Car Acceptability Prediction

Prepared by

Group 5

ST 3011
Group Project

Contents

Table of Figures

Introduction

The Car Evaluation dataset was created by Marko Bohanec and is available in the UCI Machine Learning Repository. It contains 1,728 records, each representing an evaluation of a car's acceptability based on different factors. These factors include the car's buying price, maintenance cost, number of doors, passenger capacity, luggage space, and safety level.

This dataset is entirely categorical, meaning that all the information is classified into different categories instead of numerical values. Because of this, it is well-suited for classification tasks, where the goal is to group cars into categories such as unacceptable, acceptable, good, or very good. This dataset is useful for understanding how different features influence car acceptability and can help in decision-making when choosing a car.

Attributes

There are 7 categorical features.

- **buying**: Buying price (v-high, high, med, low)
- **maint**: Maintenance cost (v-high, high, med, low)
- **doors**: Number of doors (2, 3, 4, 5-more)
- **persons**: Passenger capacity (2, 4, more)
- **lug_boot**: Luggage boot size (small, med, big)
- **safety**: Safety level (low, med, high)
- **class**: Overall car acceptability, categorized into four levels.(unacc, acc, good, vgood) This is the target variable

Aim of the Analysis

The purpose of this analysis is to understand how different features of a car affect its acceptability. We will use factors such as the car's price, maintenance cost, number of doors, passenger capacity, luggage space, and safety level to predict whether a car is classified as unacceptable, acceptable, good, or very good.

Additionally, we will find out what predictor variables have a significant relationship with the response variable. By analyzing this we can find out what factors play the most crucial role in car acceptability.

Exploratory Data Analysis

Here we will explore the distributions of variables and the relationships between those variables.

Univariate Analysis

➢ Target variable

car acceptability (class)

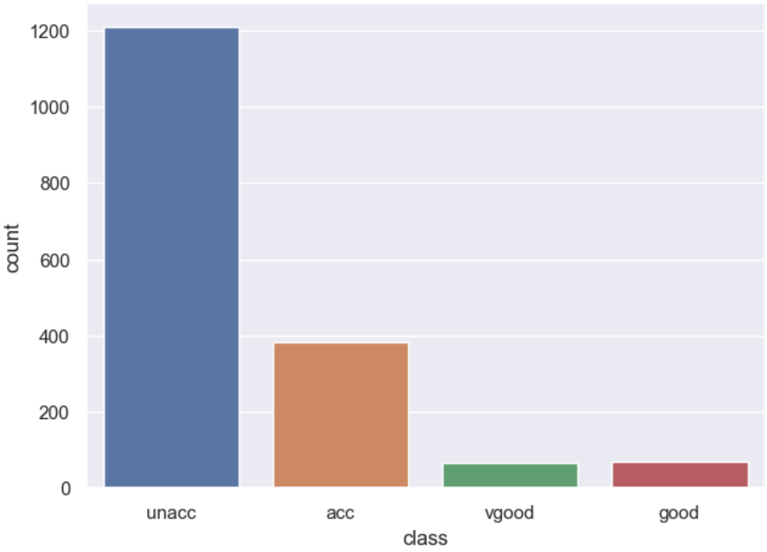


Figure 1 - Bar chart of car acceptability (Target variable)

This bar chart represents the count distribution of the variable "class" with four categories: unacc, acc, vgood, and good.

- **Unacc (Unacceptable)**: This category has the highest count, exceeding 1200 observations. It indicates that the majority of instances fall into this category.
- **acc (Acceptable)**: The second most common category, with a count significantly lower than unacc but still notable.
- **vgood (Very Good)**: This category has a small count, showing relatively few instances.
- **good (Good)**: Similar to vgood, the count is relatively low, indicating limited representation of this category.

The data is highly imbalanced, with the unacc category dominating the distribution. This might struggle to predict minority classes like vgood and good.

➢ Predictor variables

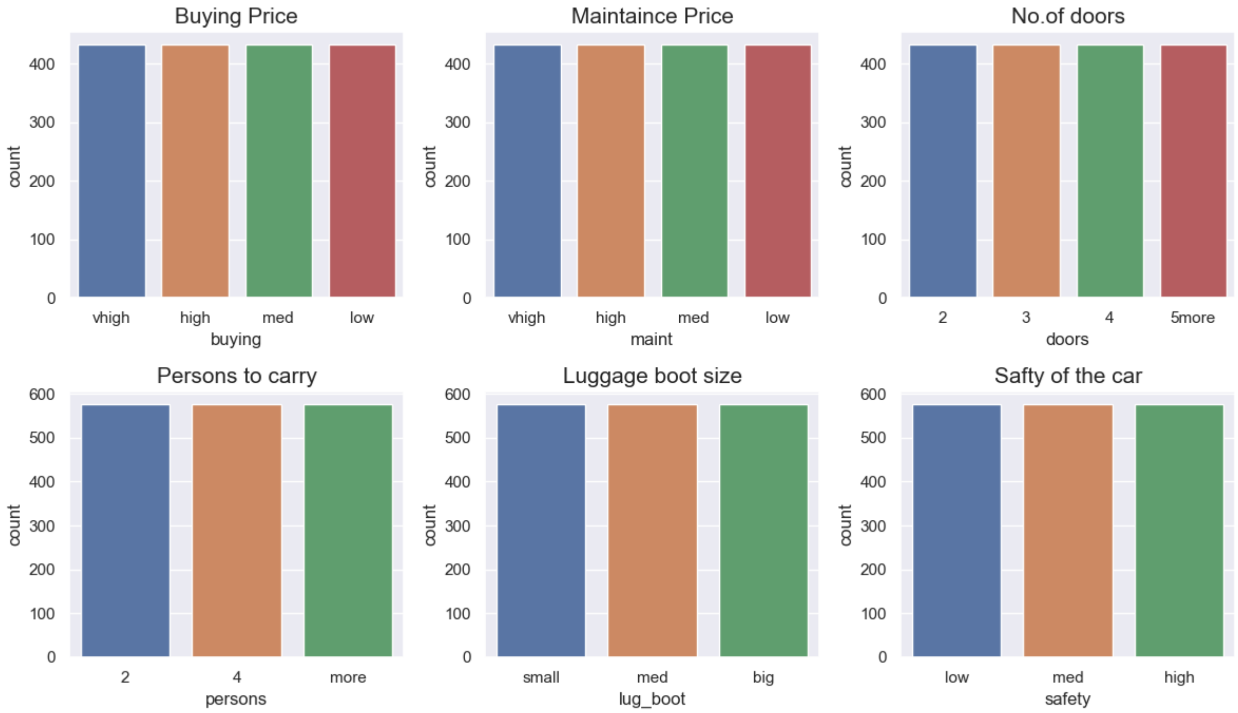


Figure 2 - Bar chart of predictor variables

- **Buying Price:** The Categories are vhigh, high, med and low. The counts are roughly uniform across all buying price categories. This indicates that the dataset has a balanced distribution for cars across different price ranges.
- **Maintenance Price:** The Categories are vhigh, high, med and low. Similar to the buying price, the maintenance price categories have approximately equal representation. This suggests a balanced distribution in the dataset.
- **Number of Doors:** The Categories are 2, 3, 4 and 5 more. The count for each category is evenly distributed. This means the number of doors does not show any significant skew in the dataset.
- **Persons to Carry:** The Categories are 2, 4 and more. The dataset appears to have an equal number of cars capable of carrying 2, 4 or more persons. This indicates no bias toward cars with different seating capacities.
- **Luggage Boot Size:** The Categories are small, med and big. Counts are evenly distributed across the three luggage boot size categories. This shows no dominant preference for luggage space among the cars.
- **Safety of the Car:** The Categories are low, med and high. Safety levels (low, medium and high) are balanced. This implies that the dataset includes cars with a variety of safety ratings without any single category dominating.

Across all variables, the dataset appears balanced. Each category receives a similar number of counts. This ensures that the data will not be biased toward a specific category.

Bivariate Analysis

➢ Car buying categories across different classes

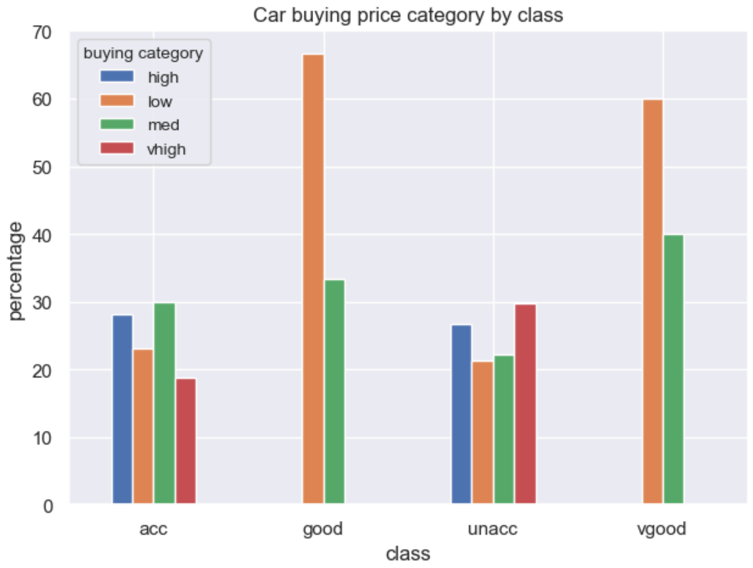


Figure 3 - Clustered bar graph of Car buying price by Class

This grouped bar chart illustrates the distribution of car-buying price categories (high, low, med, vhigh) across the four classes (acc, good, unacc, and vgood).

- **acc (acceptable) class:** The distribution is relatively balanced across high, low, and med categories. Slightly fewer cars in the vhigh category.
- **good class**: The low buying price category is high in this class, comprising around 70% of the total. med is the second most represented.
- **unacc (unacceptable) class**: The distribution appears more even across all buying categories compared to other classes. However, the vhigh category is slightly more common than others.
- **vgood (very good) class**: Similar to the good class, the low buying price category is high. med is the second largest group.

The low buying price category is a strong predictor for the good and vgood classes. This indicates that cars in these classes are often in the low price range. The vhigh buying price category is more frequently associated with the unacc class.This suggests cars with very high prices are often deemed unacceptable. The acc class exhibits a more balanced distribution across all buying categories.

➢ Safety categories across different classes

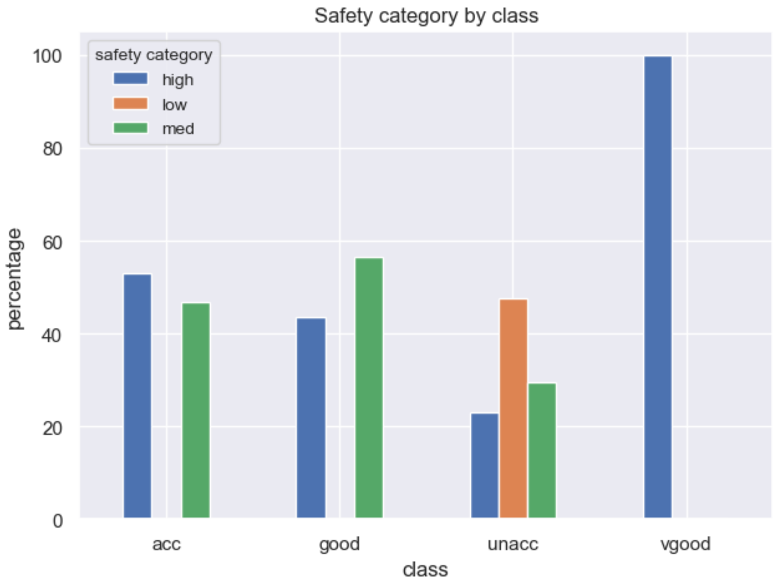


Figure 4 - Clustered bar graph of Safety by Class

The chart illustrates the distribution of **safety categories** (high, medium, low) across different **classes** in terms of **percentage**.

- **High safety** is most prominent in the **vgood** class and significantly present in **acc** and **good** classes but minimal in the **unacc** class.
- **Low safety** is dominant only in the **unacc** class. This suggests that safety concerns are linked to unacceptable classifications.
- **Medium safety** is the leading category in the **good** class and moderately present in the **acc** class but barely exists in others.

The chart highlights a clear relationship between safety categories and class ratings. Better classes (**vgood**) are associated with higher safety standards, while **unacc** aligns strongly with lower safety ratings.

➢ Door categories across different classes

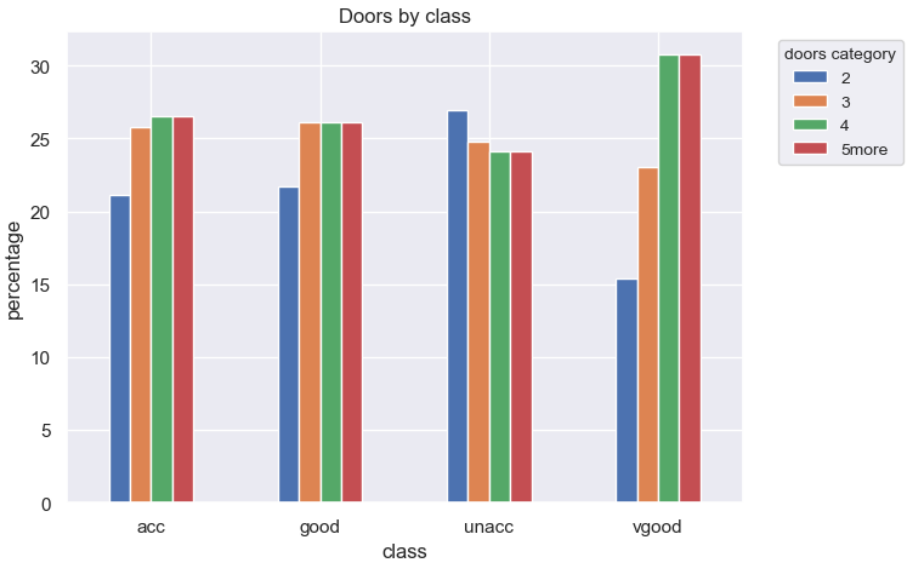


Figure 5 - Clustered bar graph of Door categories by Class

This bar plot shows the percentage distribution of door categories across different classes.

- **Class 'acc' (Acceptable)**: All door categories (2, 3, 4, and 5 more) have nearly equal percentages. This indicates that cars with any number of doors are equally acceptable within this class.
- **Class 'good'**: The percentages of the door categories (2, 3, 4, and 5more) are also nearly equal, similar to the acc class. This suggests that the number of doors has no significant influence on a car being categorized as good.
- **Class 'unacc' (Unacceptable)**: Cars with 2 doors are more frequently categorized as unacc (over 25%). Cars with 3, 4, or 5more doors have almost equal percentages.
- **Class 'vgood' (Very Good)**: Cars with 4 doors and 5more doors have the highest percentage in this class. Cars with 2 doors are rare in the vgood class. This indicates a preference for more doors in the very good category.

The number of doors seems to have some influence on the quality classification of a car, especially for vgood.

➢ People to carry across different classes

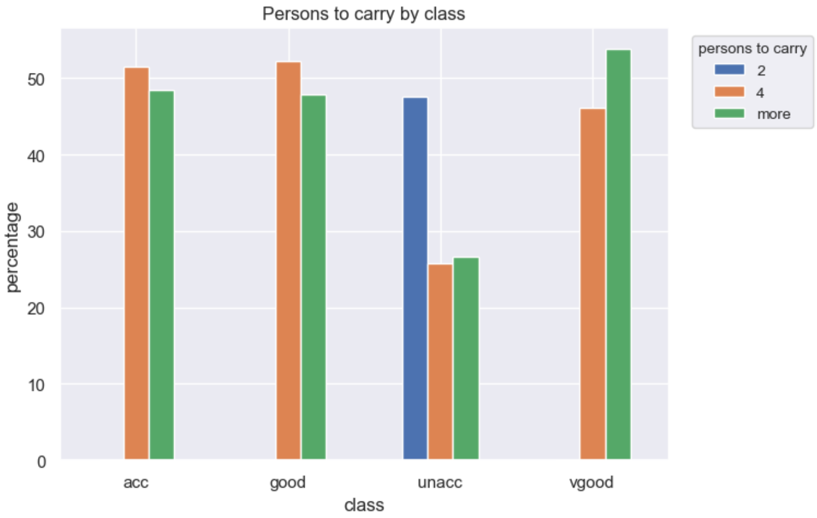


Figure 6 - Clustered bar graph of People to carry by Class

The plot represents the percentage distribution of cars categorized by the number of people they can carry (2, 4, more) across the car evaluation classes.

- **Class acc (Acceptable)**: Cars that can carry 4 persons are the most common in this class. Cars that can carry more people also have a significant proportion. Cars that carry only 2 persons are absent from this category.
- **Class good**: Similar to the acc class, cars that can carry 4 persons dominate. Cars that can carry more people also account for a substantial percentage. Again, cars that carry only 2 persons do not appear in this class.
- **Class unacc (Unacceptable):** Cars that can carry only 2 persons are overwhelmingly dominant. This makes up the largest share of this class. Cars that can carry 4 or more persons are significantly less represented.
- **Class vgood (Very Good)**: Cars that can carry more persons dominate this category. This makes them the most likely to be classified as vgood. Cars that can carry 4 persons also form a significant proportion. Cars that carry only 2 persons are entirely absent from this category.

Cars that can only carry 2 persons are mostly classified as unacc. They are absent in the acc, good, and vgood categories. Cars that can carry 4 persons are consistently well-represented in the acc, good, and vgood categories. Cars that can carry more people have a strong positive correlation with the class vgood.

➢ Luggage boot size across different classes

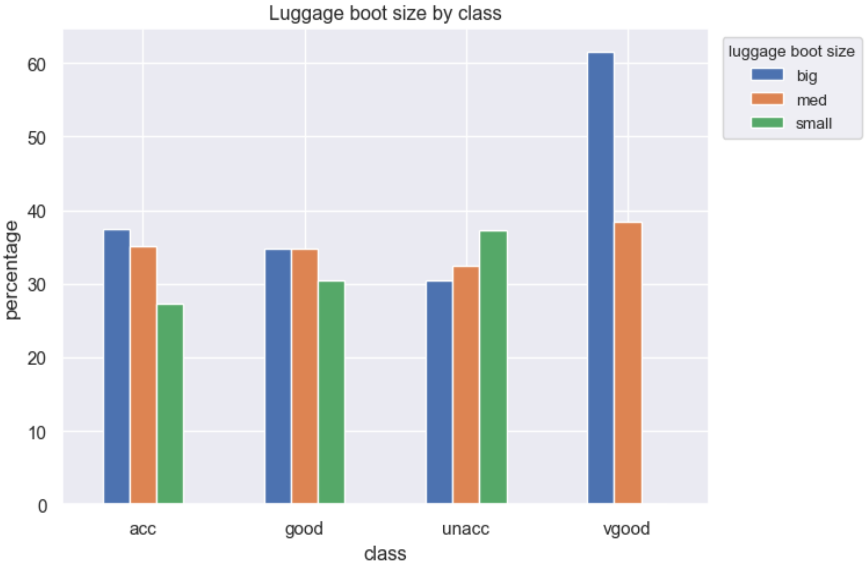


Figure 7 - Clustered bar graph of Luggage boot size by Class

The plot represents the percentage distribution of cars categorized by luggage boot size (big, medium, small) across the car evaluation classes.

- **Class acc (Acceptable):** Cars with big and medium luggage boot sizes are the most common in this category, with similar proportions. Cars with small luggage boot sizes have the lowest representation.
- **Class good:** The distribution is similar to the acc class, where big and medium luggage boot sizes dominate, and small has a lower but still notable proportion.
- **Class unacc (Unacceptable):** Cars with small luggage boot sizes are the most frequent in this class. Followed by the medium category, while big luggage boot sizes have the lowest share.
- **Class vgood (Very Good):** Cars with big luggage boot sizes dominate this category, making them the most likely to be classified as vgood. Medium luggage boot sizes also have a significant proportion.

Cars with big luggage boot sizes are strongly associated with the vgood class, while small luggage boot sizes are mostly found in unacc cars and are absent in vgood. Medium luggage boot sizes are consistently present across acc, good, and vgood categories. This suggests that a larger luggage boot size is a key factor in higher car acceptability.

➢ Maintenance cost across different classes

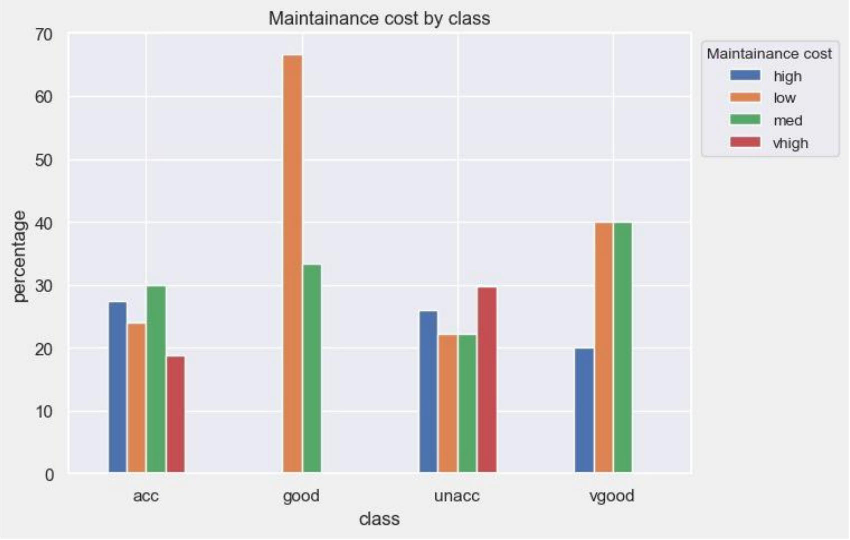


Figure 8 - Clustered bar graph of Maintenance cost by Class

The plot represents the percentage distribution of cars categorized by maintenance cost(v-high, high, med, low) across the car evaluation classes.

- **Class acc (Acceptable):** Cars with medium maintenance cost are the most common in this category. All four categories show approximately similar percentages while very high maintenance cost has the lowest representation.
- **Class good:** Only the cars with low and medium maintenance costs have fallen into this category while the highest percentage is represented by low maintenance cost and there's a huge difference between the two percentages.
- **Class unacc (Unacceptable):** Cars with very high maintenance cost are the highest in this class while cars with high maintenance cost are the second highest. Cars with low and medium maintenance costs have similar percentages.
- **Class vgood (Very Good):** Cars with low and medium maintenance costs have similar percentages and cars with high maintenance cost are the lowest. Cars with very high maintenance costs haven't fallen into this category.

Cars with high maintenance costs are strongly associated with the good class. Very high maintenance costs are mostly found in unacceptable cars and are absent in good and good classes. High maintenance costs are mostly found in acceptable cars and are absent in good class. Medium maintenance costs are mostly found in the very good class and lowest in the unacceptable class. This suggests that cars with low maintenance costs are a key factor in good car acceptability.

Advanced Methodology

This section aims to build a predictive model to predict the car class based on its features. For this, we will train a logistic regression model. Logistic regression is a widely used statistical model for classification tasks. Since the target variable consists of four categories, we use multinomial logistic regression which is an extended version of binary logistic regression.

Why Logistic Regression?

Logistic regression is a preferred baseline model when it comes to classification because of its simplicity, interpretability and statistical foundation. We can interpret the model using coefficients to show how each predictor influences the likelihood of car acceptability. Interpretability is crucial when talking about car acceptability because it allows us to understand which factors contribute the most to determining whether a car is classified into a certain category.

Advantages of Logistic Regression

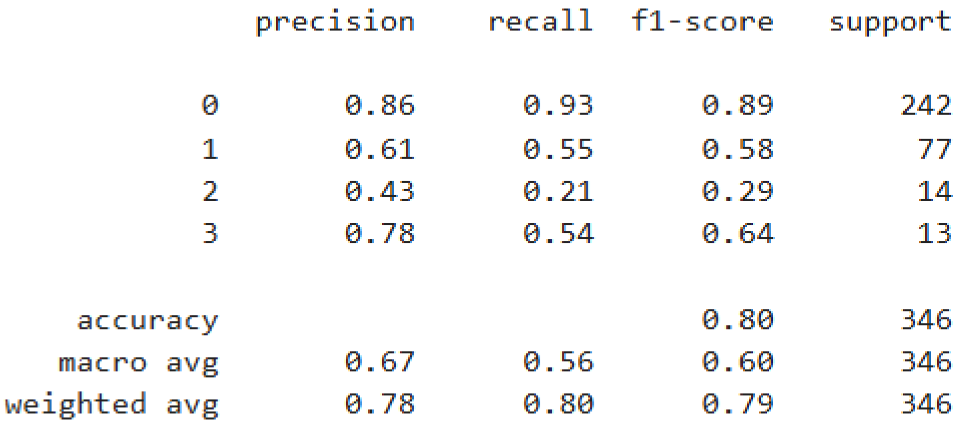
- **Interpretable:** using coefficients we can get to know how each factor affects car classification
- **Statistically robust:** Logistic regression does classification based on probability theory. Hence it allows statistical testing of the significance of predictor variables using p-values
- **Computationally efficient:** Faster training than more complex models, requiring less resources

Results with Discussion

Here we will implement the selected logistic regression and evaluate its performance using performance matrices like accuracy and F1 score.

Since all the columns are categorical, we have to encode them before feeding the data to the models. For that, the ‘OrdinalEncoder’ method in ‘category_encoders’ library will be used to encode the categorical variable because all of them are in the ordinal level. (categories have an order e.g: ‘low’, ‘middle’, ‘high’)

After encoding the variables, we divide the dataset into training and testing sets such that 80% of the observations are for training set and 20% of the observations are for the testing set. To divide the dataset, we use the ‘StratifiedShuffleSplit’ method in the ‘sklearn’ library to ensure the proportions of categories in the response variable are same in both the training set and testing set. After forming the train set and test set, we first trained a simple logistic regression model on the train set and evaluated its performance on the testing set.



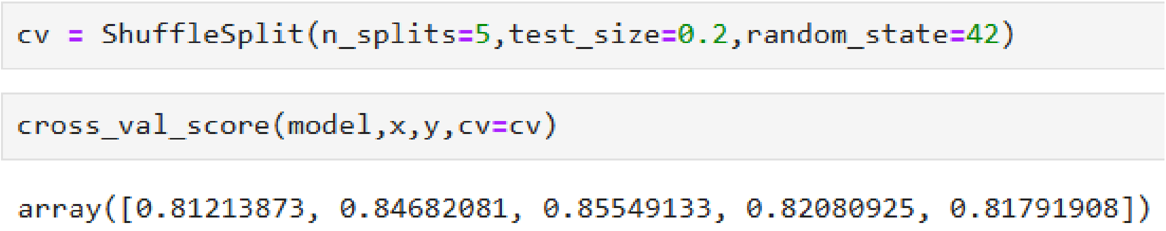
```
              precision    recall  f1-score   support

           0       0.86      0.93      0.89       242
           1       0.61      0.55      0.58        77
           2       0.43      0.21      0.29        14
           3       0.78      0.54      0.64        13

    accuracy                           0.80       346
   macro avg       0.67      0.56      0.60       346
weighted avg       0.78      0.80      0.79       346
```

Figure 9 - Output for simple logistic regression model

This is the classification report of the logistic regression model. Although the accuracy is 80%, the F1 scores for some categories are very low. Meaning that the model does not perform well in those categories. Also performed cross-validation along with shuffle split and the results are as follows.



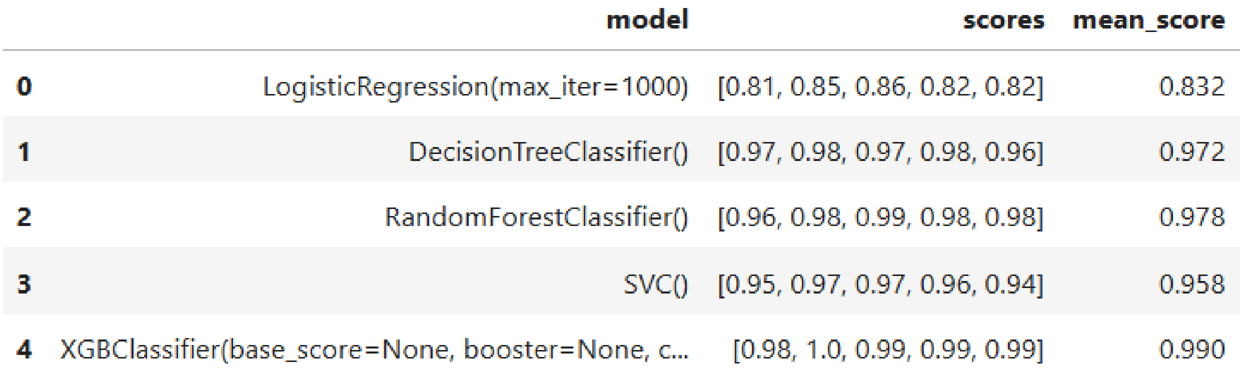
```
cv = ShuffleSplit(n_splits=5,test_size=0.2,random_state=42)
```

```
cross_val_score(model,x,y,cv=cv)
```

```
array([0.81213873, 0.84682081, 0.85549133, 0.82080925, 0.81791908])
```

Figure 10 - Output for cross validation along with shuffle split

So, we will test some other machine learning models to check what gives us the best results. For that, we will use cross-validation to evaluate each model and get the mean scores of them to find the best performing model.

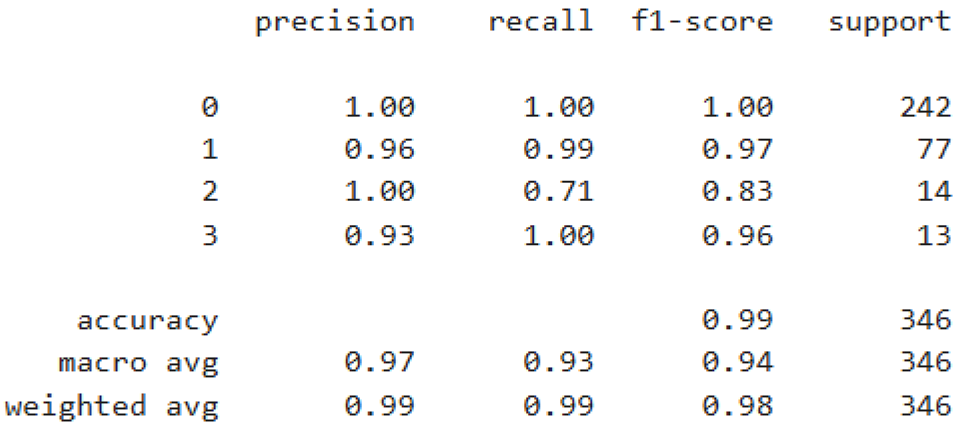


	model	scores	mean_score
0	LogisticRegression(max_iter=1000)	[0.81, 0.85, 0.86, 0.82, 0.82]	0.832
1	DecisionTreeClassifier()	[0.97, 0.98, 0.97, 0.98, 0.96]	0.972
2	RandomForestClassifier()	[0.96, 0.98, 0.99, 0.98, 0.98]	0.978
3	SVC()	[0.95, 0.97, 0.97, 0.96, 0.94]	0.958
4	XGBClassifier(base_score=None, booster=None, c...	[0.98, 1.0, 0.99, 0.99, 0.99]	0.990

Figure 11 - Evaluation of each model

Trained the model on the decision tree, random forest, support vector machine and extreme gradient boosting. And got an excellent accuracy of 99% from the extreme gradient boost model. So that model was selected as the best performing model.

Extreme gradient boosting model is trained on the training test and evaluated using the testing set and the results are as follows.



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       242
           1       0.96      0.99      0.97        77
           2       1.00      0.71      0.83        14
           3       0.93      1.00      0.96        13

    accuracy                           0.99       346
   macro avg       0.97      0.93      0.94       346
weighted avg       0.99      0.99      0.98       346
```

Figure 12 - Evaluation of the XGB model

Results suggest not only the accuracy is excellent, but also that the F1-scores for every category are good. Therefore, **Extreme Gradient Boosting** was selected as the final model.

Statistical Inference

Since all variables in the dataset are categorical, the chi-square test for independence is used to determine whether there is a significant association between categorical variables. Here all predictor variables are tested against the response variable to check whether there's a significant association between them. The test results are as follows

Variable	Chi-Square test statistic	Degrees of freedom	p-value
safety	479.322	6	2.38916e-100
buying	189.243	9	5.92806e-36
maint	142.941	9	2.54765e-26
doors	10.3848	9	0.320242
persons	371.337	6	4.03997e-77
lug_boot	53.282	6	1.02944e-09

Figure 13 - Chi-square test results for predictor variables across response variable

The hypotheses of every chi-square test are,

H_0: There is no association between two variables (two variables are independent of each other)

H_1: There is an association between the two variables (two variables are dependent on each other)

We reject the null hypothesis if the p-value is less than the significance level. If we perform the tests on 5% significance value, only the ‘door’ variable’s p-value is greater than the significance level. Therefore at 5% significance level, we have enough evidence to conclude that there’s no relationship between the number of doors and car class. And since the p-values of other variables are less than the significance value, they have a significant relationship with the response variable (car class).

Conclusion

Univariate Analysis

- Majority of cars are in the unacceptable class category in the dataset which indicates a high imbalance of the response variable. All predictor variables (buying price, maintenance cost, number of doors, passenger capacity, luggage space, and safety level) are balanced as all the categories of each variable are evenly distributed.

Bivariate Analysis

- Cars with low buying prices are more likely to be in good and very good classes while cars with high and very high buying prices are associated with only acceptable and unacceptable classes. Cars with medium buying prices are approximately evenly distributed across all classes.
- Cars with high safety level are highly associated with very good class while cars with low safety level are associated with unacceptable class.
- All door categories are likely to be evenly distributed across good, acceptable, and unacceptable classes while cars with four, and five or more doors are highly associated with very good class and cars with only two doors are weekly associated with very good class.
- Cars that could carry more passengers are more likely to be in very good class while cars that could carry only two passengers are mostly in unacceptable class.
- Cars with big luggage boot size are more likely to be in very good class, and cars with small luggage boot size are fallen into only good, acceptable, and unacceptable classes. Cars with medium luggage boot size are likely to be evenly distributed across all classes.
- Cars with lower maintenance costs are more likely to be in good class while high maintenance costs are linked to unacceptable class.

Model Building

- Several machine learning models were trained and evaluated, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, and Extreme Gradient Boosting in order to find out the best fitted model for the dataset.
- With the 99% of accuracy level and good F1-score for all categories, Extreme Gradient Boosting was selected as the final predictive model for car acceptability.

Statistical Inference

- Chi-square tests for independence were conducted to assess the relationship between each predictor variable and the response variable (car acceptability). All variables except the ‘number of doors’ showed a significant association with car acceptability at a 5% significance level.
- This suggests that the number of doors has little to no influence on a car's acceptability, while other factors like safety, buying price, maintenance cost, passenger capacity, and luggage boot size play a more critical role.

In conclusion, this analysis provides valuable insights into the factors that influence car acceptability and offers a robust predictive model for classifying cars based on their features. The findings can guide both consumers and manufacturers in making informed decisions regarding car evaluations and improvements.