International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 08 | Aug 2022 www.irjet.net p-ISSN: 2395-0072

TERM DEPOSIT SUBSCRIPTION PREDICTION

Ram Teja Inturi[1], Gurazada D V S Sriram Chandra[2]

[1]Dept. of Computer Science and Engineering, Lovely Professional University, Phagwara
[2] Dept. of Computer Science and Engineering, KL Education Foundation, Guntur

---

Abstract - *In this paper, we are applying some of the most popular classification models) for classifying the term deposit dataset. J48 is another name for C4.5, which is the extension of the ID3 algorithm. We have collected 20 datasets from the UCI repository [1] containing many instances varying from 150 to 20000. We have compared the results obtained from both classification methods Random Forest and Decision Tree (J48). The classification parameters consist of correctly classified, incorrectly classified instances, F-measures, Precision, and recall parameters. We have discussed the advantages and disadvantages of using these two models on small datasets and large datasets. The results of classification are better when we use Random Forest on the same number of attributes and large datasets i.e. with a large number of instances, while J48 or C4.5 is good with small data sets (less number of instances). When we use Random Forest on the term deposit dataset shows that when the number of instances increased from 285 to 698, the percentage of correctly classified samples increased from 69% to 96% for the dataset with the same number of attributes, which means the accuracy of Random Forest increased. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).*

Keywords **-** Classification, Random Forest, Decision Trees, Logistic Regression, Dataset, Algorithms.

1.INTRODUCTION

The application of the Decision Tree algorithm [2] is widely observed in various fields. Classification of text, comparison of text, and classification of data are the fields where they are used. Along with these, in libraries books are classified into different categories based on their type with the implementation of the Decision Tree. In hospitals, it is used for the diagnosis of diseases i.e. tumors, Cancer, heart diseases, Hepatitis, etc. Companies, hospitals, colleges, and universities use it for maintaining their records, timetables, etc. In the Stock market, it is also used for statistics.

Decision Tree algorithms are highly effective in that the rules of classification are provided such that they are human-readable. Along with all the advantages, there are some drawbacks, one of the advantages is the sorting of all numerical attributes when a node is decided to be split by the tree. Such split on sorting numerical attributes becomes somewhat costly i.e., runtime, size of memory, and efficiency especially if Decision Trees are applied on datasets with large size i.e., it has a large number of instances. In 2001, Bierman [4] proposed the idea of Random Forests which performed better when compared with other classifiers such as Support Vector Machines, Neural Networks, and Discriminant Analysis, and it also overcomes the overfitting problem.

These methods such as Bagging or Random subspaces [5][6] which are made by combining various classifiers and those methods produce diverse data by using randomization are proven to be efficient. The classifiers use randomization in the induction process to build classifiers and introduce diversity. Random Forests have gained wide popularity in machine learning due to their efficiency and accuracy in discriminant classification [7][8].

In computer vision, Random Forests are introduced by Lepit et [9][10]. In this field, his work has provided foundations for papers such as class recognition [11][12], two-layer video segmentation [13] image classification, and person identification [14][15], as they use random forests. Wide ranges of visual clues are enabled naturally by Random Forests such as text, color, height, depth, width, etc. Random Forests are considered vision tools and they are efficient in this purpose.

Random Forest as defined in [4] is the generic principle of a combination of classifiers that uses base classifiers that are L- tree-structured {h (X, Θn), N=1,2,3,...L}, where X represents the input data and {Θn} is a family of identical and not independent and also distributed random vectors. In Random Forest the classifiers use random data to construct a decision tree from the available data. For example, in Random Forest each decision tree (as Random Subspaces) is built by randomly sampling a subset, and for each decision tree, random sampling of the training dataset is done to produce diverse decision trees (Concept of Bagging).

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary

dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

In a Random Forest, the features are selected randomly for decision tree split. The correlation between each tree in random forest decreases by randomly selecting features from a dataset which increases the predictive power of the random forest and also increases efficiency. Some of the advantages of Random Forest are [16]:

- Random forest also overcomes the problem of overfitting.
- Random forests are also less sensitive to outliers in training data.
- Setting parameters is easy which therefore eliminates the pruning of the data.
- Variable importance is generated automatically.
- Accuracy is generated automatically.

Random Forest not only uses the advantages of decision trees but also uses the bagging concepts, its voting scheme [17] through which decisions are made and random samples of subsets are generated. Random Forests most of the time achieve better results than decision trees.

The Random Forest is appropriate for data modeling which is high dimensional as it can handle missing values and also can handle numerical and categorical and continuous data and also binary data. The bootstrapping process and ensembling make Random Forest strong to overcome the problems such as overfitting and makes sure that there is no need to prune the trees. Besides some advantages such as high accuracy, Random Forest is also efficient, interpretable, and not parametric for some types of datasets [2]. The model interpretability and prediction accuracy are some of the very unique features among some of the machine learning methods provided by Random Forest. By utilizing random sampling and ensembling techniques better accuracy and generalization of data.

Bagging provides generalization, which improves with the decrease in variance and improves the overall generalization error. As same as a decrease in bias is achieved by using boosting process [19]. Random Forest has some main features which have gained some focus are:

- Accurate prediction results for different processes.
- By using model training, the importance of each feature is measured.
- Pair-wise proximity between the samples is measured by the trained model.

In this article, we discuss the accuracy and other parameters when decision trees, Random Forests, and Logistic Regression are applied to the term deposit dataset. The main objective of this comparison is to create a line between these classification methods. This also helps in the selection of a suitable model. The rest of the paper is as follows: Section 2 is about Literature Review and Decision Tree related classification algorithm which also includes the Random Forest and Logistic Regression and the datasets used are described in Section 3. Section 4 deals with the results and conclusion.

2. DATASET

You are provided with the following files: 1. train.csv: Use this dataset to train the model. This file contains all the client and calls details as well as the target variable "subscribed". You have to train your model using this file. 2. test.csv: Use the trained model to predict whether a new set of clients will subscribe to the term deposit. The sample of the data set is shown in fig-1.

2.1 Variable Definition

ID Unique client ID

age-Age of the client

job-Type of job

marital-Marital status of the client

education-Education level

default-Credit in default

housing-Housing loan

loan-Personal loan

contact-Type of communication

month-Contact month

day_of_week-Day of the week of contact

duration-Contact duration

campaign number-number of contacts performed during this campaign to the client

pdays -number of days that passed by after the client was last contacted the previous number of contacts performed before this campaign

poutcome -the outcome of the previous marketing campaign

Subscribed (target) has the client subscribed to a term deposit?

	ID	age	job	marital	education	default	balance	housing	loan	\
0	26110	56	admin.	married	unknown	no	1933	no	no	
1	40576	31	unknown	married	secondary	no	3	no	no	
2	15320	27	services	married	secondary	no	891	yes	no	
3	43962	57	management	divorced	tertiary	no	3287	no	no	
4	29842	31	technician	married	secondary	no	119	yes	no	

Fig 1: Sample view of Dataset

We have applied data pre-processing to the dataset to check if any null values are present. we will apply Dataframe.isnull(). sum() to check the number of null values in each column. If no null values are present we are clear to go. If null values are present apply simple imputer on numerical values and categorical values on categorical variables. Now we will visualize data and as we know the target we will make a correlation matrix of all variables with the target variable. If the target is independent of any variables we can use variable reduction by dropping some columns. Then we will apply three classification methods to train the dataset.

3. PROPOSED METHODOLOGY

We have used three classification algorithms on this term deposit dataset. They are Logistic Regression, Decision Trees, and Random Forests.

3.1 Logistic Regression

Logistic Regression was used in the biological sciences in the early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable (target) is categorical. You can understand it by looking at the Fig-2.

For example,

- To predict whether an email is a spam (1) or (0)
- Whether the tumor is malignant (1) or not (0)

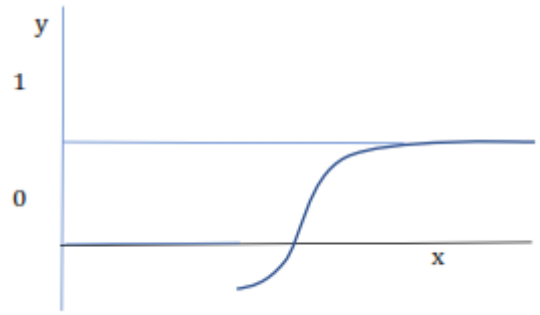

Fig 2- Logistic Regression Chart

Here in this problem, we will fit train.csv in the term deposit dataset in this model, predict the test.csv data, and save the result.

3.2 Decision Tree

Decision Trees follow a supervised classification approach. The idea came from an ordinary tree structure that was made up of a root and nodes branches and leaves. In a similar manner, the Decision Tree is constructed from nodes that represent circles and the branches are represented by segments that connect nodes. A Decision Tree starts from the root, moves downward, and generally is drawn from left to right. The node from where the tree starts is called the root node. The node where the chain ends is known as the leaf node. Two or more branches can be extended from each internal node that is the node that is not the leaf node. A node represents certain characteristics while branches represent a range of values. These ranges act as partitions for the set of values of given characteristics.

Apply the Decision Tree model on train.csv and predict test.csv data and save the results.

3.3 Random Forest

Random Forest developed by Breiman [4] is a group of non-pruned classification or regression trees made by randomly selecting samples from training data. The induction process selects random features. Prediction is done by aggregating (majority voting) the votes of each tree and the majority output will be given. Each tree is shown as described in [4]:

- By Sampling N randomly, If the no of cases in the training set is N but with replacement process, from original data. This sample will be used as the training set for making the tree.
- For M number of input variables, the variable m is selected so that m<M is satisfied at each node, m variables are selected randomly from M and the best split on this m is used for splitting. During the forest building, the value of m is made constant.
- Each tree is made to the highest possible extent. Pruning is not used.

Apply Random Forest on train.csv and predict test.csv and save the results.

Now we have applied all three models to get the model with the highest accuracy and apply hyperparameter tuning to increase accuracy.

4. RESULTS AND DISCUSSION

We have applied all three models to the dataset and we are going to compare the results. Of all three models, the random forest gives the highest accuracy as shown in Fig-3.

```
Logistic Regression
0.9048973143759874
Decision Tree
0.9042654028436019
C:\Users\Ram Teja Chowdary\Anaconda3\lib\site-packages\sklearn
\ensemble\weight_boosting.py:29: DeprecationWarning:
numpy.core.umath_tests is an internal NumPy module and should not
be imported. It will be removed in a future NumPy release.
  from numpy.core.umath_tests import inner1d
Random Forest
0.9096366508688783
```

Fig3-Results

Random Forest gives the highest accuracy so I am going to increase the accuracy by using hyperparameter tuning as in Fig-4. We are trying to increase max_estimators and see whether the accuracy increases or not.

```
print('Hyper Parameter Tuning')
f1=RandomForestClassifier(criterion='gini',n_estimators=70,random_state=1,n_j
f1.fit(xtrain,ytrain)
pred1=f1.predict(xtest)
print('Random Forest after Hyper Parameter Tuning 1')
print(accuracy_score(ytest,pred1))
f2=RandomForestClassifier(criterion='gini',n_estimators=90,random_state=1,n_j
f2.fit(xtrain,ytrain)
pred2=f2.predict(xtest)
print('Random Forest after Hyper Parameter Tuning 2')
print(accuracy_score(ytest,pred2))
f3=RandomForestClassifier(criterion='gini',n_estimators=100,random_state=1,n_
f3.fit(xtrain,ytrain)
pred3=f3.predict(xtest)
print('Random Forest after Hyper Parameter Tuning 3')
print(accuracy_score(ytest,pred3))
```

Fig 4- Hyperparameter tuning

In the figure-5 we can see the increase in accuracy after hyperparameter tuning.

```
Random Forest
0.9096366508688783
Hyper Parameter Tuning
Random Forest after Hyper Parameter Tuning 1
0.9091627172195893
Random Forest after Hyper Parameter Tuning 2
0.9102685624012639
Random Forest after Hyper Parameter Tuning 3
0.9105845181674566
```

Fig 5- Results

5. CONCLUSION

From the results, we can say that the Random Forest has increased classification performance and yields results that are more accurate and precise in the cases of a large number of instances in datasets. These scenarios also include the missing values problem in the datasets and besides accuracy, it also removes the problem of the over-fitting generated by the missing values in the datasets. Therefore, for the classification problems, if one has to do classification by choosing one among the tree-based classifiers set, we suggest using the Random Forest with great confidence for the majority and diverse classification problems.

6. REFERENCES

[1] A. Asuncion and D. Newman, "The UCI machine

[2] Yanjun Qi, “Random Forest for Bioinformatics”. (2010)

[3] Yael Ben, “A Decision Tree Algorithm in Artificial Intelligence”,2010.

[4] Breiman L, Random Forests classifier, Machine Learning, 2001.

[5] “Bagging predictors," Machine Learning, vol. 24,1996.

[6] T Ho, "constructing decision tree forests,",1998.

[7] Amit Y, Geman D: Shape quantization and recognition with randomized trees,1997.

[8]“Comparison of Decision Tree methods for finding active objects” Yongheng Zhao in classification (2012).

[9] Lepetit V, Fua P: Keypoint recognition using randomized trees. (2006)

[10] Ozuysal M, Fua P, Lepetit, V.: Fast keypoint recognition in ten lines of code. (2007)

[11] Winn J, Criminisi A: Object class recognition at a glance. (2006)

[12] Shotton J, Johnson R.: Semantic texton forests for image categorization and segmentation. (2008)

[13] Yin P, Criminisi A, Essa, I.A.: Tree-based classifiers for bilayer video segmentation. (2007)

[14] Bosh, X.: Image classification using Random Forests and ferns. (2007)

[15] Apostolof, N, Zisserman, A: Who are you? - real-time person identification. (2007)

[16] Introduction to Decision Trees and Random Forests in classification, Ned Horning.

[17] Breiman, L: Random Forests. Machine. Learning. (2001)