Machine Learning

Project On

TITANIC DATASET ANALYSIS AND SURVIVAL PREDICTION

TEAM MEMBERS:

Gargi Patel

Meenal

Purvi Bharani

ABSTRACT

The sinking of the RMS Titanic caused the death of thousands of passengers and crew is one of the deadliest maritime disasters in history. 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. The interesting observation which comes out from the sinking is that some people were more likely to survive than others, like women, children were the one who got the priority to rescue. The objective is to first explore hidden or previously unknown information by applying exploratory data analytics on available dataset and then apply different machine learning models to complete the analysis of what sorts of people were likely to survive. After this the results of applying machine learning models are compared and analyzed on the basis of accuracy.

INTRODUCTION

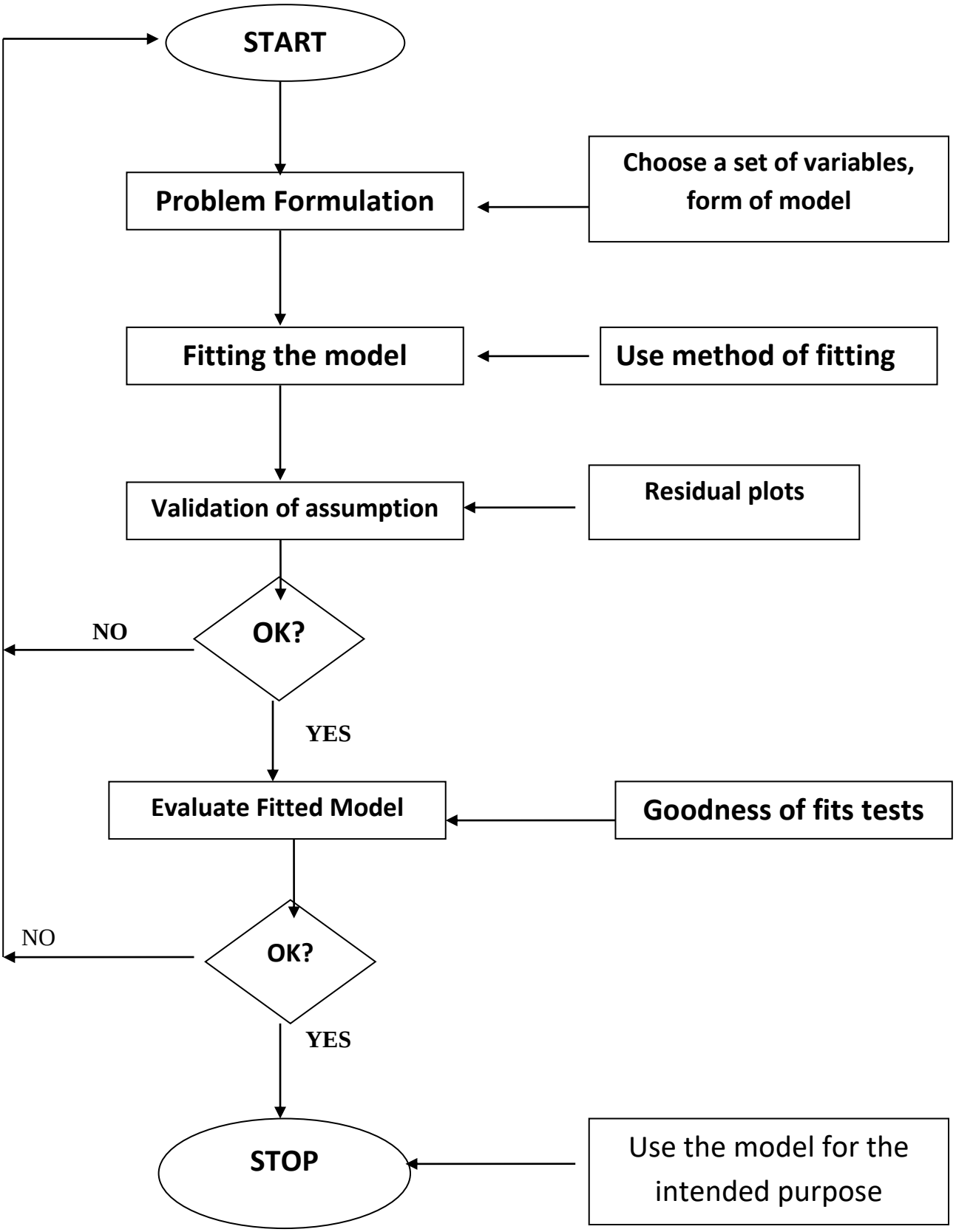
The most infamous disaster which occurred over a century ago on April 15, 1912, that is well known as sinking of "The Titanic". The collision with the iceberg ripped off many parts of the Titanic. Many classes of people of all ages and gender where present on that fateful night, but the bad luck was that there were only few life boats to rescue. The dead included a large number of men whose place was given to the many women and children on board. The men travelling in second class were dead on the vine.

Machine learning algorithms are applied to make a prediction which passengers survived at the time of sinking of the Titanic. Features like ticket fare, age, sex, class will be used to make the predictions. Predictive analysis is a procedure that incorporates the use of computational methods to determine important and useful patterns in large data. Using the machine learning algorithms, survival is predicted on different combinations of features.

The objective is to perform exploratory data analytics to mine various information in the dataset available and to know effect of each field on survival of passengers by applying analytics between every field of dataset with "Survival" field. The predictions are done for newer data sets by applying machine learning algorithm. The data analysis will be done on applied algorithms and accuracy will be checked. Different algorithms are compared on the basis of accuracy and the best performing model is suggested for predictions.

DATA ANALYTICS AND ITS CATEGORIES

PROCESS FLOW



DESCRIPTION OF DATA

This is our dataset which have following features:PassengerId: Id of every passenger.Survived: This feature have value 0 and 1. 0 for not survived and 1 for survived.Pclass: There are 3 classes of passengers. Class1, Class2 and Class3.Name: Name of passenger.Sex: Gender of passenger.Age: Age of passenger.SibSp: Indication that passenger have siblings and spouse.Parch: Whether a passenger is alone or have family.Ticket: Ticket no of passenger.Fare: Indicating the fare.Cabin: The cabin of passenger.Embarked: The embarked category.Initial: Initial name of passenger.

DATA CLEANING AND PREPROCESSING

Before applying any type of data analytics on the dataset, the data is first cleaned. There are some missing values in the dataset which needs to be handled. In age column, the missing values were replaced by mean of the existing ages. And in Embarked column the missing values were substituted with the mode of that column.

The next thing that we did was to encode the categorical features using different encoding techniques. The sex column was encoded using Label Encoder. Males were assigned a value of 0 and 1 represented the female category. Further, one hot encoding was done on Embarked column to separate the three ports from where the passengers boarded the ship. After one hot encoding the encoded features were concatenated with the original dataset to get the final data for further processing.

VISUALIZATION OF THE DATASET

In exploratory data analysis, dataset is explored to figure out the features which would influence the survival rate. The data was deeply analyzed and various graphs were plotted using matplotlib and seaborn libraries. These were done to get a visualization of different features and how they affect the output.

FEATURE ENGINEERING

Feature engineering is the most important part of data analytics process. It deals with, selecting the features that are used in training and making predictions. A bad feature selection may lead to less accurate or poor predictive model. The accuracy and the predictive power depend on the choice of correct features. It filters out all the unused or redundant features.

To get the relevant features that would impact the output, we used correlation and plotted a heatmap. The visualization resulted in selecting the features which had a correlation value greater than (+/-) 0.1. This was done and we got the top five features that played a huge role in predicting the correct output. The following features were used-

Pclass	0.338481
Sex	0.543351
Fare	0.257307
Embarked_C	0.168240
Embarked_S	0.149683

These features will be the input values for training our model. And, Survival column will be the response column. If wrong features were selected then even the good algorithm may produce the bad predictions. Therefore, feature engineering acts like a backbone in building an accurate predictive model.

FEATURE SCALING

Our next step was to scale the features using min max scaler to bring all the values in our dataset in the range of 0 to 1. Feature scaling was performed so that our model does not discriminate on the basis of different range of values in different columns. This might have resulted in neglecting the features that had small values and giving most of the importance to features which had larger range of values.

MACHINE LEARNING

What is the machine learning Model?

The machine learning model is nothing but a piece of code; an engineer or data scientist makes it smart through training with data. So, if you give garbage to the model, you will get garbage in return, i.e. the trained model will provide false or wrong predictions.

1. Gathering Data

Data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem we did data pre-processing of the data.

2. Data pre-processing

Data pre-processing is the most important step that helps in building machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis.

We took the five features selected above and put them in a variable called ‘x’. And the survived column, whose values were to be predicted was put in a variable called ‘y’. Moving on to further steps, we divided the data into two parts i.e. train data and test data. The ration of these two was kept 8:2.

3. Researching the model that will be best for the type of data

Our main goal is to train the best performing model possible, using the pre-processed data. Various machine learning models are implemented to validate and predict survival.

Supervised Learning:

In Supervised learning, an AI system is presented with data which is labelled, which means that each data tagged with the correct label.

The supervised learning is categorized into 2 other categories which are “Classification” and “Regression”.

Classification:

Classification problem is when the target variable is categorical (i.e. the output could be classified into classes — it belongs to either Class A or B or something else).

A classification problem is when the output variable is a category, such as “red” or “blue” , “disease” or “no disease” or “spam” or “not spam”.

These are some machine learning algorithms that we have used:

Logistic Regression

Logistic regression is the technique which works best when dependent variable is dichotomous (binary or categorical). The data description and explaining the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables is done with the help of logistic regression. It is used to solve binary classification problem, some of the real life examples are spam detection- predicting if an email is spam or not, health-Predicting if a given mass of tissue is benign or malignant, marketing- predicting if a given user will buy an insurance product or not.

Random Forest Classifier

Random forest algorithm is supervised classification algorithm. The algorithm basically makes a forest with large number of trees. The higher the number of trees in the forest gives the higher accuracy results. Random forest algorithm the sum of true positive and false positive (event that makes false prediction and subject result is also false).

Naive Bayes Classifier

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Decision Tree

Decision tree is a supervised learning algorithm. This is generally used in problems based on classification. It is suitable for both categorical and continuous input and output variables. Each root node represents a single input variable (x) and a split point on that variable. The dependent variable (y) is present at leaf nodes. For example: Suppose there are two independent variables, i.e. input variables (x) which are height in centimeter and weight in kilograms and the task to find gender of person based on the given data.

Support Vector Machine

Support Vector Machine (SVM) falls in supervised machine learning algorithm. This algorithm is used to solve both classification and regression problems. The classification is performed by constructing hyper planes in a multidimensional space that separates cases of different class labels. For categorical data variables a dummy variable is created with values as either 0 or 1.

K-nearest neighbors (KNN)

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well

- Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

MODEL EVALUATION

The accuracy of the model was evaluated using "accuracy score" and "R2 score" of all the models. And we came to know that random forest classifier was the best model. The next two models that could be used for prediction were the Decision Tree Classifier and Logistic Regression.

A confusion matrix is a table layout that allows us to visualize the correctness and the performance of an algorithm. Hence, a confusion matrix was then made for the random forest classifier. We also calculated the precision, recall and F-score values for the same.

The results of the confusion matrix were 147 correctly predicted values and 32 values were predicted incorrectly. The precision of a model is how correctly the values of the target variable are predicted out of the total values. Our model was 77.63% precise. The recall here would represent that out of all those who survived, how many times our model predicted the survival of a person

correctly. And that came out to be 79.72%. Finally, the F-score is the harmonic mean of precision and recall values. The F-score of Random Forest Classifier was calculated as 78.66%.

CONCLUSION

For a better view for analysis, we then represented the predicted values and the actual target variable's values along with the input features in the form of a dataframe. Hence, with all the data processing, analyzing, model building and predicting values from different models we can say that Random Forest Classifier is the best suited model for predicting the survival of people on the titanic ship. And this model can now be further used on different input parameters and predictions can be made whether a person would survive or not.