Predicting Target Value

Amisha Tiwari
3rd year Student
IET, Alwar(Raj.)

General Terms
Data Analytics, Exploratory Data Analytics, Machine Learning, Model Evaluation, Data Science.

Keywords
Data mining, ggplot, Logistic Regression, Random Forest, Feature Engineering, Support Vector Machine, Confusion Matrix.

INTRODUCTION

This project is on predicting the Text whose Target is "Blockchain". Where "Text" and "Target" are two columns. Text and Target both are object type variable.

Machine learning algorithms are applied to make a prediction of Text that whether its Target is "Blockchain" or not. Features Text and Target 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, Target is predicted on combinations of these 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 Target of data by applying analytics between every field of dataset with "Target" 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.

1. DATA ANALYTICS AND ITS CATEGORIES

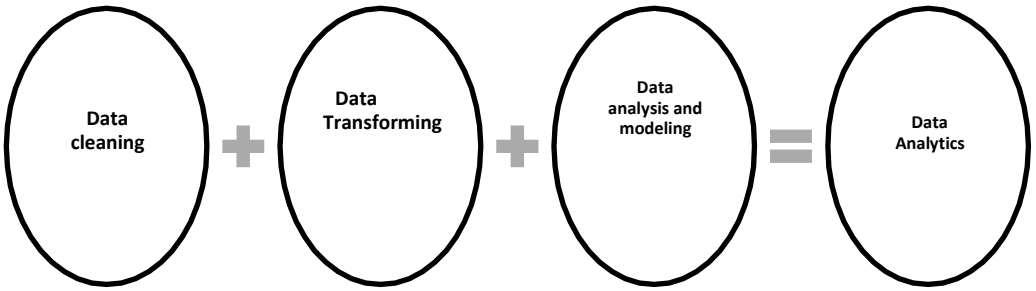

Fig 1: Data Analytics

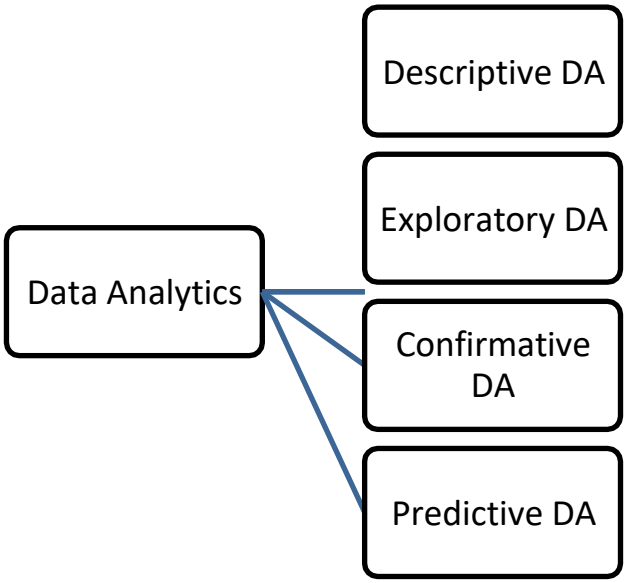

Fig 2: Categories of Data Analytics

2. PROCESS FLOW

There is a step by step approach to choose a particular model for the current problem. We need to decide whether a particular machine learning model is suitable for our problem or not. Here we can see process flow being followed.

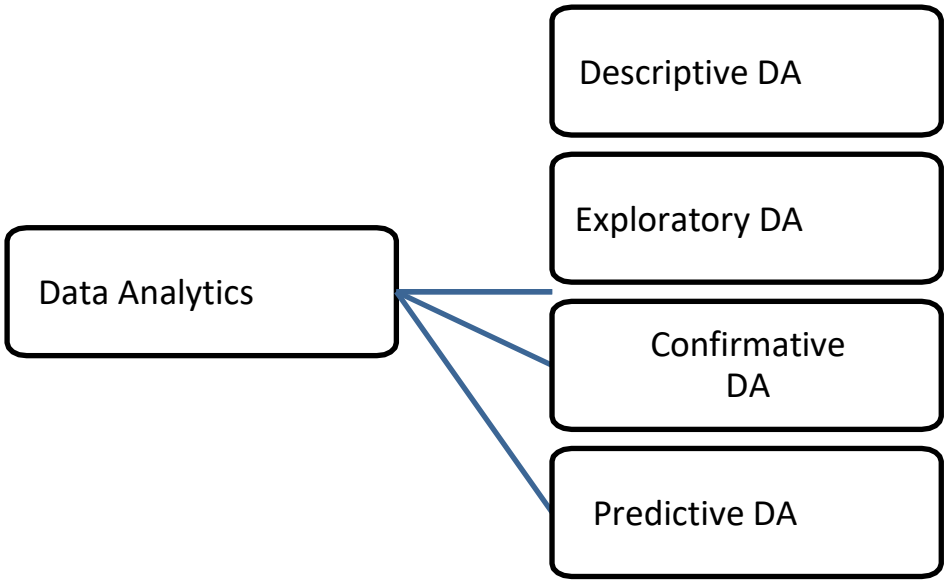

Fig 2: Categories of Data Analytics

PROCESS FLOW

There is a step by step approach to choose a particular model for the current problem. We need to decide whether a particular machine learning model is suitable for our problem or not. Here we can see process flow being followed.

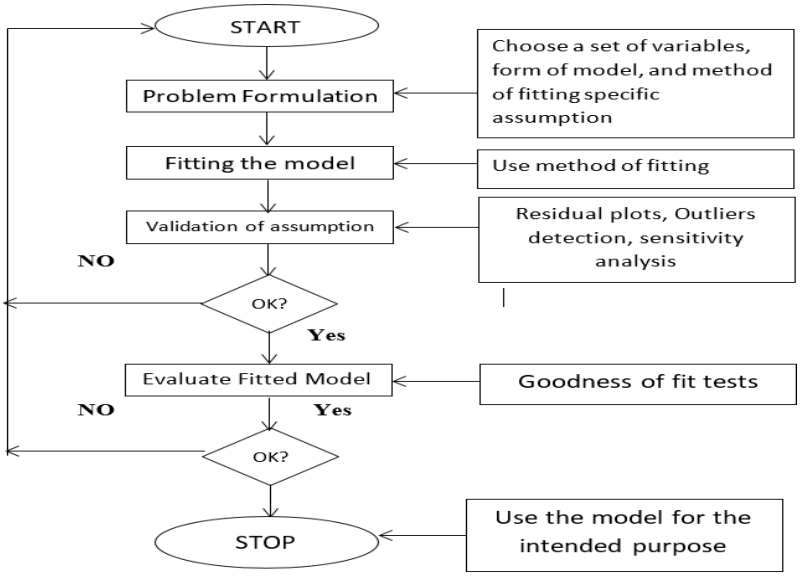

Fig 3: Process of fitting a Machine Learning Model

Table 1. Description of each attribute in our dataset

Attribute	Description	Data Type
Text	Contains text from blockchain domain	Object
Target	Target class	Object

Now let us explore our dataset by knowing the influence of each attribute on survival of target class. We will create histograms, Bar plots to achieve this.

2. DATA CLEANING

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. These missing values can be seen with the function isnull().sum(). This gives the total number of missing values in a the columns.

These missing values can be treated in many way the most simple and easy way is to delete the row where missing values exist but in case of large data or where each row is important, then dropping the values seems not good and it directly affects on accuracy. So there is another approach for handling these missing value is to take average of the column and put the average value in missing space. In case of categorical variable mode() is used to fill the column. In attribute "Text" there are 3 missing values which are replaced with random sample from existing Text. In case of column "Target" there are no missing values, so no need to clean this data set.

3. EXPLORATORY DATA ANALYSIS

We are going to perform exploratory data analysis for our problem in the first stage. In exploratory data analysis dataset is explored to figure out the features which would influence the Target class. The data is deeply analyzed by finding a relationship between attributes Text and Target.

4. METHODOLOGY

a. 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. In feature engineering the domain knowledge is used to find features in the dataset which are helpful in building machine learning model. It helps in understanding the dataset in terms of modeling. 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.

Based on the exploratory analysis above, following features Text and Target are used. Target column is chosen as response column. These features are selected because their values have an impact on the target class. These features will be the value of "x" in the bar-plots. Therefore, feature engineering acts like a backbone in building an accurate predictive model.

b. Machine Learning Models

Various machine learning models are implemented to validate and predict survival.

a. 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.

b. 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. (Hypothetical example, for demonstration purpose only).

c. Random Forest

Random forest algorithm is supervised classification algorithm. The algorithm basically makes forest with large number of trees. The higher the number of trees in the forest gives the higher accuracy results. Random forest algorithm can be used for both classification and regression problems. For instance, it will take random samples of 100 observation and 5 randomly chosen initial variables to build a model. The same process is repeated a number of times, then the final prediction is made according to the observations. Final prediction is a function (mean) of each prediction.

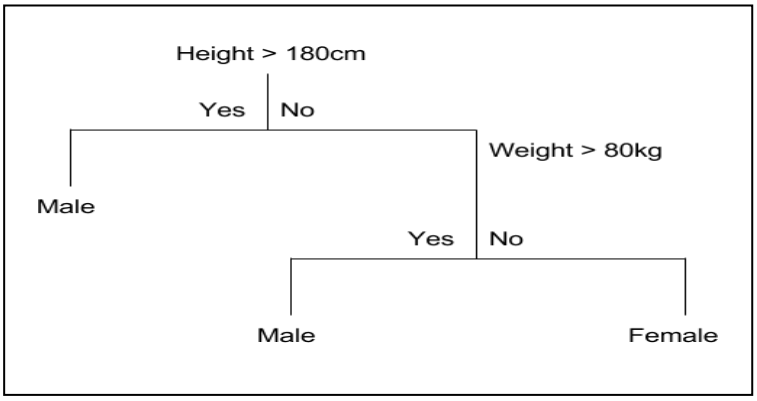

Fig 4: Example of a Decision Tree

There are two types of decision tree based on the type of target variable.

1. Categorical Variable Decision Tree: The tree in which target variables have categorical values.
2. Continuous Variable Decision Tree: The tree in which the target variable has continuous values.

d. 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. So, a categorical dependent variable consisting three levels, say (A, B, C) can be represented by a set of three dummy variables:

A: {1, 0, 0}; B: {0, 1, 0}; C: {0, 0, 1}

8. MODEL EVALUATION

The accuracy of the model is evaluated using "confusion matrix". A confusion matrix is a table layout that allows to visualize the correctness and the performance of an algorithm.

8.1 Confusion Matrix

A confusion matrix is a method to verify how accurately the classification model works. It gives the actual number of predictions which were correct or incorrect when compared to the actual result of the data. The matrix is of the order N*N, here N is the number of values. Performance of such models is commonly evaluated using the data in the matrix.

Sensitivity: It defines the percentage of actual positive which are correctly identified, and is complementary to the false negative rate. Sensitivity= true positive/(true negative + false positive). The ideal value for sensitivity is "1.0" and minimum value is "0.0"

Specificity: It measures the proportion of negatives which are correctly identified, and is complementary to the false positive rate. Specificity= true negatives/(true negatives + false positives). The ideal value for specificity is "1.0" and least value is "0.0".

Positive Predictive Value: It gives the performance measure of the statistical test. It is a ratio true positive (event that makes true prediction and subject result is also true) and the sum of true positive and false positive (event that makes false prediction and subject result is also false).

Negative Predicted Value: It is the ratio of true negatives (the event which makes negative prediction and result is also false) and sum of true negative and false negative (event that makes false prediction and subject result is positive).

8.2 Accuracy: It gives the measure of percentage of correct prediction done by the model/algorithm. The best value is "1.0" and the worst value is "0.0".

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	Positive Predictive Value	a/(a+b)
	Negative	c	d	Negative Predictive Value	d/(c+d)
		Sensitivity	Specificity	**Accuracy=**	
		a/(a+c)	d/(b+d)	**(a+d)/(a+b+c+d)**	

Fig 5: Generalized confusion matrix

In R mathematical calculations are performed and accuracy using each model is found. Here are the accuracies we achieved for each model.

9. PREDICTION

Here we can choose any of the models to predict Target class. Since we have evaluated all models by using confusion matrix we will predict by using model which has highest accuracy.

We performed prediction on data dataset by using logistic model and SVM.

10. Conclusion

Data cleaning is the first step while performing data analysis. Exploratory data analytics helps one to understand the dataset and the dependency among the attributes. EDA is used to figure out the relationship between the features of the dataset. This is done by using various graphical techniques. The one used above is ggplot and histograms.

In feature engineering the actual parameters to be used while designing the training model and prediction model is found out on the basis of exploratory data analytics process.

Machine Learning models predict the values of Target. Logistic regression technique is used in making predictions in classification problem.

The confusion matrix gives the accuracy of all the models, the logistic regression is proves to be best among all with an accuracy of 0.837261504. This means the predictive power of logistic regression in this dataset with the chosen features is very high.

It is clearly stated that the accuracy of the models may vary when the choice of feature modelling is different. Ideally logistic regression and support vector machine are the models which give a good level of accuracy when it comes to classification problem.

11. FUTURE WORK

This project involves implementation of data analytics and machine learning. This project work can be used as reference to learn implementation of EDA and machine learning from very basic.

In future the idea can be extended by making more advanced graphical user interface with the help of newer libraries like shiny in R. An interactive page can be made, i.e. if the value of a attribute is changed on the scale the values corresponding to its graph (ggplot or histogram) will also change. We can also draw much focused conclusions by combining results we obtained.