# **TOOLS AND METHOD**

In this chapter, the tools and method used for this project will be explained.

## **Machine Learning Algorithms**

In predicting the survival of passengers during the Titanic shipwreck, we identified and evaluated three different machine-learning algorithms that are suitable for this type of dataset. An empirical comparison was performed using these three learning algorithms, the algorithms are described below.

### **Two-Class Boosted Decision Tree**

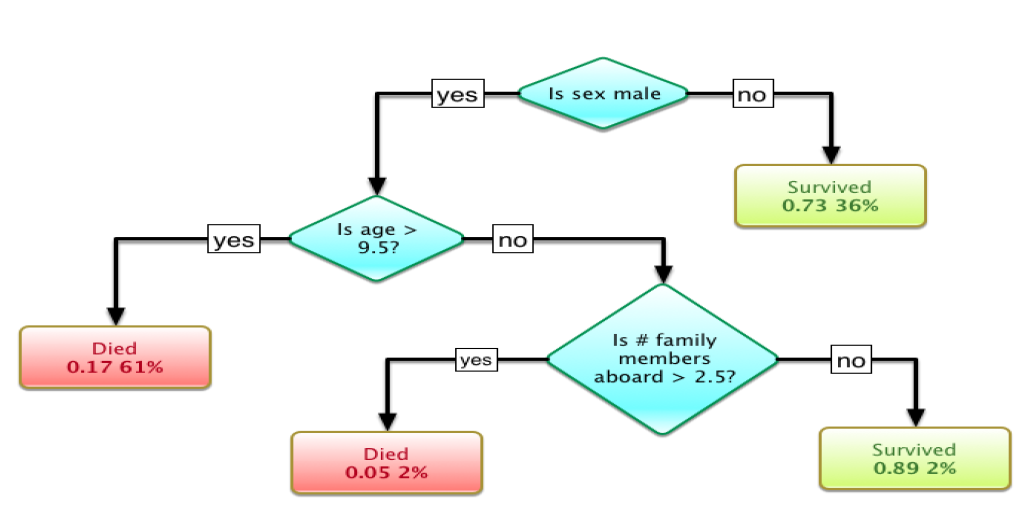
Two-Class Boosted Decision Tree is based on boosted decision tree algorithm; it was recently developed as an alternative to Random Forest. The trees in Random Forest are trained independently while the trees in boosted decision trees are trained sequentially by boosting. It uses the ensemble method to construct more than one decision tree. A boosted decision tree is an ensemble learning method in which the second tree corrects the errors of the first tree, the third tree corrects the errors of the first and second trees and it continues like that. The predictions are based on the entire ensemble of trees together that makes the prediction.

Boosted decision tree is very easy to understand, configure and interpret. It is very easy to get top performance on a wide variety of machine learning task. It can be visualized and requires little data preparation; decision trees are able to handle both numerical and categorical data.

On the other hand, boosted decision tree is a memory intensive learner and cannot process a very large amount of dataset like some linear learners. They are not robust so a small change in training data can cause a big change in the tree making a change in the final prediction. Decision tree does not express concepts easily such as XOR, multiplexer problem or parity making these concepts hard to learn. Decision trees can generate complicated trees, which are not well established from the training data. This is call overfitting.

The limitations in decision trees can be improved by using techniques such as tree pruning. Tree pruning reduces the size of decision trees by removing sections that provide little power to the classified instances. The size of the learning tree is reduced without reducing the predictive accuracy as measured by a cross-validation set.

Since we will be using the Titanic dataset, we need to know the kind of machine learning problem we will be solving. This is a classification problem because the response class is a categorical value that is zero for deceased and one for survived. In particular, it is a two-class classification problem since there are only two possible results: zero or one. The Titanic dataset fits well to decision tree because the predictor classes are numeric and categorical.



A tree showing survival of passengers on the Titanic (source: https://commons.wikimedia.org/wiki/File:Titanic\_Survival\_Decison\_Tree\_SVG.png)

### **Logistics Regression Algorithm**

Logistics regression algorithm also referred as Logit regression or Logit model is a regression analysis method used when the dependent (outcome of the prediction) variable is a dichotomous variable (binary variable). This means that the dependent variable must have only two possible types of outcomes (A or B, Yes or No, 0 or 1). This is one among the different algorithm chosen within the scope of the project and solely dependent on the type/genre of the dataset under the scope of study.

Logistics regression is considered a classification algorithm that is highly suitable and appropriate in prediction involving classified /categorical dataset when using a supervised learning algorithm procedure – In supervised learning, the output variable and the input variable are used to train the machine. Invariably using an algorithm (in this case a logistics regression) to learn the mapping function from input variable to the output variable in order to get desired prediction result.

In a logistics regression, we are estimating an unknown probability (p) of the dependent variable for any given linear combination of the independent variable; this is mathematically referred as probability estimation. For this to be achieved, a relationship must be established between the dependent variable and the independent (predictor) variable. To be able to link together this combination of variables, we need a function and this function is called the Logit function. This is nothing but the natural logarithm of the odds ration of the variables.

Odds are the number of times success occurred to the number of times failures occurred in given trials(s). Mathematically it is the number of success for every failure. Odds that is greater than a 1 indicates success and odds less that a 1 indicates more likely a failure. While the odds ratio represents how the odds of a variable changes with one unit change or increase of an independent while holding all other variables constant. This reflects the constant effect of an independent variable on a dependent variable with the likelihood one outcome will occur.

Odds = Probability of success (X = P

Probability of failure (X) 1-P

Given that the: Logit (P) = In (P / 1-P)

However, the natural log of the odd ratio is just an equivalent of the linear function of the independent variables.

Logit (P) = In (P / 1-P) = ß0 + ß1X1 +ß2X2…+ßnXn

Knowing the implication of using a linear regression in calculating probability of binary dependent variable with dichotomous outcome, that makes it difficult to model variable with restricted range and doesn’t produce a normal distribution, provide probability estimation beyond zero (0) and one (1) and in some situation, produce a negative probability. All these problem leads to impossible prediction. For a reasonable prediction to me made in logistics regression, two necessary condition must be met: Probability must be positive (P>= 0) and probability must be less than 1 (P<= 1).

There are different approaches to achieve this but in logistics regression, using an exponential approach and the division of a number by itself plus an integer 1.

Transforming the linear regression in a logistics linear regression, we have

P = eß0 + ß1X1 +ß2X2 = Exp (Y\*)

1 + eß0 + ß1X1 +ß2X2 1+Exp (Y\*)

P = Probability, ß0 = Intercept, ß1...ßn = Regression coefficient, X1…Xn = Independent variables and e = the Euler number or exponential.

The equation above is called the Estimation Regression Equation. This equation and its regression coefficient are calculated from the train dataset using the method of Maximum Likelihood Estimation. (MLE). MLE is a method of estimating parameters of statistical model given an observation, by finding particular values that makes the observation outcome the most probable and maximize the likelihood of making the observation given the model. () The approach seeks the values for the coefficient that minimize the error in the probability prediction model.

Meanwhile, the essence of using logistics regression as a classification algorithm is to predict binary outcome, not actually to predict probability. To transform or be able to interpret the probability outcome into binary outcome of 0 or 1prediction, a threshold is defined. Usually for the logistics regression, the threshold is always at 0.5, which is derived from the sigmoid function or the logit model. To make prediction with a logistics regression using outcome of the probability, we assume that our prediction outcome is zero when the probability is less than threshold; 0 if P(x) < 0.5, and a 1 if the probability is greater than the threshold; 1 if P (x) > 0.5.

To summarise;

• Logistics regression is appropriate for dataset with one or more dependent variables.

• It is appropriate for binary dependent variables with only two types of outcome.

• It is a classification algorithm, suitable for supervised learning, and for modelling and predicting categorical variables.

• It establishes relationship between the dependent and independent variables by looking at the past result using maximum likelihood to predict accuracy and efficiency.

• It uses odds ratio to determine consistency and represent the constant effect of the independent variables on the dependent variables.

• It estimates the value of the unknown probabilities.

• Probability in logistics regression must always be positive or less than one.

• Perdition are made by interpreting the probability of the outcome in comparison to the sigmoid function threshold.

### **Support Vector Machine**

Support Vector Machine (SVM) also referred to as Support Vector Networks in machine learning, is a supervised learning algorithm and it uses mainly 2 statistical techniques; Classification and Regression analysis. Supervised learning is a technique of training a machine with data (train data) to predict the outcome of new occurrences belonging to the family of this dataset (test data). In unsupervised learning, data are not labelled and naturally, points are clustered. In the case of unsupervised learning, support vector clustering is used which is an improvement to support vector machine.

Given a set of training data, each of the records is categorized into two making it a linear binary classifier. SVM builds a model out of the training data. A new test data is assigned to one of the categories so outcome is not based on probability (except when using Platt's sequential minimal optimization method).

Each of the data are mapped as points in space and the SVM model is a clear division of these points into 2 separate categories with maximum separation width, called Hyperplane. Test data are then categorized in the two halves divided by the hyperplane.

SVM develops a hyperplane in a dimensional space that is a good separation using largest functional margin in a space of data points. Hyperplanes are not just sketched, they are mathematically determined through different methods like linear program solution or methods like Fisher linear discriminant. See below the equation of the linear program solution that is commonly used.

Linear program solution If yi= +1; wxi + b ≥ 1, If yi= -1; wxi + b ≤ 1,

For both i; yi (wi + b) ≥ 1; *In the equation above, x is a vector point and w is weight*.

The points along the maximum margins are called support vectors and the line with the largest margins between support vectors is called optimal hyperplane, see Figure 1.



Source: <https://www.researchgate.net/figure/268232391_fig5_Figure-5-Hyperplane-blue-line-representation-in-SVM-Red-and-blue-circles-represent>.

SVM does not only perform linear classification it can also effectively perform non-linear classifications especially in a high-dimensional feature space. Kernel trick is the method used by SVM to perform the non-linear classification. Kernel trick performs a dot product of vectors and allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. There are different kernels that can be used and the kernel selected will determine how effective SVM is. Mostly used kernels include Polynomial, Gaussian radial basis function (RBF). RBF is the most popularly used which adds bumps around the data points.

Polynomial kernel K (x, x') =(x · x'+ c) d; *degree-d, c is constant of influence*

RBF kernel K (x, x') = exp (− γ|| x − x'|| 2); *two samples x and x'. γ is a hyperparameter that sets the “spread” of the kernel*.

To compute the SVM classifier (*i.e. soft margins used to determine the hyperplane*) expressions need to be minimized. Methods of minimizing the expressions are primal, dual which are constrained optimizations. However, approaches like coordinate descent and sub-gradient descent are now used. When the dimension of the feature space is high, coordinate descent is mostly efficient and it works from the dual problem while sub-gradient method is mostly efficient when the training examples are many and can work directly with the expressions. The soft-margin methodologies outlined are examples of an empirical risk minimization algorithm for hinge loss, which makes SVM to have an edge in statistical analysis.

This shows the SVM uses some techniques that are related to other fundamental classification algorithms like least squares and logistic regression especially in the computation of loss function to minimize expression and to find good approximations.

Characteristics of SVM is not limited to the aforementioned but also include, Its ability to handle large feature spaces especially when dealing with large data sets, SVM usually computes sparseness of solution, Also in SVM, the dimensions of the feature space do not determine the complexity of a given problem and Soft margins can be used to control overfitting.

Possible shortcomings of SVM include: SVM is only precisely suitable for two-class tasks and algorithms that reduces classes of tasks into two have to be used in cases of multi-class tasks, Parameters of the model that provides the solution are not easy to interpret as well and Full labeling of input data is always required.

## **Platforms**

Two platforms are used for this research; Azure Machine Learning Studio and Python.

### **Azure Machine Learning Studio**

This is a collaborative, drag-and-drop platform used for building, testing and deploying predictive analytics solutions on data. It is develop by Microsoft and described as a platform where data science, predictive analytics, cloud resources and data meet. The studio contains a large number of machine learning algorithms, with modules that assist with data input, output, preparation and visualization. A model can be trained by using the components of azure machine learning to build a predictive analytics experiment. The model training and cross-validation features that are required for this project are supported by Azure Machine Learning Studio.

### **Python**

This is a high-level programming language and it Open Source – free to install. Python was originally a general purpose language. But, over the years, with strong community support, this language got dedicated libraries for data analysis and predictive modeling.

<https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/>

Pandas: An open source library sponsored by the NUMFOCUS project. It is use for data manipulation.

Numpy: A cross platform and open source library in python for scientific computing, also used as a multi-dimensional container of generic data.

Matplotlib: A 2D plotting library used for visualizing data in python.

Scikit-learn: A free library for machine learning in python, it has so many algorithms for classification, regression and clustering. Python also supports the split and train and the cross-validation approach of predictive analytics.

## **Data Processing**

In this research, we are not generating data from scratch but using an already existing dataset obtained from Kaggle website. The Titanic dataset contains 1309 rows and 14 columns (variables) and survival was based on certain criteria such as age, sex, passenger class and if a passenger made it to the life boat.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| pclass | The ticket class 1 represents upper class, 2 middle class and 3 lower class |
| survived | If passenger survived or not 1 for survived and 2 for deceased |
| name | The passengers name |
| sex | Sex of passenger either male or female |
| age | Passengers age |
| sibsp | Number of siblings or spouse on the ship |
| parch | Number of parents or children on the ship |
| ticket | Ticket number |
| fare | Passenger fare |
| cabin | Cabin number |
| embarked | Port where passenger enter the ship C = Cherbourg, Q = Queenstown, S = Southampton. |
| boat |  |
| body |  |
| Home.dest |  |

### **Variable Selection**

Variable selection is intended to select the “best” subset of predictors. But why bother?

Redundant predictors should be removed. The principle of Occam’s Razor states that among several plausible explanations for a phenomenon, the simplest is best. Applied to regression analysis, this implies that the smallest model that fits the data is best.

In this research we used forward stepwise variable selection. Forward stepwise variable selection is the simplest data-driven model building approach is called *forward selection*. In this approach, one adds variables to the model one at a time. At each step, each variable that is not already in the model is tested for inclusion in the model.

### **Transforming categorical variables**

[Label Encod](http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html#sklearn.preprocessing.LabelEncoder)ing is a method that helps to normalize labels. It can be used to transform non-numerical labels (as long as they are hashable and comparable) to numerical labels. In this research we have transformed categorical variables into numerical values.

### **Replacement of missing values**

Imputation transformer for completing missing values to increase the robustness of the training. In this research, there is only one missing value in survival variable and the missing value for survived variable was replaced with the mode. Missing values in other variables were replaced with 0.

## **Cross-validation**

This is a model validation method for evaluating how the results of a statistical analysis will generalize to an independent data set. It is used in situations where the objective is prediction and where an approximation wants to be made on how accurately a predictive model will perform in practice. Cross validation is use to assess the robustness of a classifier. In this research we used ten kfold cross validation which has been proven to be statistically good enough in estimating the performance of the classifier (Witten and Frank, 2000). The training set in the ten kfold cross validation is spilt uniformly into 10 different subsets, one set is used as the test set and the remaining nine are used to train the learner. The process is repeated ten times with different subsets being used as the test set.

## **Model training and Testing**