iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4540

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Bike Buyer Prediction using Machine Learning Techniques

Jessica Saritha[1], L. Aparna[2], S. Salomi[3], Rudra Teja[4]
[1]*Assistant Professor,* [2, 3, 4]*Student, Dept of CSE, JNTUACEP, Pulivendula, AP, India*

Abstract: ***In this project, I am working on prediction of bike buyers. Transportation has been part of our life as long as humanity exists. The most common road vehicle for transportation is automobile. Automobile includes cars, motorbikes etc.***
Bikes are the affordable vehicles for all kinds of people. As the need of transportation increases the usage of bikes also increased which results in the sudden rise of automobile industries. The automotive industry continues to face set of challenges. Most manufacturing operations in automotive industries are still largely dependent on experiences-based human decisions. I am highly interested to apply Machine Learning in the automotive industry to make a remarkable ability to bring out hidden relationships among datasets and make predictions.
Keywords: Machine Learning, Bikes, automobiles, Classification algorithms, SVM, Decision tree, Random Forest, f_score.

I. INTRODUCTION

The problem with every manufacturing company is about the amount of production of the product. If they couldn't understand the requirement of the product before manufacturing it results in the incurring of losses to the companies. In this project we focused on automobile industry in which bikes are key roles. Our project helps the industry to predict the correct requirement of the product based on customers behavior in the previous deals. Given a dataset containing various 13 attributes of 6995 customers , define classification algorithms. To apply different classification algorithms on the bike buyer dataset then choose the best algorithms
Based on the accuracy which can identify whether a person is willing to buy the bike or not. The dataset that discussed in this paper working is downloaded from https://drive.google.com/open?id=14Us9JSvbKm9xoS85rYFK0_u jREeelvOs . The number of instances are 6998. Bike Buyer is a class label used to divide into groups (Bike Buyer or not).

II. RELATED WORK

The related academic work is found at Bike sharing project . It is somewhat similar to my project and I got inspired by this to apply machine learning in automotive industry as well. In this article it is taken as regression task and my task follows classification.
My main aim is to predict the bike buyers. For doing this I selected my private dataset and that can be found in the following link Bike buyer dataset
The related academic work can be found in the link :-https://towardsdatascience.com/predicting-no-of-bike-share-users-machine-learning-data-visualization-project-using-r-71bc1b9a7495

III. PROPOSED SYSTEM

Using data mining techniques predicting bike buyer is a time consuming task which degrades performance, By Appling different machine learning techniques An early prediction of bike buyer will increase the possibilities of buying bikes based on accuracy and Fscore to find the best suitable algorithms for predicting bike buyers which gives best performance. It added a greater advantage to automobile field.

Some of the classification algorithms used are:

1) SVM
2) Decision trees
3) Support Vector Machine
4) Random Forest

A. Metrics

I want to use accuracy as evaluation metric for prediction of bike buyer. The performance of a model cannot be assessed by considering only the accuracy, because our data is highly unbiased. Therefore this experiment considers the F1 score along with the accuracy for evaluation.

Thus, here we will use F-beta score as a performance metric, which is basically the weighted harmonic mean of precision and recall.
Precision and Recall are defined as:
Precision=TP/ (TP+FP), Recall=TP/ (TP+FN), where
TP=True Positive
FP=False Positive
FN=False Negative
In the same vein, F-beta score is:
F-beta score = (1+β^2)*precision*recall/((β^ 2*precision)+recall)
We can use F-beta score as a metric that considers both precision and recall
Implementation

IV. DATASET

The Bike Buyer Dataset comprised of 13 different attributes of 6995 customers. The patients were described as either 'Yes' or 'No' on the basis of bike buyer or not. The detailed description of the dataset is shown in Table. The table provide details about the attribute and attribute type. As clearly visible from the table, all the features except marital status, region, gender, education, occupation are real valued integers. The numerical attributes information is as follows:

	ID	Yearly Income	Children	Cars	Commute Distance	Age
count	6996.000000	6997.000000	6997.000000	6997.000000	6997.000000	6997.000000
mean	17744.435249	57020.151493	1.108761	1.586823	4.209233	45.107332
std	4337.428859	32080.449720	1.599842	1.146782	2.920171	11.916654
min	2.000000	0.000000	0.000000	0.000000	1.000000	25.000000
25%	14249.750000	30000.000000	0.000000	1.000000	1.000000	36.000000
50%	17406.500000	60000.000000	0.000000	2.000000	4.000000	44.000000
75%	20609.500000	70000.000000	2.000000	2.000000	6.000000	53.000000
max	29476.000000	170000.000000	5.000000	4.000000	13.000000	96.000000

A. Dataset Description

The dataset is quite interesting because it is a good mixture of categorical and numerical attributes. There are total 13 attributes and 6995 instances are there. There are few missing values also. There are mainly two classes' bike buyers or not. The dataset I am using is unbalanced because among 6995 customers [5997] are 'non-buyers' and [998] are 'buyers'.

B. Data Preprocessing

In this step of data pre-processing we will pre-process the data. We will read the data from dataset and replace the null values. We will know the information about all the data types. We will know the mean standard deviation and various metrics regarding to the numerical data.
To know the correlation between the numerical attributes we will plot the graphs to visualize the data (this context we will use pair plot and heatmap). We will now remove the true outliers. We will consider the data which is three standard deviations away from the mean as the outliers and remove them. Now we need to perform a One Hot Encoding of the categorical variables to prepare the data for classification. We can do this easily by using OneHotEncoder from the sklearn.preprocessing module. Normalising to rescale the features to a standard range of values using MinMaxScaler. After that the whole dataset is divided into training and testing data using train_test_split from sklearn.model_selection.

V. CLASSIFICATION TECHNIQUES

Three supervised learning approaches are selected for this problem. Care is taken that all these approaches are fundamentally different from each other, so that we can cover as wide an umbrella as possible in term of possible approaches. For example- We will not select Random Forest and RandomForest together as they come from the same family of 'ensemble' approaches. The choice of algorithms was influenced from these source:
https://stackoverflow.com/questions/2595176/which-machine-learning-classifier-to-choose-ingeneral
For each algorithm, we will try out different values of a few hyper parameters to arrive at the best possible classifier. This will be carried out with the help of grid search cross validation technique. For these dataset we apply the different supervised learning algorithms there are:

A. Random Forest Classifier

Tree models are known to be high variance, low bias models. In consequence, they are prone to overfit the training data. This is catchy if we recapitulate what a tree model does if we do not prune it or introduce early stopping criteria like a minimum number of instances per leaf node. Well, it tries to split the data along the features until the instances are pure regarding the value of the target feature, there are no data left, or there are no features left to spit the dataset on. If one of the above holds true, we grow a leaf node. The consequence is that the tree model is grown to the maximal depth and therewith tries to reshape the training data as precise as possible which can easily lead to over fitting. Another drawback of classical tree models like the (ID3 or CART) is that they are relatively unstable. This instability can lead to the situation that a small change in the composition of the dataset leads to a completely different tree model

1) Advantages

a) Reduction in over fitting: by averaging several trees, there is a significantly lower risk of over fitting.

b) Less variance: By using multiple trees, you reduce the chance of stumbling across a classifier that doesn't perform well because of the relationship between the train and test data.

2) Disadvantages

a) It takes more time to train samples.

B. Support Vector Machine

SVM aims to find an optimal hyper plane that separates the data into different classes, using a method called as kernel to project data points belonging to a particular class into different dimensions, so that a hyperplane can easily pass through and maintain the largest possible distance between itself and these data points.

1) Advantages

a) Performs well with high dimensional data. SVM's are very good when we have no idea on the data.Works well with even unstructured and semi structure data like text, Images. The kernel trick is strength of SVM. By using the kernel function to solve the complex problem

b) The SVM model have generalization in practice, the risk of over fitting is less in SVM.

2) Disadvantages

a) Choosing the good kernel function is not easy and it take long training time for large datasets

b) Difficult to understand and interpret the final model, variable weights and individual impact.

C. Decision Tree Classifier

The decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, its also widely used in machine learning.

1) Advantages

a) Able to handle categorical and numerical data.

b) Doesn't require much data pre-processing, and can handle data which hasn't been normalized, or encoded for Machine Learning Suitability.

c) Simple to understand ,visualize and interpret.

2) Disadvantages

a) Complex Decision Trees do not generalize well to the data and can result in over fitting.

b) Unstable, as small variations in the data can result in a different decision tree.

VI. RESULTS

Since the size of dataset is small at present , there is not much difference between training and testing times of different algorithms. However, for the sake of comparison, these times have been displayed in the 'Implementation' sub-heading of 'Analysis' section. Decision Tree Classifier consumes maximum time during training and good accuracy score, F_score. From this dataset use three models based on the accuracy_score,F_score we decide DecisionTree is best suitable for this dataset.

Accuracy: The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

1) Recall

Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN).

Recall is given by the relation:

$$\text{Recall} = \frac{TP}{TP + FN}$$

2) Precision

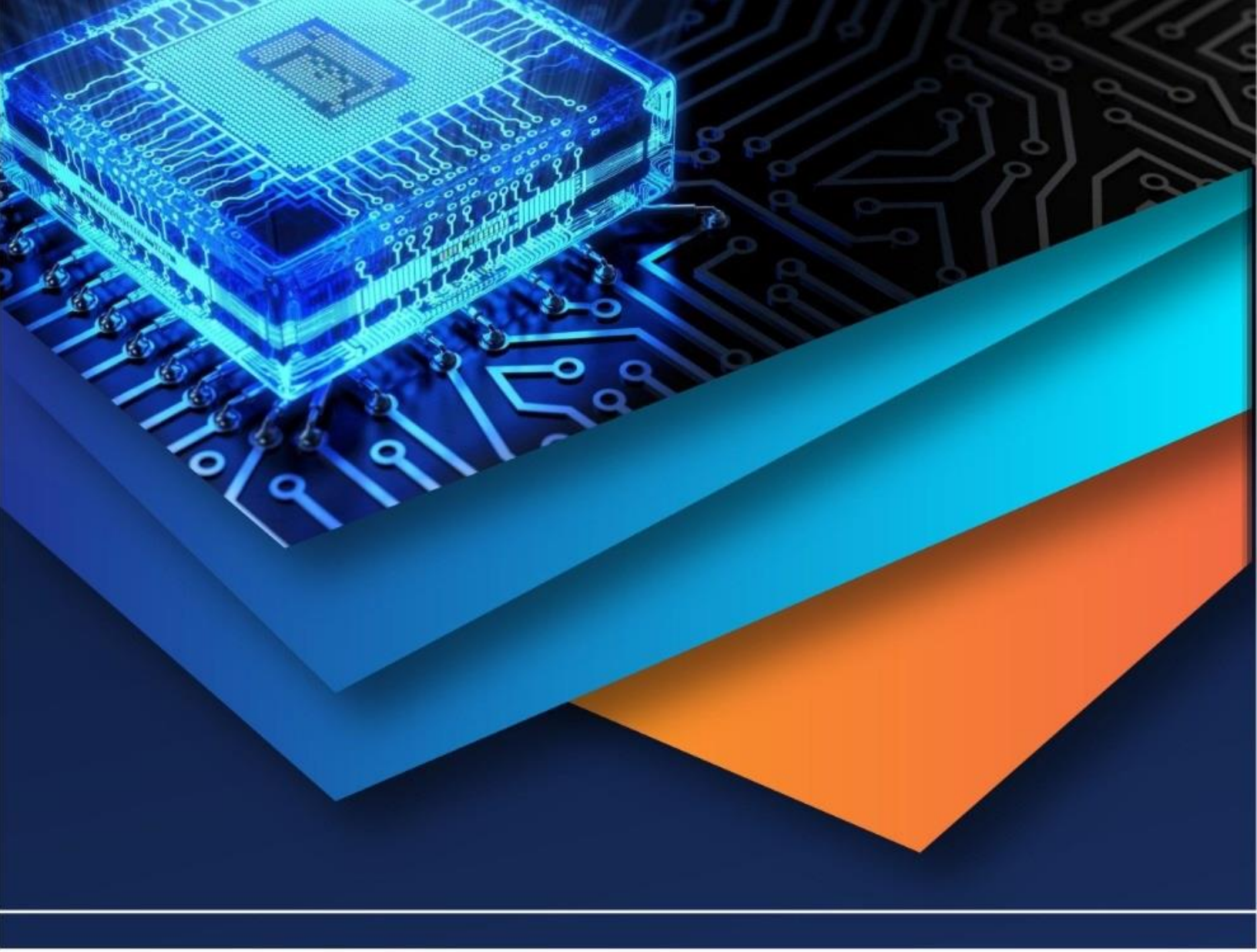
To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labeled as positive is indeed positive (small number of FP).

Precision is given by the relation:

$$\text{Precision} = \frac{TP}{TP + FP}$$

VII. CONCLUSION

In this project, we have proposed methods for prediction of bike buyers using machine learning techniques. The four machine learning techniques that were used include SVM, Random Forest, Decision tree, KNN. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metric. Decision Tree techniques resulted highest f_score of 0.2758. From the above results Decision Tree plays a key role in shaping improved classification accuracy of a dataset.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY