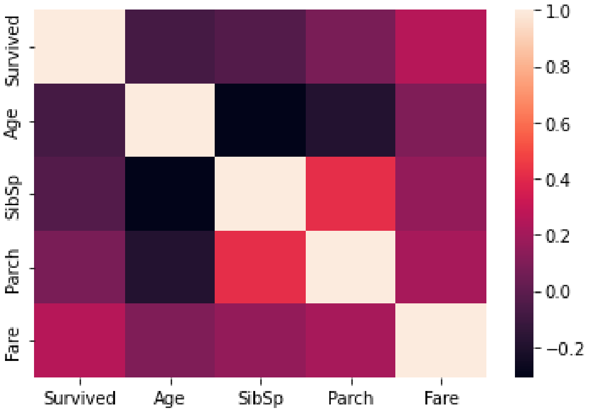


Figure 5: Covariance matrix

This is useful in determining which features most strongly vary with survival rates and interestingly the number of sibling/spouses and or parents had a strong correlation with surviving. Note categorical values such as sex cannot be used in this data analysis method. Now that the data is understood it must be put in a manner that a machine learning algorithm can best understand, this is done with feature engineering.

3.3 Feature Engineering

Feature engineering includes two main processes. One, is to develop new features and to better describe current features. The other is to chose the best combination of features which yield the highest accuracy. The first feature to be studied would be simply removing all features which contain NaN values. The accuracy of these features could be tested within the model building section to determine their efficacy (described briefly in Section 3.4.

3.3.1 Removing NaN values: This is done by sampling and dropping the columns/records that contained NaN values. This feature simply removed the features which included nan values which was cabin, age and embarked. Having the models determine the accuracy rate of this feature it was found to achieve roughly a 75% accuracy which is rather high.

3.3.2 Inputting Average Age: This feature looked to fill in the missing age data with the calculated average. Other methods such as using the median or using a constant were also practiced however the mean produced the greatest accuracy results.

3.3.3 Assuming Cabin Based on Class: This section looked to replace NaN cabin values within data based on which class they were in. If the diagram of the titanic in Appendix C is analysed, it is seen that certain cabins submerged before others, this would have a great effect survival rates. However, as stated early 77% of this data is missing and when accuracy was determined using this feature it seemed to drop the overall performance of the algorithm and as such was not used.

3.3.4 Title Extraction: This feature looked to extract a passengers title from their name. Titles included categories like Dr, Sir, Countess... It was seen that individuals with higher ranking titles had far higher survival chances, this is seen in Figure 12 in Appendix B. This feature was thought to have added a fair amount of accuracy, however it only improved accuracy with a select group of algorithms which were not used.

3.3.5 Best Feature Combination: After many attempts of mixing many features, a lot which have not been mentioned, the highest accuracy combination was found. It included whether they were alone, their age (mean filled in if absent), their class and their sex.

3.4 Model Building

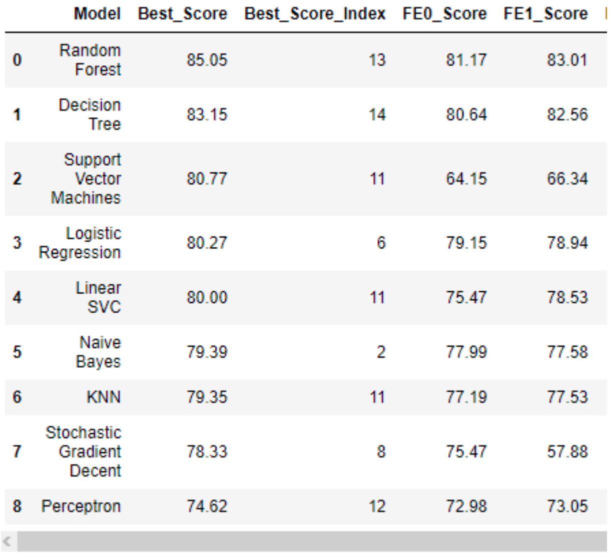
Model building was headed by Tristan Baesal. Various models were tested including Random Forest, Logistic Regression, Support Vector Machine, KNN, Gaussian Naive Bayes, Perception, Linear SVC, Stochastic Gradient Descent and Decision Tree.

The training set is first applied to models, after that the models can be applied to the test set in order to obtain accuracy results. Firstly, the provided feature engineered training set is split into sections for training and validation. These will be used to determine the overall performance of that feature with that specific model. Overfitting is the process by which the Machine Learning algorithm effectively learns too much from the training set which creates noise and lowers the overall accuracy. This was seen when including the title feature. The specific results can be seen below in Table 2.

Table 2: Model Accuracy check against Feature Engineering

Model	**Best Score**	**Best FE**
Random Forest	85.05	13
Decision Tree	83.27	14
Support Vector Machine	80.77	11
Logistic Regression	80.27	6
Linear SVC	80.00	11
Naive Bayes	79.39	2
KNN	79.35	11
Stochastic Gradient Descent	79.30	11
Perception	74.62	12

The summarised results can be seen below in Figure 6.



	Model	Best_Score	Best_Score_Index	FE0_Score	FE1_Score
0	Random Forest	85.05	13	81.17	83.01
1	Decision Tree	83.15	14	80.64	82.56
2	Support Vector Machines	80.77	11	64.15	66.34
3	Logistic Regression	80.27	6	79.15	78.94
4	Linear SVC	80.00	11	75.47	78.53
5	Naive Bayes	79.39	2	77.99	77.58
6	KNN	79.35	11	77.19	77.53
7	Stochastic Gradient Decent	78.33	8	75.47	57.88
8	Perceptron	74.62	12	72.98	73.05

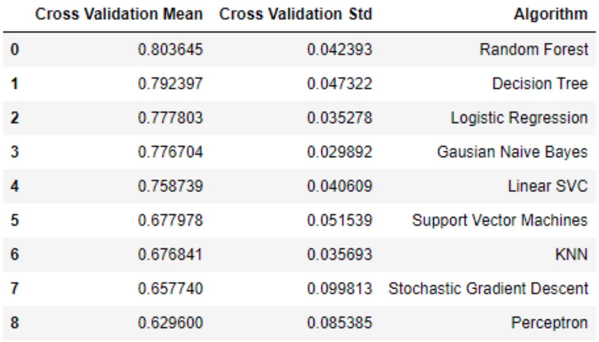
Figure 6: Models accuracies per feature

KNN, Random Forest and Logistic Regression returned the most reliable accuraccies and were to be used moving forward into model tuning and ensembling.

3.5 Model Tuning

From Figure 6 above, it can be seen that Random Forest produced the most accurate results. Feature 13(fe13) had the most accurate results for this mode however, feature engineering 6, 12 and 14 were chosen to be carried further as they were the most accurate across multiple models and these results after tuning and ensembling would be compared again to that of using just the Random Forest Model.

The high training accuracy is not realistic, as it does not represent how the system will treat actual test data. There is most likely high bias within each model and as such tuning should be done to increase overall accuracy. K-fold cross-validation uses a resampling procedure to evaluate machine learning models on a limited data sample. It was seen after this procedure certain models' accuracies dropped including logistic regression. The highest accuracy results would be used for model tuning, these results can be seen below in Figure 7



	Cross Validation Mean	Cross Validation Std	Algorithm
0	0.803645	0.042393	Random Forest
1	0.792397	0.047322	Decision Tree
2	0.777803	0.035278	Logistic Regression
3	0.776704	0.029892	Gausian Naive Bayes
4	0.758739	0.040609	Linear SVC
5	0.677978	0.051539	Support Vector Machines
6	0.676841	0.035693	KNN
7	0.657740	0.099813	Stochastic Gradient Descent
8	0.629600	0.085385	Perceptron

Figure 7: K-fold cross validation

Random Forest, Log Regression and KNN were tuned. Model tuning is also known as hyperparameter optimisation. Hyperparameters are variables which control the training of an algorithm by managing weights features. Model tuning provides optimized values for hyperparameters, which maximize your model's predictive accuracy[12]. The most weighted parameters for the selected features can be seen in Figure 8 below for Random Forest.

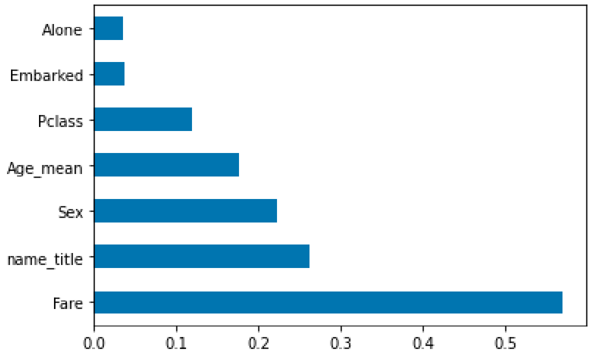


Figure 8: Hyper-parameterisation for RF

3.6 Ensembling

There were a few methods of ensembling, however the voting classifier returned the greatest results. The voting classifier effectively combines different machine learning algorithms and uses the average predicted probabilities (soft vote) to determine a class value. In this case Random Forest and Logistic Regression were used together to determine the survival of an individual. KNN was found to in fact lower accuracy in ensembling and so it was removed from the selection. This classifier works well if the models perform relatively similar. The intent is to balance out each algorithms weakness. This method results in our Best Prediction with a public score of 0.775.

The soft voting method was been implemented as:

$$\hat{y} = \arg\max_i \sum^m j = 1wjpij,$$

where wj is the weight that can be assigned to the jth classifier.

3.7 Model Submission and Results

With all the previous steps completed the test data can be fully processed. the results of the test where generated in a csv file and uploaded to Kaggle. The highest accuracy of the group was recorded at 75.5% using the voting classifier as described above. This placed the submission as 7953 out of 14,118. This result was deemed acceptable however there were some ideas that would be noted to improve performance which will be discussed in Section 5

4. ANALYSIS AND DISCUSSION

The results from the feature engineering were not as great as expected. Within model building validation the accuracies found where very dependant with feature engineering. The final Deep Forest algorithm was found to have an accuracy of 84%, 78% and 84% for the respective features fe6, fe13 and fe14 that were used in tuning. Finding a better combination of features is thought to be able to improve accuracy by a greater amount. Ensembling was difficult to manage. Attempts often resulted in an overall decrease in accuracy. The approach of being able to test and document multiple features and models simultaneously was very effective in choosing a final design. An overall rating of 7953 placed the solution in the top 56th percentile. There are some improvements which could be made as discussed below.

5. FUTURE RECOMMENDATIONS

The use of a Convolutional Neural Network, should be explored. This method makes use of various node layers in order to make predictions. Each node is linked and have various thresholds and weights. When processing, if a threshold is met the node is set to active. When this occurs the data proceeds to the next layer of nodes. There are three layers in this process, a convultion layer, a pooling layer and a fully connected layer. The convolution layer applies a filter to the input to create a feature map which is able to summarizes the presence of detected features in the input[13]. The pooling layer effectively slides a two-dimensional filter over each channel of the feature map and summarises the features within this region. Lastly, the fully connected layer applies labels to the input data based on what is received from the previous two layers.

6. CONCLUSION

An effective machine learning solution for the Kaggle Titanic ML Competition was created. It used concepts of feature engineering using detailed data analytics as a guideline. Multiple different models were built and tested in order to determine the most effective algorithms. Hyperparameters were then set using effective model tuning. Ensembling was then used to combine the Random Forest with Logistic Regression in order to achieve the teams most accurate solution. Overall the solution placed 7953. It is recommended to investigate the promise of Convolutional Neural Networks to attempt to improve accuracy. Overall the system was informative and educational.

7. References

[1] Expert.AI Team (2020). What is Machine Learning? A definition - Expert System. [online] Expert.ai. Available at: https://www.expert.ai/blog/machine-learning-definition/.

[2]Simplilearn.com. (n.d.). The Complete Guide to Machine Learning Steps. [online] Available at: https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps.

[3] kaggle.com. (n.d.). Titanic: Machine Learning from Disaster. [online] Available at: https://www.kaggle.com/c/titanic.

[4] www.mathworks.com. (n.d.). What Is Machine Learning? — How It Works, Techniques Applications. [online] Available at: https://www.mathworks.com/discovery/machine-learning.html: :text=Machine%20learning%20is%20a%20data.

[5]KDnuggets. (n.d.). Random Forest® vs Decision Tree: Key Differences. [online] Available at: https://www.kdnuggets.com/2022/02/random-forest-decision-tree-key-differences.html: :text=The%20critical%20difference%20between%20the.

[7]Donges, N. (2021). A Complete Guide to the Random Forest Algorithm. [online] Built in. Available at: https://builtin.com/data-science/random-forest-algorithm.

[8]www.tutorialspoint.com. (n.d.). Machine Learning - Logistic Regression - Tutorialspoint. [online] Available at: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_logistic$_r$*egression.htm* : : *text* = *Logistic*%20*regression*%20*is*%20*a*%20*supervised*.

[9]www.javatpoint.com. (n.d.). Logistic Regression in Machine Learning - Javatpoint. [online] Available at: https://www.javatpoint.com/logistic-regression-in-machine-learning.

[10]S. Ray. "Commonly used Machine Learning Algorithms (with Python and R Codes).",2017. URL https://www.analyticsvidhya.com/ blog/2017/09/common-machine-learningalgorithms/. [Online; accessed 27-October-2021].

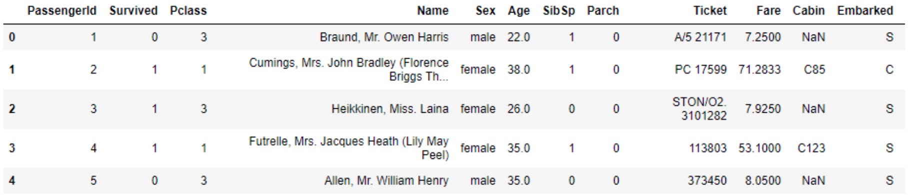
[11]Understanding the Covariance Matrix — DataScience+. 2022. Understanding the Covariance Matrix — DataScience+. [ONLINE] Available at: https://datascienceplus.com/understanding-the-covariance- matrix/. [Accessed 09 May 2022].

[12]www.mlexam.com. (2021). Model tuning - ML exam study guide. [online] Available at: https://www.mlexam.com/model-tuning/: :text=Model

[13]Jason Brownlee (2019). How Do Convolutional Layers Work in Deep Learning Neural Networks? [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/.

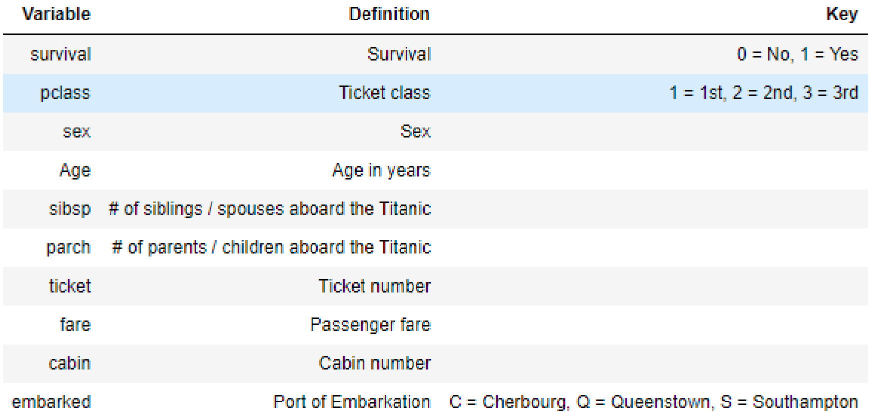
Appendices

A Appendix



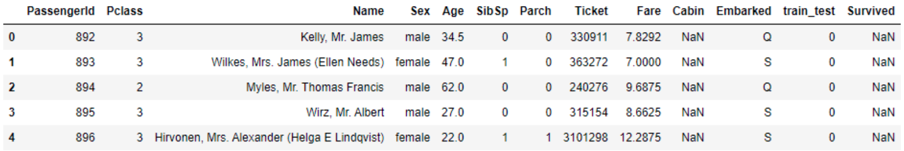
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 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 9: Training Data Set



Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

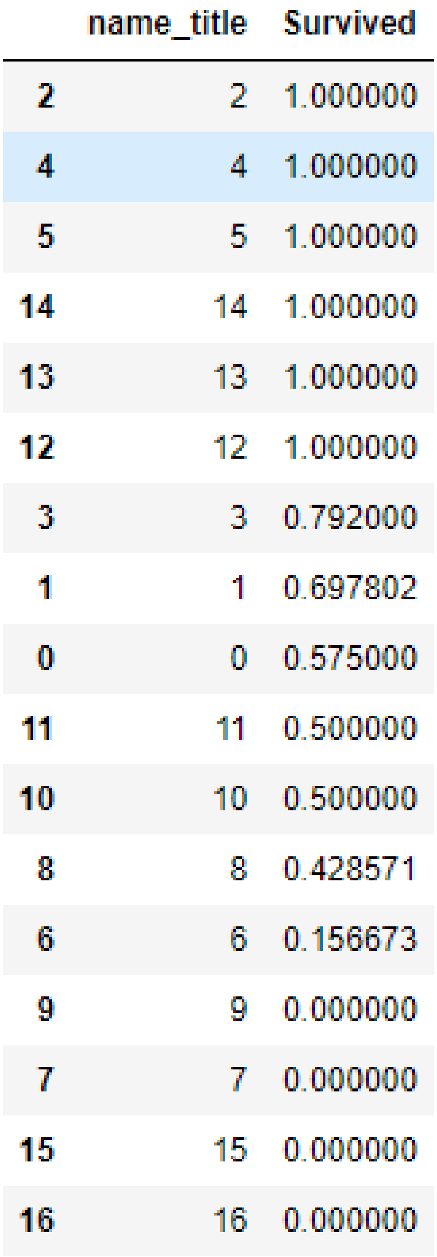
Figure 10: Data Descriptions



	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	train_test	Survived
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q	0	NaN
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S	0	NaN
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q	0	NaN
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S	0	NaN
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S	0	NaN

Figure 11: Testing Data Set

B Appendix



	name_title	Survived
2	2	1.000000
4	4	1.000000
5	5	1.000000
14	14	1.000000
13	13	1.000000
12	12	1.000000
3	3	0.792000
1	1	0.697802
0	0	0.575000
11	11	0.500000
10	10	0.500000
8	8	0.428571
6	6	0.156673
9	9	0.000000
7	7	0.000000
15	15	0.000000
16	16	0.000000

Figure 12: Titles vs Survival

Data distribution:

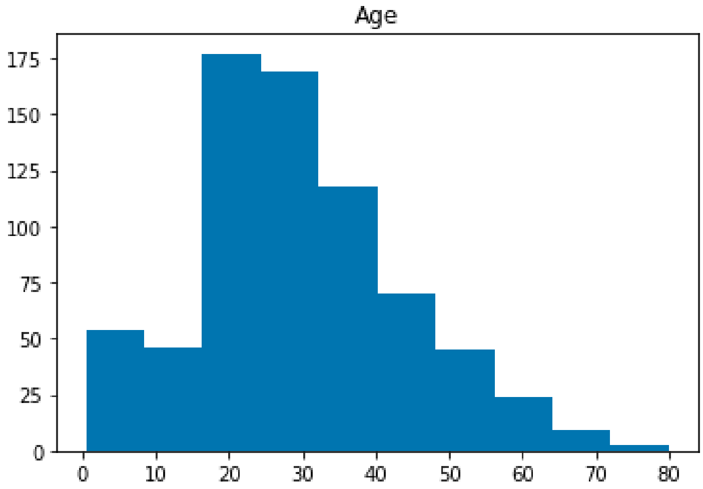


Figure 13: Age Distribution

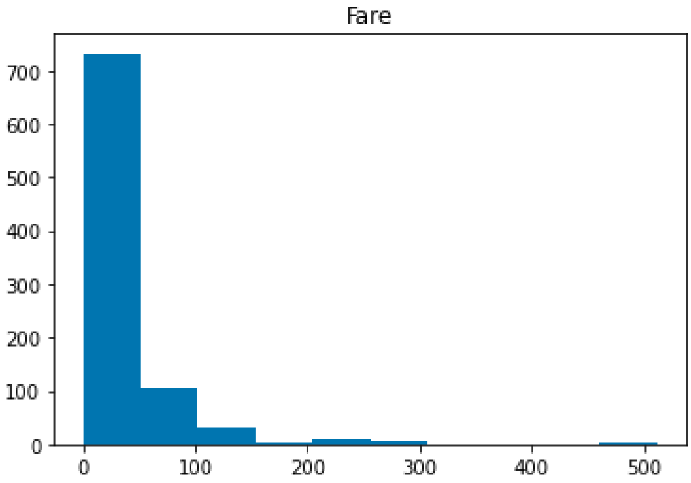


Figure 14: Fare Distribution

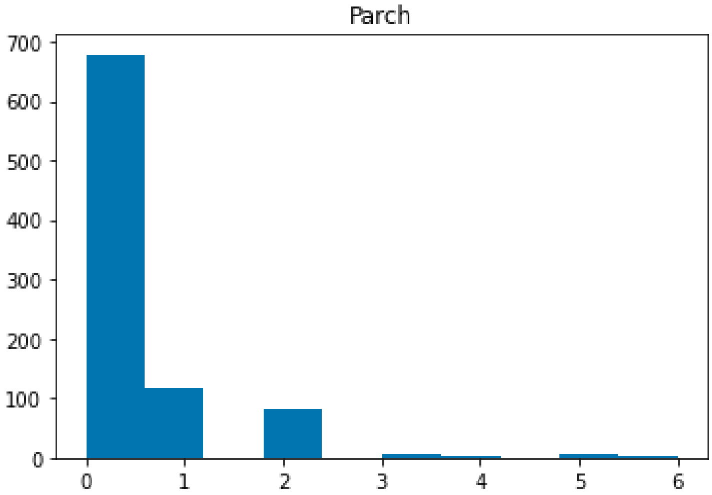


Figure 15: Parents/Children Distribution

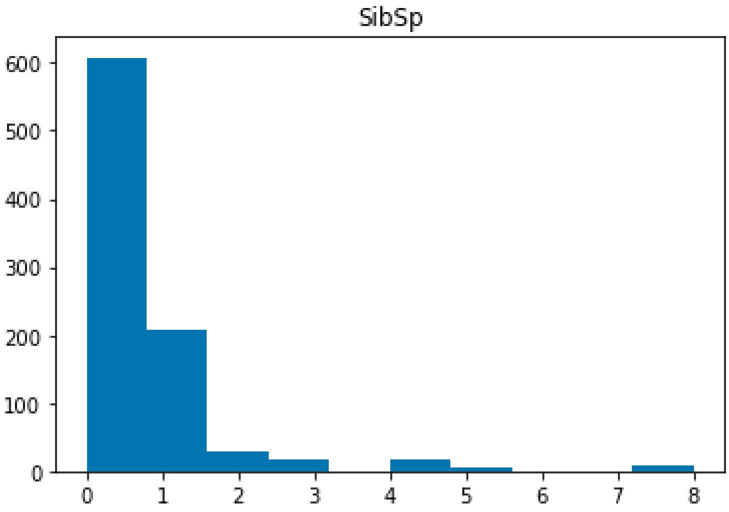
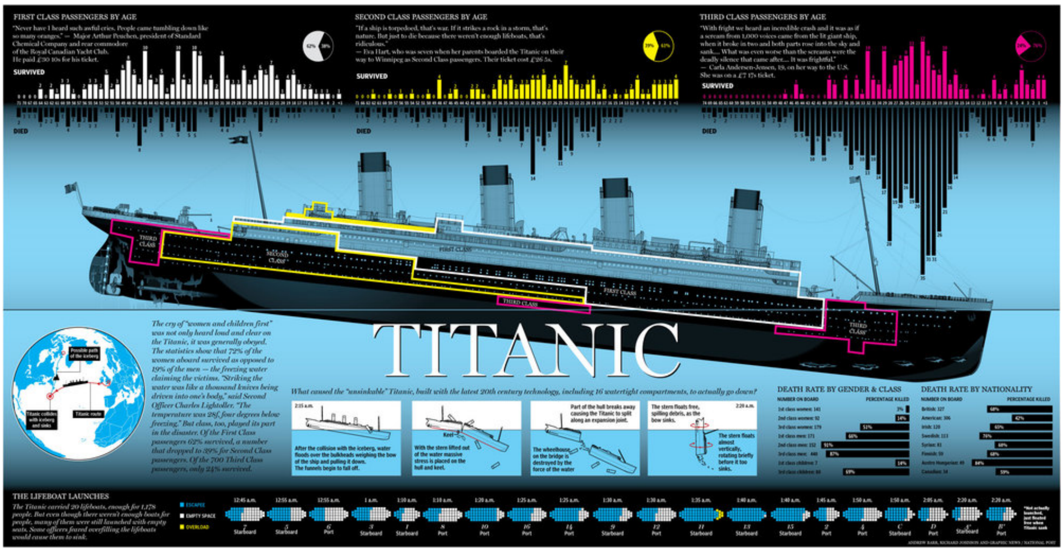


Figure 16: Sibling/Spouse Distribution

C Appendix



"Never have I heard such awful cries. People came tumbling down like so many oranges." — Major Arthur Peuchen, president of Standard Chemical Company and rear commodore of the Royal Canadian Yacht Club. He paid £30 10s for his ticket.

"If a ship is torpedoed, that's war. If it strikes a rock in a storm, that's nature. But just to die because there weren't enough lifeboats, that's ridiculous."
— Eva Hart, who was seven when her parents boarded the Titanic on their way to Winnipeg as Second Class passengers. Their ticket cost £26 5s.

"With fright we heard an incredible crash and it was as if a scream from 1,000 voices came from the lit giant ship, when it broke in two and both parts rose into the sky and sank.... What was even worse than the screams were the deadly silence that came after.... It was frightful."
— Carla Andersen-Jensen, 19, on her way to the U.S. She was on a £7 17s ticket.

The cry of "women and children first" was not only heard loud and clear on the Titanic, it was generally obeyed. The statistics show that 72% of the women aboard survived as opposed to 19% of the men — the freezing water claiming the victims. "Striking the water was like a thousand knives being driven into one's body," said Second Officer Charles Lightoller. "The temperature was 28f, four degrees below freezing." But class, too, played its part in the disaster. Of the First Class passengers 62% survived, a number that dropped to 39% for Second Class passengers. Of the 700 Third Class passengers, only 24% survived.

What caused the "unsinkable" Titanic, built with the latest 20th century technology, including 16 watertight compartments, to actually go down?

After the collision with the iceberg, water floods over the bulkheads weighing the bow of the ship and pulling it down. The funnels begin to fall off.

With the stern lifted out of the water massive stress is placed on the hull and keel.

Part of the hull breaks away causing the Titanic to split along an expansion joint.

The wheelhouse on the bridge is destroyed by the force of the water

The stern floats free, spilling debris, as the bow sinks.

The stern floats almost vertically, rotating briefly before it too sinks.

The Titanic carried 20 lifeboats, enough for 1,178 people. But even though there weren't enough boats for people, many of them were still launched with empty seats. Some officers feared overfilling the lifeboats would cause them to sink.

Figure 17: Titanics sinking information