Titanic Survivor Predictions

Diogo Oliveira, Joshua Jones, Leandro Matos

Rua de São Tomé 79, Braga, 4705-578, Portugal
Rua Quinta dos Lagos 12, Braga, 4700-289, Portugal
Travessa da Vieirinha 54, Barcelos, 4700-289, Portugal

Abstract

Click here and insert your abstract text.

Keywords: Type your keywords here, separated by semicolons ;

Nomenclature

A	AI (Artificial Intelligence)
B	IPCA (Instituto Politécnico do Cávado e do Ave)
C	ML (Machine Learning)
D	OW (Orange Workflow)
E	ANN (Artificial Neural Network)

Contextualization

Our project is about the Titanic disaster, for better context we need to understand what happened.

The Titanic disaster of 1912 remains one of the most famous shipwrecks in history, claiming the lives of over 1,500 passengers and crew. In this report, we will be using the Orange Workflow to study a dataset about the Titanic survivors, with the goal of understanding the factors that may have influenced their chances of survival.

OW is a data visualization and analysis tool that allows users to work with various data types, including tabular data. By using Orange, we can easily load, visualize, and analyse the Titanic dataset, which includes information such as passenger demographics, ticket class, and cabin number.

The first step in our analysis will be to load the Titanic dataset into Orange and explore the data to get a general understanding of the distribution of variables and the relationships between them. Next, we will use Orange's data pre-processing tools to clean and prepare the data for analysis. Once the data has been prepared, we will use Orange's Machine Learning capabilities to train and evaluate models that predict the chances of survival based on the numerous factors in the dataset. This will allow us to identify which factors had the greatest impact on survival and understand how they interacted with each other.

Finally, we will use Orange's visualization tools to present the results of our analysis in an easily interpretable way. The results of this analysis can be used to gain insight into the factors that influenced the survival of passengers on the Titanic and can help inform future research on the topic.

Overall, this report demonstrates the power of Orange to analyse large datasets quickly and easily, providing valuable insights into complex problems. In this particular case, we used the Orange workflow to study a dataset about the Titanic survivors, and understand how different factors such as age, gender, and class, among others, affected their chances of survival.

Please cite this article as: First author et al., Article title, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.00.000

Art Study

Machine learning is a branch of artificial intelligence that involves training algorithms to automatically learn from data and make predictions. The Orange Workflow, which is a popular data visualization and analysis tool, provides an easy-to-use interface for working with machine learning algorithms, including loading and preprocessing data, training models, and visualizing results.

When applied to the Titanic disaster, machine learning and the Orange Workflow can be used to analyze historical data and identify patterns that may have contributed to the disaster.

By using Machine Learning algorithms to predict the chances of survival based on factors such as passenger demographics and ticket class, we can gain insight into how these factors influenced the outcome of the disaster. This can help us identify the most important factors that contributed to the high death toll and inform future research on the topic.

One of the main advantages of using machine learning and the OW is that it allows us to analyze large amounts of data quickly and easily. This can help us identify patterns and relationships that might otherwise be difficult to spot, providing valuable insights that can help inform decision-making and prevent similar disasters from happening in the future.

Machine Learning Types

Supervised learning

In supervised learning, the machine is taught by example. The operator provides the machine learning algorithm with a known dataset that includes desired inputs and outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs. While the operator knows the correct answers to the problem, the algorithm identifies patterns in data, learns from observations and makes predictions. The algorithm makes predictions and is corrected by the operator – and this process continues until the algorithm achieves a high level of accuracy/performance.

Under the umbrella of supervised learning fall: Classification and Regression.

1. **Classification:** In classification tasks, the machine learning program must draw a conclusion from observed values and determine to
what category new observations belong. For example, when filtering emails as ‘spam’ or ‘not spam’, the program must look at existing observational data and filter the emails accordingly.

2. **Regression:** In regression tasks, the machine learning program must estimate – and understand – the relationships among variables. Regression analysis focuses on one dependent variable and a series of other changing variables – making it particularly useful for prediction and forecasting.

Reinforcement learning

Reinforcement learning focuses on regimented learning processes, where a machine learning algorithm is provided with a set of actions, parameters, and end values. By defining the rules, the machine learning algorithm then tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal. Reinforcement learning teaches the machine trial and error. It learns from past experiences and begins to adapt its approach in response to the situation to achieve the best possible result.

For this project we used the following ML types:

- Artificial Neural Networks (Reinforcement Learning)
  An Artificial Neural Network (ANN) comprises 'units' arranged in a series of layers, each of which connects to layers on either side. ANNs are inspired by biological systems, such as the brain, and how they process information. ANNs are essentially many interconnected processing elements, working in unison to solve specific problems.
  ANNs also learn by example and through experience, and they are extremely useful for modelling non-linear relationships in high-dimensional data or where the relationship amongst the input variables is difficult to understand.

- Logistic Regression (Supervised learning – Classification)
  Logistic Regression focuses on estimating the probability of an event occurring based on the previous data provided. It is used to cover a binary dependent variable, that is where only two values, 0 and 1, represent outcomes.

- Decision Trees (Supervised Learning – Classification/Regression)
  A Decision Tree is a flow-chart-like tree structure that uses a branching method to illustrate every possible outcome of a decision. Each node within the tree represents a test on a specific variable – and each branch is the outcome of that test.

- Naïve Bayes Classifier Algorithm (Supervised Learning - Classification)
  The Naïve Bayes Classifier is based on Bayes' theorem and classifies every value as independent of any other value. It allows us to predict a class/category, based on a given set of features, using probability.

- Random Forests (Supervised Learning – Classification/Regression)
  Random Forests or 'Random Decision Forests' is an ensemble learning method, combining multiple algorithms to generate better results for classification, regression and other tasks. Each individual classifier is weak, but when combined with others, can produce excellent results. The algorithm starts with a 'decision tree' (a tree-like graph or model of decisions) and an input is entered at the top. It then travels down the tree, with data being segmented into smaller and smaller sets, based on specific variables.

Problem Solving

To solve a Titanic Survivor Prediction problem using Orange Workflow, we had to load the Titanic dataset into Orange. This was done by using the File widget to import the data from a TAB file.

To accomplish this, we adapted the dataset, halving the passenger number to 1,146 passengers, including their demographics and ticket class. We then used Orange Workflow to load, visualize, and preprocess the data, before applying several machine learning algorithms, including Neural Networks, Logistic Regression, Decision Trees, Naive Bayes, and Random Forests. These algorithms were used to predict the survival rates of passengers based on the different factors in the dataset.

The network was trained to learn the relationships between these factors and the chances of survival for each passenger. Once the network was trained, it was used to make predictions on the survival rates of the passengers in the dataset, based on the factors it had learned.

Neural Networks are particularly useful in this case because they are able to learn and adapt to complex, non-linear relationships between the input and output data. They are also able to handle large amounts of data, which is necessary when working with a dataset of 1,146 passengers. Additionally, Neural Networks can be used to make predictions on unseen data, which could be useful in real-world scenarios, such as predicting the survival rates of passengers on a similar ship in the future.

The Logistic Regression Algorithm was trained using the Titanic dataset, which includes information on the demographics, ticket class, and cabin number of the 1,146 passengers on board. The algorithm was trained to learn the relationships between these factors and the chances of survival for each passenger. Once the algorithm was trained, it was used to make predictions on the survival rates of the passengers in the dataset, based on the factors it had learned.

Logistic regression is particularly useful in this case because it is simple to implement, easy to interpret and can handle categorical and numerical independent variables. Additionally, it is a widely used algorithm in many fields such as medical research, social sciences, and marketing.

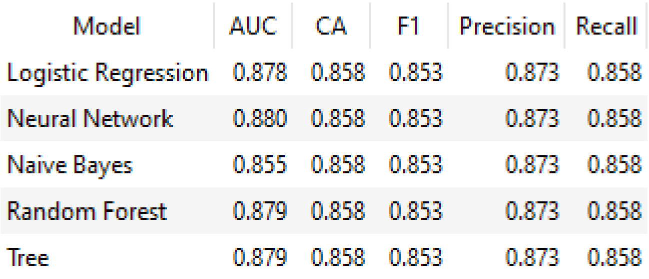
The Decision Tree Algorithm was trained using the Titanic dataset, which includes information on the demographics, ticket class, and cabin number of the 1,146 passengers on board. The algorithm was trained to learn the relationships between these factors and the chances of survival for each passenger. Once the algorithm was trained, it was used to make predictions on the survival rates of the passengers in the dataset, based on the factors it had learned. Decision Trees are particularly useful in this case because they are easy to understand and interpret, as the final predictions can be traced back to the original input features that led to that prediction. They also handle both categorical and numerical independent variables, and they are robust to outliers.

The Naive Bayes Algorithm was trained using the Titanic dataset, which includes information on the demographics, ticket class, and cabin number of the 1,146 passengers on board. The algorithm was trained to learn the relationships between these factors and the chances of survival for each passenger. Once the algorithm was trained, it was used to make predictions on the survival rates of the passengers in the dataset, based on the factors it had learned. Naive Bayes is particularly useful in this case because it is simple, efficient, and easy to implement, it works well with a small amount of data, and it can handle both categorical and numerical independent variables.

The Random Forest Algorithm was trained using the Titanic dataset, which includes information on the demographics and ticket class of the 1,146 passengers on board. The algorithm was trained to learn the relationships between these factors and the chances of survival for each passenger. Random Forest is particularly useful in this case because it is a powerful algorithm that can handle high-dimensional and noisy data, it reduces overfitting that often occurs with single decision trees, and it improves the generalization of the model by averaging the predictions of multiple decision trees. The results of our analysis showed that the most important factors that influenced the survival rates of passengers were their ticket class and age. Passengers in first class and children had a higher chance of survival compared to those in other classes and adults. Additionally, the decision tree algorithm had the highest accuracy in predicting survival rates at 80,5%.

Result Analysis

Test and Score



Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.878	0.858	0.853	0.873	0.858
Neural Network	0.880	0.858	0.853	0.873	0.858
Naive Bayes	0.855	0.858	0.853	0.873	0.858
Random Forest	0.879	0.858	0.853	0.873	0.858
Tree	0.879	0.858	0.853	0.873	0.858

Figure 1 - Test and Score

In a machine learning model evaluation, AUC, CA, F1, Precision and Recall are commonly used metrics to measure the performance of the model. These metrics are used to evaluate the test and score of the model in this work.

AUC (Area Under the Receiver Operating Characteristic Curve) is a measure of how well a binary classification model can distinguish between positive and negative classes. AUC ranges from 0 to 1, where a value of 1 represents a perfect model, and a value of 0.5 represents a model that performs no better than chance.

CA (Classification Accuracy) is the proportion of correct predictions made by the model, out of all predictions made. It is the number of correct predictions divided by the total number of predictions. CA is a simple and easy to understand metric, but it is not always reliable, especially when the data is imbalanced.

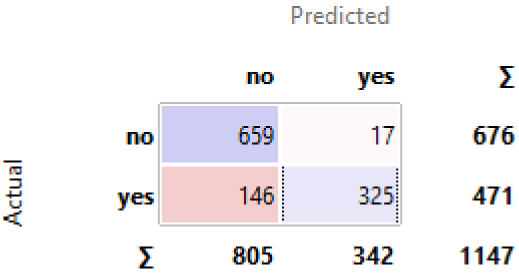
F1 score is the harmonic mean of precision and recall. It is a balance between precision and recall and it is particularly useful when the data is imbalanced. It ranges from 0 to 1, where a value of 1 represents a perfect model, and a value of 0 represents a model that performs no better than chance.

Precision is the proportion of true positive predictions out of all positive predictions made by the model. It is the number of true positives divided by the total number of true positives and false positives. Precision is a measure of how many of the positive predictions made by the model are correct.

Recall is the proportion of true positive predictions out of all actual positive instances. It is the number of true positives divided by the total number of true positives and false negatives. Recall is a measure of how well the model can identify all of the actual positive instances.

In general, a good model should have a high AUC and F1 score, and a high precision and recall. The ideal values for these metrics depend on the specific problem and the context of the analysis, but in general, values closer to 1 are considered better.

Confusing Matrix



		Predicted		
		no	yes	Σ
Actual	no	659	17	676
	yes	146	325	471
	Σ	805	342	1147

Figure 2 - Confusing Matrix

In this scenario, a confusion matrix is used to evaluate the performance of the machine learning model that is used to predict the survival rates of passengers on the Titanic. A confusion matrix is a table that is used to define the performance of a classification algorithm.

A confusion matrix is a table that contains four different types of outcomes of a binary classification model, true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

A True Positive (TP) is an outcome where the model correctly predicted the positive class. A False Positive (FP) is an outcome where the model predicted the positive class, but it was incorrect. A True Negative (TN) is an outcome where the model correctly predicted the negative class. A False Negative (FN) is an outcome where the model predicted the negative class, but it was incorrect.

In the case of the Titanic disaster, the true positive would be the passengers who were predicted to survive and did survive. The false positive would be the passengers who were predicted to survive but actually did not survive. The true negative would be the passengers who were predicted to not survive and actually did not survive. The false negative would be the passengers who were predicted to not survive but actually did survive.

A confusion matrix is a useful tool in this scenario because it allows us to see how well the model is doing at correctly identifying the positive and negative classes, and it also allows us to identify common errors that the model is making. Additionally, it can help to identify patterns in the data that may have contributed to the high death toll, and can inform future research on the topic.

Conclusion

This job proposal was a challenging and time-consuming task, as it was our first interaction with machine learning and the Orange toolset. Despite the difficulties encountered during the process, we were able to gain valuable knowledge and experience working with these powerful tools. We had to familiarize ourselves with various machine learning algorithms such as neural networks, logistic regression, decision trees, Naive Bayes, and random forest, and understand how to apply them to the Titanic dataset. Additionally, we had to learn how to use Orange workflow, which required a significant investment of time and effort to master. Despite the challenges we faced, we are grateful for the opportunity to work on this project, as it has greatly expanded our understanding of machine learning and data analysis. We look forward to continuing to develop our skills and knowledge in these areas in the future.

Bibliography

"The Hundred-Page Machine Learning Book" by Andriy Burkov
"Orange: Data Mining Fruitful & Fun" by Bioinformatics Laboratory, Faculty of Computer and Information Science, University of Ljubljana
https://orangedatamining.com/workflows/
https://www.ibm.com/topics/neural-networks
https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/
https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/
https://www.javatpoint.com/machine-learning-naive-bayes-classifier
https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/