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.2019.4494

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

Car Buying Decision using Machine Learning Algorithms

V. Kavitha[1], P. Divya Sai[2], S. Revathi[3], P. Vishnu Kumar Reddy[4]
[1]*Adhoc Lecturer,* [2,3,4]*Student, Dept of CSE, JNTUACEP, Pulivendula, AP, India*

Abstract: ***The main aim is to predict the best decision to buy a car. Cars are essentially part of our everyday lives. There are a different variety of cars produced by various manufacturers in the industry, therefore, the buyer has a choice to make. The choice made by buyers or drivers will mostly depend on the price, safety, and how luxurious or spacious the car is. We will use different classification algorithms and find the algorithm that will make best prediction in this aspect. The key motivation of this projecr is find the best suited algorithm for the car evaulation dataset. Accuracy is used as the evaluation metric and the best model is chosen based on the accuracy.***
Keywords: Classification, Logistic Regression, K-Nearest Neighbour, Decision Tree.

I. INTRODUCTION

When an individual considers of buying a car, there are many aspects that could impact his/her choice on which kind of car he/she is interested in. There are different selection criteria for buying a car such as a prize, maintenance, comfort, and safety precautions, etc. Safety, cost, and luxury are important factors to consider in buying cars. These factors vary based on type, model, and manufacturer of the car. However, these factors are so crucial in aspect like accident number reduction. Standard equipments are part of the factors to consider when buying a car. Standard equipment's include conveniences, performance enhancers, and safety equipment. Safety as mentioned in the factors, is really indispensable, also as much as conveniences which in the case of this study falls under the attributes, door, maintenance, and luggage boot.

In this approach, the data is modelled as a classification problem where various algorithms are used and its efficiency is calculated by splitting the data into train and test sets.

II. LITERATURE SURVEY

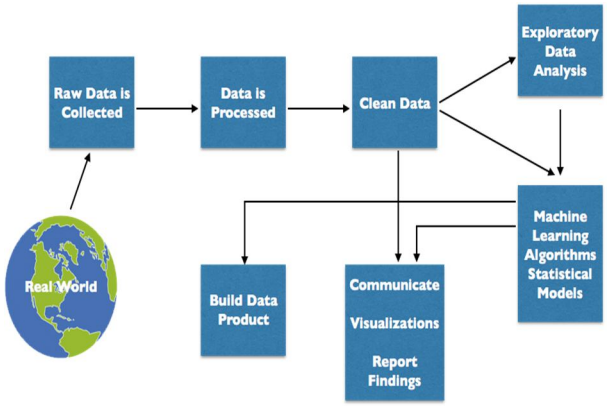
In this study, we have studied various methods of car evaluation dataset. Each technique has a different output with different accuracies. Understanding the idea in settling on a choice in procuring a car is fundamental to everybody particularly the first buyers through purchasers or any individual who lack knowledge of how the car business functions. A study conducted by [1] on the car evaluation dataset employs various data mining technique to investigate the performance of various classifiers. In the research conducted by [2], they also mine customer feedbacks and extract interesting patterns from the dataset and created clusters. The observations as a claim by [3] that summarization task is different from traditional text summarization. They proposed a set of techniques for mining and summarizing product reviews based on data mining and natural language processing methods. Their experimental results indicate that the proposed techniques are very promising in performing their tasks.

A. Ronald F., 2004 "Decision Model for Car Evaluation Final Project in Pattern Recognition."
B. Marko B. and Rajkovic V. 1988 "Knowledge acquisition and explanation for multi-attributed decision making." 8th Intl Workshop on Expert Systems and their Applications.
C. Eduard A. S. and E. K. Özyirmidokuz 2015 has Proposed a system named as "Mining Customer Feedback Documents" international journal of Knowledge engineering.

III. PROPOSEDWORK

By using data mining techniques predicting about the term deposit is a time consuming task. So in the proposed system we will use different supervised classification models to find the accuracy given by the each model and finally selects the best model which gives the highest accuracy. Some of the classification models which used are Logistic Regression, K Nearest Neighbour, Decision Tree. The architecture of this application is

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com



IV. EVALUATION METRIC

We will use accuracy score as evaluation metric to predict the outcome of chess end games. It is defined as the number of correct predictions made as a ratio of all predictions made. Accuracy is most common evaluation metric for classification problems.

Number of correct predictions
Accuracy= ----------
Total number of predictions

V. DATASET

The dataset used in this study which is a collection of the records on specific attributes on cars donated by Marco Bohanec in 1997 was obtained from the UCI dataset repository.

The car evaluation dataset as described in the The Car Evaluation Database contains examples with the structural information removed, i.e., directly relates CAR to the six input attributes namely buying, maint, doors, persons, lug_boot, safety.

This data is downloaded from https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

Number of Instances: 1728

Number of Attributes: 6

A. Attribute Names

buying, maint, doors, persons, lug-boot, safety.

B. Attribute Values

buying v-high, high, med, low
maint v-high, high, med, low
doors 2, 3, 4, 5-more
persons 2, 4, more
lug_boot small, med, big
safety low, med, high

Data set characteristics	Multivariate	Number of instances	1728
Attribute characteristics	Categorical	Number of attributes	6
Task to be done	Classification	Missing value attribute	None

VI. DATA ANALYSIS

The objective of data analysis step is to increase the understanding of the problem by better understanding the problems of the data. There are two approaches to describe a given dataset. Summarizing and Visualizing data.

A. *Data Exploration*

The dataset used in this study which is a collection of the records on specific attributes on cars donated by Marco Bohanec in 1997 was obtained from the UCI dataset repository. This data is downloaded from https://archive.ics.uci.edu/ml/datasets/Car+Evaluation
There are 1728 instances and 6 attributes in my dataset.
The information about each attribute is explained below:

1) Buying (buying price)
2) Maint (price of maintenance)
3) Doors (number of doors)
4) Persons (capacity in terms of persons to carry)
5) Lug_Boot (the size of luggage boot)
6) Safety (estimated safety of the car)

VII. ALGORITHMS AND TECHNIQUES

The algorithms which I am going to use are mentioned in my proposal namely Logistic Regression, K-Nearest Neighbour, Decision tree.

A. *Logistic Regression*

Logistic Regression is a technique borrowed by machine learning from the field of statistics. Logistic regression predicts the probability of an outcome that can only have two values .The prediction is based on the use of one or several predictors (numerical and categorical).Moreover logistic regression is a predictive analysis. When selecting the model for the logistic regression analysis, another important consideration is the model fit. However, adding more and more variables to the model can result in overfitting, which reduces the generalizability of the model beyond the data on which the model is fit. I can see two main advantages of logistic regression. The first is you can include more than one explanatory variable (dependent variable) and those can either be dichotomous, ordinal, or continuous. The second is that logistic regression provides a quantified value for the strength of the association adjusting for other variables (removes confounding effects).

B. *Strengths*

Logistic regression also performs better (than Naive Bayes) if your features are not conditionally independent. But logistic regression has the advantage over decision trees and SVM of allowing you to update your model as you receive new data and producing probabilities so that you can measure the confidence level of the model's predictions.

C. *Weaknesses*

Logistic regression doesn't perform well when the feature space is too large and/or there is a large number of categorical features. It also requires you to perform transformations for non-linear features and may be influenced by outliers since it relies on the entire data set.

1) *Parameters:* https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.Logi sticRegression.html
2) *Reference:* https://www.statisticssolutions.com/what-islogistic-regression/
3) https://machinelearningmastery.com/logistic-regression-formachine-learning/
4) *K-Nearest Neighbour:* The k-Nearest Neighbour algorithm is a simple, easy to implement a supervised machine learning algorithm that can be used to solve both classification and regression tasks. The KNN algorithm assumes that similar things exist in close proximity. KNN has no model other than storing the entire dataset, so there is no learning required. When KNN is used for classification, the output can be calculated as the class with the highest frequency from the K most similar instances. Each instance in essence votes for their class and the class with the most votes is taken as the prediction.
5) *Advantages:* The K-Nearest Neighbour (KNN) Classifier is a very simple classifier that works well on basic recognition problems.
6) *Disadvantages:* The main disadvantage of the KNN algorithm is that it is a lazy learner, i.e. it does not learn anything from the training data and simply uses the training data itself for classification. To predict the label of a new instance the KNN algorithm

will find the K closest neighbours to the new instance from the training data, the predicted class label will then be set as the most common label among the K closest neighbouring points. The main disadvantage of this approach is that the algorithm must compute the distance and sort all the training data at each prediction, which can be slow if there are a large number of training examples. Another disadvantage of this approach is that the algorithm does not learn anything from the training data, which can result in the algorithm not generalizing well and also not being robust to noisy data. Further, changing K can change the resulting predicted class label.

7) *Reference Link:* https://scikitlearn.org/stable/modules/neighbors.html#nearestneighborsclassification
8) *Decision Tree:* Decision tree is a general predictive modeling tool that has application spanning a number of different areas. They are constructed via an algorithmic approach that identifies ways to split a dataset based on different conditions. Decision trees are a non-parametric supervised learning method used for both classification and regression tasks.
9) *Strengths:* It is very easy to understand and interpret. The data for decision trees require minimal preparation.
10) *Weaknesses:* Sometimes a decision tree may become complex. The outcomes of decisions can be based mainly on your expectations. So this can lead to unrealistic decision trees. Since a decision tree can handle both numerical and categorical data, it's a good choice of algorithm. The goal is to create a model that predicts the value of the target variable by learning simple decision rules. https://www.hackerearth.com/practice/machine-learning/machine-learningalgorithms/ml-decision-tree/tutorial
11) *Parameters:* https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

As my application is classification oriented. So, techniques that are used are taken from Classification models.

VIII. METRICS

I want to use an accuracy score as my evaluation metric for predicting the best classifier for my dataset.

Accuracy is a common metric for binary classifiers, it takes into account both true positives and true negatives with equal weight. Accuracy measures how often the classifier makes the correct prediction. It's the ratio of the number of correct predictions to the total number of predictions (the number of test data points).

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Datasize}}$$

It would seem that using **accuracy** as a metric for evaluating a particular model's performance would be appropriate.

Further, I would like to use f-score if necessary.

F score is a measure of the test's accuracy. It considers both precision and recall of the test to compare the score.

** Precision ** tells us what proportion of messages we classified as spam, actually were spam. It is a ratio of true positives (words classified as spam, and which are actually spam) to all positives (all words classified as spam, irrespective of whether that was the correct classification), in other words, it is the ratio of

`[True Positives/(True Positives + False Positives)]`

** Recall (sensitivity) ** tells us what proportion of messages that actually were spam were classified by us as spam. It is a ratio of true positives (words classified as spam, and which are actually spam) to all the words that were actually spam, in other words, it is the ratio of `[True Positives/ (True Positives + False Negatives)]`

1) *Confusion Matrix:* A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.
2) The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.
3) *True Positives (TP):* These are cases in which we predicted yes (they have the disease), and they do have the disease.
4) *True Negatives (TN):* We predicted no, and they don't have the disease.
5) *False Positives (FP):* We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
6) *False Negatives (FN):* We predicted no, but they actually do have the disease. (Also known as a "Type II error.")
7) *Reference link:* https://blog.exsilio.com/all/accuracy-precision-recall-f1-scoreinterpretation-of-performance-measures/ https://en.wikipedia.org/wiki/F1_score

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

IX. METHODOLOGY

A. Data Preprocessing

Before data can be used as input for machine learning algorithms, it should often be cleaned, formatted, and restructured — this is typically known as preprocessing. Fortunately, for my dataset, there are no invalid or missing entries we must deal with. It refers to transformations applied to our data before feeding it to the algorithm. The technique that is used to convert raw data into a clean data set. The data as obtained from the UCI dataset repository have to be cleaned and to ensure that it is in the standard quality before the model creation is initiated. So clean the data i.e. removing unwanted data or replacing null values with some constant values or removing duplicates if any. Then finding the correlation for each feature with the target variable. Data transformation is a very crucial process in data preprocessing. It involves normalization and aggregation. Normalization: Normalization makes training less sensitive to the scale of features, so we can better solve for coefficients. After a pre-processed dataset is split into two halves of varying sizes at different times for use as training and testing datasets for model creation and selection of which of the models performs best. The data set used for training is mainly a portion of the dataset from which the classifying algorithm used learns the class/result of the model created from each model. The whole data is divided into training and testing data using train_test_split from sklearn.model_selection.

B. Implementation

The implementation process can be divided into two main stages.

1) The classifier training

2) Tuning the parameters of the best defined classifier

Classifier Training stage involves selecting the classifier for model creation with above-mentioned algorithms namely
Logistic Regression K-Nearest Neighbor algorithm Decision tree

Firstly I tried logistic regression the accuracy was 66.4%, later after regularizing the parameters using cross-validation scores, learning curves the accuracy increased to approximately 71%.

Secondly, I tried the KNN algorithm with a rapid increase in accuracy i.e., 90%. Finally, I used Decision Tree Classifier to get better accuracy of 96.5% which crossed my benchmark accuracy.

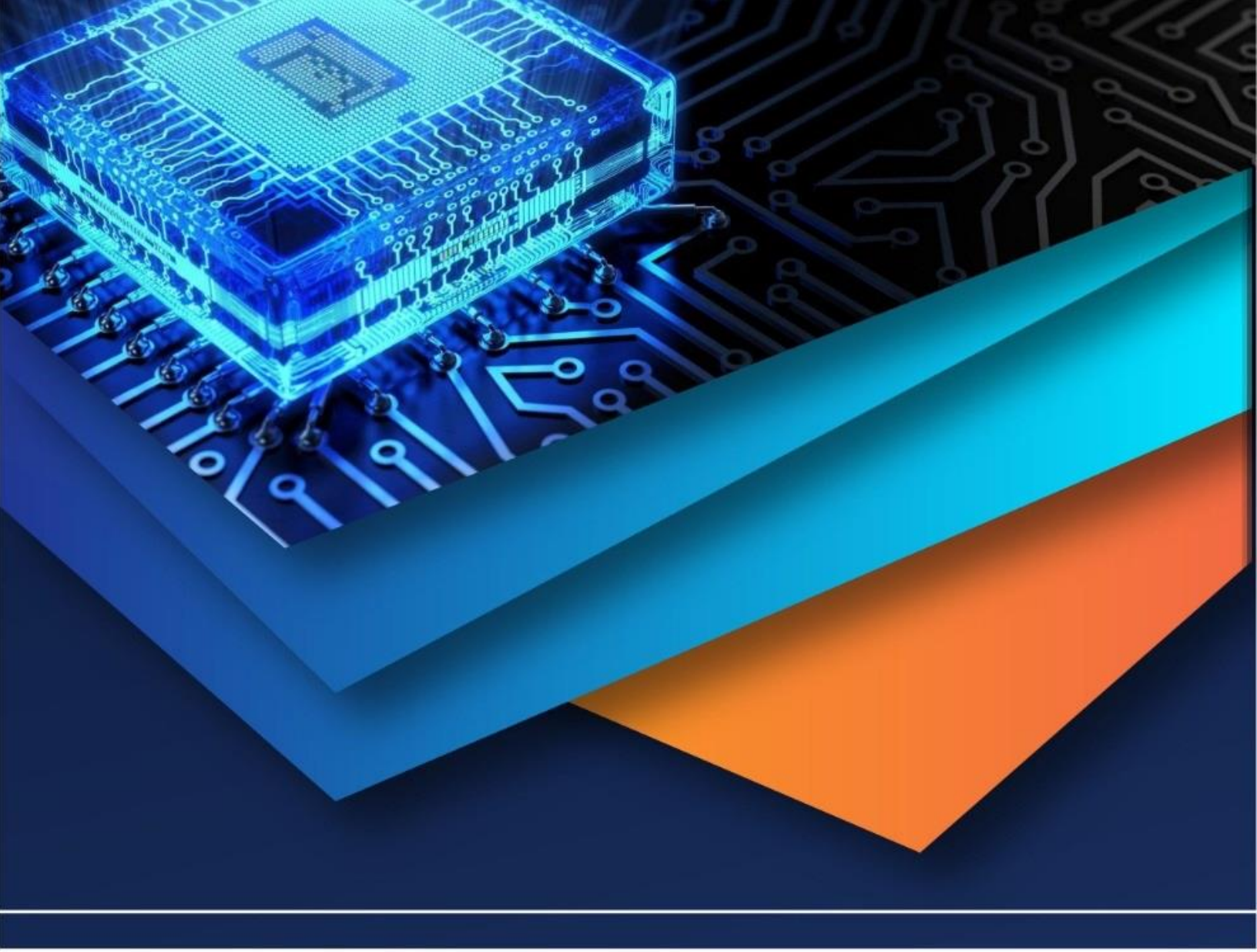
Tuning the parameters stage involves the parameter of the best-defined classifier are tuned using appropriate methods to get the best accuracy score.

a) After getting the necessary parameters of the best-defined classifier the training data is fitted to the Classifier.

b) Now the scores like accuracy score, f1 score etc. are obtained from the data using the Classifier.

X. CONCLUSION

In this paper, I evaluated the different classifiers on car evaluation dataset. Based on the customer feedback about the cars used, the model is very appropriate to judge the best car segment as per the requirement of the customer. In future, research can be use more refine technique to give more accuracy and deal with the some other issue like choose the nature of feeling, also assemble the traverse of the testing dataset and can take a gander at the more auto evolution as enormous number of flexible car are available in market. Not simply with compact brand however for other thing we can perform same investigation.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY