SURVIVAL PREDICTION FOR TITANIC DATA USING MACHINE LEARING ALGORITHMS

[1]P. Ravindra, [2]Dr. P. Vijayapal Reddy,
[1]Assistant Professor, [2]Professor
[1,2]Computer Science Engineering, Matrusri Engineering College, Hyderabad, India

Abstract : RMS Titanic sinking is one of the most infamous shipwrecks in history. During its maiden voyage, the Titanic sank after colliding with an iceberg, killed many passengers including crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class**.** In this paper we are going to make the predictive analysis of what sorts of people were likely to survive and using some tools in machine learning to predict which passengers survived the tragedy.

Index Terms **- Machine learning, Logistic Regression, Decision Trees, Feature Engineering, Titanic Dataset**

I. INTRODUCTION

Machine learning[8] means the application of any computer-enabled algorithm that can be applied against a data set to find a pattern in the data. This encompasses basically all types of data science algorithms, supervised, unsupervised,segmentation, classification, or regression". few important areas where machine learning can be applied are Handwriting Recognition, Language Translation, Speech Recognition, Image Classification, Autonomous Driving.

Some features of machine learning algorithms can be observations that are used to form predictions for image classification, the pixels are the features, For voice recognition, the pitch and volume of the sound samples are the features and for autonomous cars, data from the cameras, range sensors, and GPS.

II. LITERATURE SURVEY

Every machine learning algorithm works best under a given set of conditions. Making sure your algorithm fits the assumptions / requirements ensures superior performance. You can't use any algorithm in any condition. Instead, in such situations, you should try using algorithms such as Logistic Regression, Decision Trees, SVM, Random Forest etc. logistic regression and decision trees are the models used in this paper for prediction.

Logistic Regression[2][4][9][11] is used to model the probability of an event occurring depending on the values of the independent variables which can be categorical and numerical and to estimate the probability that an event occurs for a randomly selected observations versus the probability that the event does not occur and it is used to predict the effects of series of varibles on a binary response variable and it is used to classify observations by estimating the probability that an observation is in a particular category. It is most commonly used in social and biological sciences.

The performance of Logistic regression[1] model can be measured using AIC (AkaikeInformation Criteria), Null Deviance and Residual Deviance, Confusion Matrix and McFadden R2 is called as pseudo R2. AIC is an analogous metric of adjusted R^2 in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value. Null Deviance and Residual Deviance measure indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model. Confusion Matrix is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. McFadden R2 is used to analyze data with a logistic regression, an equivalent statistic to R-squared does not exist. However, to evaluate the goodness-of-fit of logistic models, several pseudo R-squared's have been developed. To find the accuracy of model in confusion matrix the formula is

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ positives + false\ negatives}$$

Decision tree[3] is a hierarchical tree structure that can be used to divide up a large collection of records into smaller sets of classes by applying a sequence of simple decision rules. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous(mutually exclusive) classes.The attributes of the classes can be any type of variables from binary, nominal, ordinal, and quantitative values, while the classes must be qualitative type (categorical or binary, or ordinal). In short, given a data of attributes together with its classes, a decision tree produces a sequence of rules (or series of questions) that can be used to recognize the class. One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a node. With each successive division, the members of the resulting sets become more and more similar to each other. Hence, the algorithm used to construct decision tree is referred to as recursive partitioning. Decision tree applications are prediction tumor cells as benign or malignant,

classify credit card transaction as legitimate or fraudulent, classify buyers from non -buyers ,decision on whether or not to approve a loan, diagnosis of various diseases based on symptoms and profiles.

III. METHODOLGY

The data we collected is still raw-data which is very likely to contains mistakes ,missing values and corrupt values. Before drawing any conclusions from the data we need to do some data preprocessing which involves data wrangling and feature engineering . Data wrangling is the process of cleaning and unify the messy and complex data sets for easy access and analysis .Feature engineering[7] process attempts to create additional relevant features from existing raw features in the data and to increase the predictive power of learing algorithms

Our approach to solve the problem starts with collecting the raw data need to solve the problem and import the dataset into the working environment and do data preprocessing which includes data wrangling and feature engineering then explore the data and prepare a model for performing analysis using machine learning algorithms and evaluate the model and re-iterate till we get satisfactory model performance then compare the results within the algorithm and select a model which gives a more accurate results.

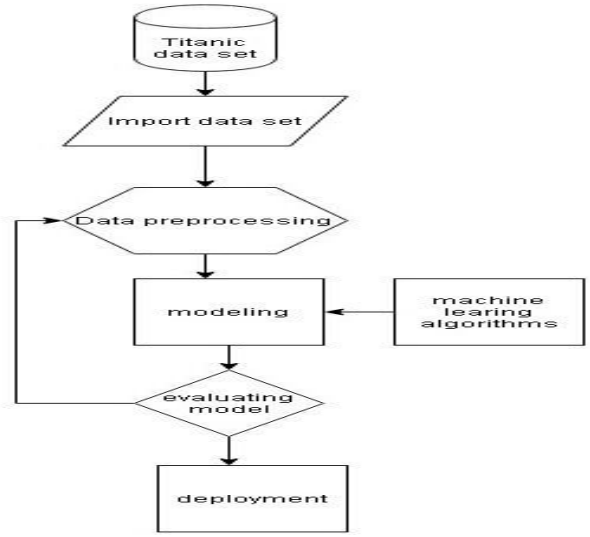

Fig:1 Operational flow chart

IV. EXPERIMENTAL ANALYSIS AND DISCUSSION

4.1 Data set description:

The original data has been split into two groups : training dataset(70%) and test dataset(30%).The training set is used to build our machine learning models. The training set includes our target variable, passenger survival status along with other independent features like gender, class, fare, and Pclass. The test set should be used to see how well our model performs on unseen data. The test set does not provide passengers survival status. We are going to use our model to predict passenger survival status. The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model we trained to predict whether or not they survived the sinking of the Titanic.

4.2 Results:

After training with the algorithms , we have validate our trained algorithms with test data set and measure the algorithms performance with godness of fit with confusion matrix for validation. 70% of data as training data set and 30% as training data set. The accuracy of predicting the survival rate using decision tree algorithm(83.7%) is high when compared with logistic regression(81.3%) for a given data set

Confusion matrix for decision tree

Trained data set

	References	
predictions	0	1
0	395	71
1	45	203

Test data set

	References	
predictions	0	1
0	97	20
1	12	48

Confusion matrix for logistic regression

Trained data set

Test data set

	References	
predictions	0	1
0	97	12
1	21	47

4.3 Enhancements and Reasoning:

Predicting the survival rate with others machine learning algorithms[8] like Random Forests , various types Support Vector Machines[8] may improve the accuracy of prediction for the given data set.

	References	
predictions	0	1
0	395	12
1	21	204

V. CONCLUSION

The analysis revealed interesting patterns across individual-level features. Factors such as socioeconomic status, social norms and family composition appeared to have an impact on likelihood of survival. These conclusions, however, were derived from findings in the given data set.

REFERENCES

[1] Atakurt, Y., 1999, Logistic Regression Analysis and an Implementation in Its Use in Medicine, Ankara University Faculty of Medicine Journal, C.52, Issue 4, P.195, Ankara
[2] Bircan H., Logistic Regression Analysis: Practice in Medical Data, Kocaeli University Social Sciences Institute Journal, 2004 / 2: 185-208
[3] M Jamel Selim S Z The construction of decision tree vol. 61 pp. 177-188 1994.
[4] J C. Bezdek Introduction of statistical model 1973.
[5] K S A 1- Sultan S Z Selim Application of decision tree vol. 26 no. 9 pp. 1357-1361 1993.
[6] Abdelghani Bellaachia and Erhan Guven. Predicting breast cancer survivability using data mining techniques.
[7] https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/
[8] Michalski R S,et al.Machine Learning:Challenges of the eighties.Machine Learning,1986,99-102.
[9] Vapnik V.N. The Nature of Statistical Learning Theory[M]. New York□Springer-Verlag.1995
[10] V. Kumar, M. Steinbach, and P. N. Tan, “Introduction to data mining,” Pearson, Addison Wesley, . London, 2006.
[11] V. Vapnik, “Statistical learning theory,” Wiley, New York, 1998.
[12] V. Kumar, M. Steinbach, and P. N. Tan, “Introduction to data mining,” Pearson, Addison Wesley, . London, 2006.