**Car Evaluation Database**

Group 11

Amit Adhikari (u2044461)

Karuna Shahi (u2020002)

Sujit Dhakal (u2011202)

Janesh Swar (u2013146)

**Abstract:** Cars are becoming very important means of transport in this fast-paced world. There are different types of cars in market made by different company with different features and style. When someone consider buying a car, their choice is dependent on the different features of car like maintenance cost, number of pax and doors, comfort, and safety. In this paper, we applied different data mining classification model to the car evaluation dataset to determine which types of cars is unacceptable, acceptable, good, and very good for the buyers.

**1. Introduction**

Cost, safety, and comfort are very important factors while buying cars. Depending on the model, manufacture date and company those factors change and make difference while decision making. Safety is the most important factor, also as much as conveniences which in this case is maintenance, door, and luggage boot, which one should consider for while buying a car. The attribute ‘buying’ refers to whether the cost of the is acceptable or not at the given price according to different factors of that car.

Data mining is the new technology with a lot of potential to help companies to predict future as it finds the hidden predictive in progression from very large database. Data processing tools helps to predict the future sales and trends, which helps the company to make informed decisions. Future analysis by the data processing will answer any business questions quickly and is step away from the old analysis by retrospective tools. They use hidden pattern and anticipatory information that normal data specialists may overlook because it is outside of their expectations.

A diagram of a process

Description automatically generated

Figure 1: Data Mining Process

**1.1 Dataset Attributes**

We used the Car evaluation dataset created by Marco Bohanec in 1997 which is available in UC Irvine Machine Learning repository, it is collection of specified attributes of the car.

The concept of attributes used in this dataset are:

|  |  |  |
| --- | --- | --- |
| Car (car Acceptability) | | |
| Price (overall price) | Tech | |
| Buying (buying price) | Comfort (comfort in car) | Safety (Safety in car) |
| Maint (maintenance price) | doors (no. of doors) |  |
|  | Persons (persons capacity) |  |
|  | Lug\_boot (size of luggage boot) |  |

Table 1: Dataset Attributes

Since all the attribute is not important in the dataset, removing unnecessary variables is the first step while managing the dataset. So best way to drop the attributes is to drop the one which will only distract from our aim. The car evaluation is directly relay on only six attributes i.e. ‘buying’, ‘maint’, ‘door’, ‘persons’, ‘lug\_boot’, ‘safety’.

|  |  |
| --- | --- |
| buying | vhigh, high, med, low |
| maint | vhigh, high, med, low |
| doors | 2, 3, 4, 5more |
| persons | 2, 4, more |
| lug\_boot | Small, med, big |
| Safety | Low, med, high |

Table 2: Attributes Value

The target variables classes give us the decision of the car like unacceptable, acceptable, good, very good. The class distribution is:A close up of text

Description automatically generated

Figure 2: Class Distribution

Attribute Type: Categorical

Associated Task: Classification

**2. Literature Review**

[1] used various data mining techniques to investigate the performance of various classifiers in a study on the car evaluation dataset. [2] conducted research in which they looked at customer feedback, extracted interesting patterns form the dataset, and created clusters. According to [3], the summarization task is different form the standard text summarization. Based on data mining and natural language processing methods, they proposed a set of techniques for mining and summarising product reviews. Their experimental results show that the proposed techniques are very useful in term of performance. Also, reviews are important not only to the regular customers but also to the manufacturers. The author proposed an approach for exploring large database of written customer feedback based on labelled clusters in graphical fashion in paper [4], claiming an increase in the accuracy of the used methods.

**3. Methodology**

**3.1 Data Collection**

We used the Car Evaluation Dataset from UC Irvine Machine Learning Repository. The dataset has 1728 instances, 6 features and has no missing values. We are using pandas modules to read the dataset in python environment.

**3.2 Data Preprocessing**

The dataset we used form UC Irvine repository is standard quality, clean and doesn’t have any outliers. Normally the dataset contains some missing values or more values called outliers, which effect the test and accuracy or sometimes it fails to run.

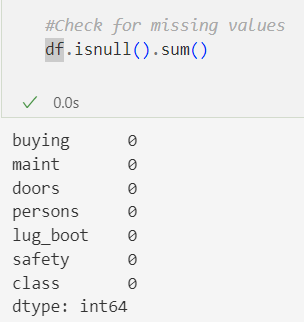


Figure3: Checking for missing value.

The above output detects the missing values in the dataset. As there aren’t any missing values in the dataset, so we don’t need to do data cleaning. Before we start doing data analysis it always better to check the dimensions of the dataset by looking at the description of the variables.

**3.2.A Exploring the attributes and variables:** In exploratory data analysis, we have to have to read the data information and find out the number of variables and cases, and their attributes datatypes.

**3.2.B Data Transformation:** When we load the dataset some of the variables doesn’t fit well in the dataset like class variables. So, for all the variables that are in string type need to be change in integer type for further operation. We converted the string type to integer by giving specific number to all the variables (Label Encoding).

Encoding Data

Description automatically generated

Figure 4: Encoding Dataset

**3.3 Data Exploration**

The process where data is analysed in visual exploration and all characteristics of the data is explained is known as data exploration. The data exploration can be done in different ways like univariate, bivariate, and multivariate.

Plot of each feature of dataset with the class

**Class**

A graph with blue bars

Description automatically generated with medium confidence

Figure 5: Class Distribution

In the above graph, out of 1728 instances of dataset 1210 (70%) is unacceptable, 384 (22%) is acceptable, 69 (3.9%) is good, and 65 (3.7%) is very good. As 70% of the cars is in unacceptable class this is skewed distribution.

**Buying**

A graph of different colored bars

Description automatically generated

Figure 6: Buying Distribution

In buying histogram the very high and high cost of car probably make the car to be unacceptable.

**Maintenance**A graph of different colored bars

Description automatically generated

Figure 7: Maintenance Distribution

The very high and high cost of car is probably making the car to be unacceptable.

**Doors**

A graph of different colored bars

Description automatically generated

Figure 8: Door Distribution

Number of doors are almost uniformly distributed however 2 doors are making the car unacceptable.

**Persons**

A graph of a bar graph

Description automatically generated with medium confidence

Figure 9: No. of person Distribution

Seating capacity is one of the most important factors in car selection. A we can see, the car with 2-person capacity is making the cars unacceptable.

**Luggage Boot Space**

A graph of a bar chart

Description automatically generated with medium confidence

Figure 10: Boot space distribution

The small luggage boot is reason of less comfort and is making the car unacceptable according to this histogram.

**Safety**

A graph of a bar chart

Description automatically generated

Figure 11: Safety Distribution

Everyone looks for safety while buying a car. As you can see, low safe car is unacceptable.

**Heat Map of Dataset**

**A screenshot of a computer screen

Description automatically generated**

**3.4 Splitting of dataset and randomization**

We divided the dataset into training and testing set with 75% training data and 25% testing data.

**3.5 Data Modelling (Classification)**

We have used different non-parametric and parametric models like Random Forest, ANN, and CNN to determine which classifier is best for our dataset. We have trained and tested the dataset and made prediction. The details process of the experiments is below:

1. **Random Forest**

Random Forest algorithm is a supervised learning algorithm that is used in regression and classification problem. I am using Random Forest classifier which is a combination of many decision tree classifier. Decision Tree on their own is weaker, so I am combining 2 decision tree to make a strong learner.

1. Training and Testing dataset

We have divided the dataset 75% for the training and 25% for the testing with the instance of 1296 and 432 respectively.

A screenshot of a computer program

Description automatically generated

Figure 12: Testing and Training dataset

1. Importing Classifier

We have imported Random Forest classifier from sklearn. We are using default hyperparameters for the model except specifying the number of decision tree (n\_estimators) and a random state value for reproducibility purposes.

A screenshot of a computer

Description automatically generated

Figure 13: Using Random Forest

1. Validation

We used accuracy score function from sklearn library to find the testing accuracy of the model. To get more details of the performance of the model, we used confusion matrix and classification report like precision, recall, f1 score, and support.

A screenshot of a computer code

Description automatically generated

Figure 14: Model Accuracy

A screenshot of a computer

Description automatically generated

Figure 15: Classification Report

A chart with a number of boxes

Description automatically generated with medium confidence

Figure 16: Confusion Matrix

1. ROC Curve

The ROC curve is used to plot the true positive rate (TPR) verses the false positive rate (FPR) at the different threshold value. The Roc curve for our model is:

A graph of a line graph

Description automatically generated with medium confidence

Figure 17: ROC Curve

1. **ANN**

We have used ANN model using TensorFlow and Keras for multiclass classification task. This model is using Adam optimizer with a learning rate od 0.01 for training and to maximize to accuracy of training and testing. We have used stratified k-fold cross-validation with 4 folds. This model is defined as:

* Input layer

The input layer has 6 input features (‘input\_dim=6’).

* Hidden layer 1

(‘dense\_1’) represent the first hidden layer with 12 nodes, ‘ReLU’ activation and ‘He normal’ as weight initialization with a factor of ‘5’.

* Hidden Layer 2

(‘dense\_2’) represent the hidden layer with 8 nodes, ‘ReLU’ activation, and ‘He normal’ initialization for weight with a factor of ‘5’.

* Output Layer

(‘dense\_3’) represents the output layer with 4 nodes for multiclass classification as the dataset has four classes (unacc, acc, good, vgood), ‘softmax’ activation, and ‘He normal’ initalizaition with a factor of ‘5’.

A diagram of a network

Description automatically generated

Figure 18: Model Architecture of ANN

In this model we have used 3 epochs, 32 batch size to train the data. Then we have calculated the accuracy score for each fold and calculated mean accuracy at last.

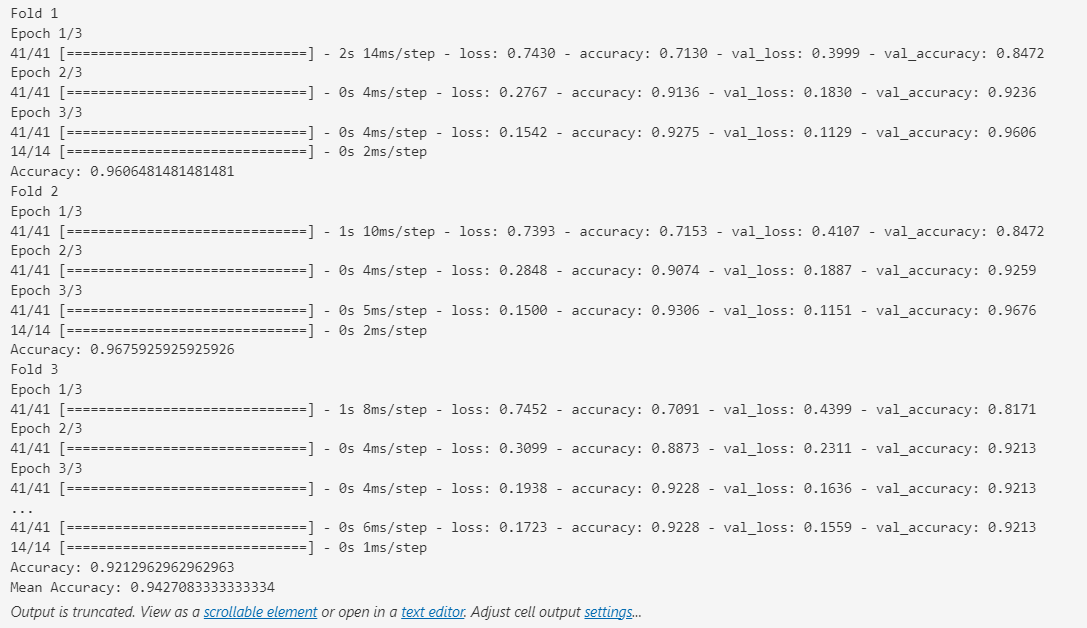


Figure 19: ANN Accuracy in 4-fold

1. Training and validation Accuracy Curve

A graph of a line graph

Description automatically generated with medium confidence

Figure 20: Training and validation Accuracy Curve

1. Model Summary

All the layers are connected densely in this model. It has 224 parameters in total.

No. of parameters = No. of Input \* No. of Neurons + No. of Neurons

A screenshot of a computer

Description automatically generated

Figure 21: ANN Model Summary

1. Accuracy curve

Model Accuracy

A graph with a line

Description automatically generated

Figure 22: Model Accuracy curve

Model Loss

A graph with a red line

Description automatically generated

Figure 23: Model Loss Curve

1. Classification Report and Confusion matrix

To get better understanding of the model we get the classification report and confusion matrix.

A screenshot of a computer

Description automatically generated

Figure 24: Classification Report of ANN

A chart with different colored squares

Description automatically generated

Figure 25: Confusion matrix

1. **CNN**

We have imported different libraries for this model like scikit-learn, TensorFlow, Keras. After loading the data, we used one-hot encoding to convert categorical labels into binary matrix. The we split the data for training and testing set and build CNN model with 1D convolutional layer, max pooling, flattening, and two dense layers. Also, the model used softmax activation function for the output layer as it is multi-class classification. we have used 25% of data for the testing and while training we have used 3 epoch and batch size of 32.

* Input layers

The input layer is the first layer of the neural network, and it is represented by 1D convolutional layer. It has 32 filters, size of convolutional layer is 3, ReLU si used as an activation function, and it process the input data with a shape (input\_shape=X.shape[1],1).

* Hidden Layer

MaxPooling1D: It perform max pooling operation and has pool size of 2 to reduce the spatial dimension of input data.

Flatten: This layer helps to flatten the output into one-dimension array and often helps to transfer from convolutional layer to densely connected layer.

Dense\_1: This is a fully connected hidden layer (dense) with ReLU as an activation function and 128 neurons.

* Output Layer

The output layer is a dense layer and the number of neuron equal to the number of classes. The softmax activation function is used and the model predict the class with the highest probability is the output.

A screenshot of a computer code

Description automatically generated

Figure 26: Creating CNN Model

1. Accuracy

CNN has the best accuracy among the model compared. We have used 3 epoch and batch size is 32.

A number and numbers on a white background

Description automatically generated

1. Classification Report and Confusion Matrix:

To get better understanding of the model we get the classification report and confusion matrix. As our dataset is small, while using the CNN we get very good accuracy.

A screenshot of a computer screen

Description automatically generated

Figure 27: Classification Report of CNN

A chart with a blue and green squares

Description automatically generated with medium confidence

Figure 28: Confusion Matrix

* 1. **Result**

The result is based on the above model analysis. As our dataset was small (1728 instance) it was easy to predict with deep learning models. We have used Random Forest, ANN, and CNN to predict the accuracy of the model. While splitting the dataset, we used 25% as a testing dataset and 75% as a training.

|  |  |
| --- | --- |
| Model | Accuracy |
| Random Forest | 90.50% |
| ANN | 92.12% |
| CNN | 92.40% |

Table 3: Accuracy Comparison

We got high accuracy in ANN and CNN than Random Forest because these are deep learning model and it perform dense calculation and produce the result.

A graph of different colored bars

Description automatically generated

Figure 29: Accuracy Comparison Graph

**4. Conclusion**

In this report, we have displayed different classifier used in car evaluation dataset. We compare parametric and non-parametric models and found that parametric models are better and has more accuracy. The main challenges in this assignment were handling the dataset as our dataset was small and we perform ANN and CNN, which is generally used in bigger dataset. So, we had to maintain the epoch and folds to avoid overfitting and get real and accurate predictions.

If we had more time, we could have thought of using big dataset which would be more suitable for ANN, CNN. Also, we could have split the dataset in different part while training and testing like 75% - 25%, 50% - 50%, or 25% - 75% and compare all the accuracies and select the better performing splits. Also, for the non-parametric model, we could have used few different models like Decision Tree, KNN, or Naïve Bayes, which will help us to find best model for the dataset.

In future, the research on this can be better to give more accurate prediction and focus on some other issue for car evaluation like chose favourite company or local company, choose the nature of feeling.

**5. Members Contributions and Comments:**

Amit Adhikari (u2044461): I am leading this group project. On our first meeting, I suggested this dataset and do the research about it and assign everyone their task and fixed the deadlines. In every meeting, I was cross checking everyone work and providing them feedback. I did the ANN and CNN coding and report and helped other members for other part of coding when needed. Everyone did great job and was very punctual.

Karuna Shahi (u2020002): I was assigned to do the coding and report non-parametric model. Also, I reported the first two meeting and helped team member when needed. For the non-parametric, I choose Random Forest because it was outside of course, and I wanted to test our skills and knowledge. I am very happy with this team as everyone did great.

Sujit Dhakal (2011202): I was assigned to do the coding and report of Exploratory Data Analysis. I did research about the data analysis and try to view and explain the dataset every possible way. Also, I reported the third meeting. I tied manual encoding and suggested it to my team for better accuracy. Our group was great, we meet all the deadlines and help each other.

Janesh Swar (2013146): I was assigned to improve the coding and helping Amit for parametric model. Also, I was assigned to do final report and scheduling meeting, taking notes and provide feedback to all the members. I was also helping other members when needed. Everyone was active, our meetings were great. We used to solve all the problems in meeting, and everyone was meeting the deadlines.

**6. Reference**

[1] Ronald F., 2004 "Decision Model for Car Evaluation Final Project in Pattern Recognition."

[2] Marko B. and Rajkovic V. 1988 "Knowledge acquisition and explanation for multi-attribute decision making." 8th Intl Workshop on Expert Systems and their Applications.

[3] Eduard A. S. and E. K. Özyirmidokuz 2015 has purposed a system named as “Mining Customer Feedback Documents” international journal of knowledge engineering.

[4] Minqing H. B. L. Department of Computer Science from University of Illinois at Chicago 851 South Morgan Street Chicago, IL 60607-7053 “Mining and Summarizing Customer Reviews.” American association for artificial intelligence.

https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

https://www.dataquest.io/blog/sci-kit-learn-tutorial/