TITANIC SURVIVOR PREDICTION

Submitted by:

Abhilekh Verma

Contents

ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. I had also taken help from YouTube & online videos.

INTRODUCTION

ABOUT TITANIC SHIP

Titanic was among the one of the biggest passenger's ship in the world. It was called as an Unsinkable ship however on 15 Apr 1912 its was sink. Total 2224 passengers & Crew were aboard more than 1500 died.

PROBLEM STATEMENT

The Titanic Problem is based on the sinking of the 'Unsinkable' ship Titanic in early 1912. It gives you information about multiple people like their ages, sexes, sibling counts, embarkment points, and whether they survived the disaster. Based on these features, must predict if an arbitrary passenger on Titanic would survive the sinking or not.

OBJECTIVE

- Our main objective is to predict the survivors on titanic ship using machine learning algorithms.
- All the parameters will be analysed through Machine Learning algorithms like Logistic Regression, Decision tree classifier, Random Forest Classifier etc which will help to predict the survivors.

DATA DESCRIPTION

- The dataset contains the detailed study of Age, Sex, Fare, Passenger Class, SibSp, Survived columns.
- The source of data is taken from GitHub.

DATA SAMPLE LOOK USING PANDA

	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

METHEDOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the survivors because it gives the best outcome.
- The dependent variable is Survived, whereas other variables i.e., Sex, Age, sibsp, embarked etc. are independent variables.

METRIC USAGE

a. Logistic Regression.
b. Decision Tree Classifier.
c. AdaBoost Classifier.
d. Random Forest Classifier.
e. Gaussian NB

Hardware and Software Requirements and Tools Used

a. Hardware Requirement:
   i. Intel core i5
   ii. 8 GB Ram
b. Software Requirement:
   i. Python 3.x with packages:
      1. Pandas: Data analysis and manipulation tool
      2. NumPy: Provide support for mathematical functions, random number etc.
      3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
      4. Seaborn: is a library mostly used for statistical plotting in python.
      5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

Following models are used for solving the problem:

a. accuracy score: this function computes subset accuracy, the set of labels predicted for a sample must exactly match the corresponding set of labels in true.
b. Logistic Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results).
c. Random Forest Classifier: a collection of decision trees classifiers that each do their best to offer the best output.
d. Decision Tree Classifier: is a classification model that can be used for simple classification tasks where the data space is not huge and can be easily visualized.
e. GaussianNB: Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data.
f. Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
g. Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.

TESTING OF IDENTIFIED APPROACH(Algorithms)

a. Train Test Split
b. Logistic Regression
c. Random Forest Classifier
d. AdaBoost Classifier
e. Decision Tree Classifier
f. GaussianNB
g. Grid Search CV
h. Cross Validation

KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION.

1. Analysed data for any outliers and removed it by z-score method.
2. Analysed data for any skewness.
3. Cross Validation for cross validates the accuracy-score from overfitting.
4. Hyper parameter tuning using Grid Search CV for making the prediction better.

CONCLUSION

- Removed variables like ‘Name’, ‘Ticket’, ‘Fare’, ‘Cabin’, as they are not affecting the target variable much.
- Women, children, and first-class passengers as well as people with a small family had a better chance at survival.
- And, getting an accuracy of 79%.

REFERENCES

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc.

It also helped me to learn how to apply various model techniques on data and enable predications.