Dhanya Sri Vasantha(G01461366)

Car Evaluation Analysis

I. Project description

I chose car evaluation analysis dataset for this project. The main purpose of choosing this project is to get a clear understanding of the change in factors for the acceptability of the car. There are many factors involved in accepting a car. Some people don't know what factors need to be considered to select a desired car. This analysis gives a rough estimation of how to select the car with their personal requirements.

The factors which change the acceptability are discussed in the research questions with proper analysis of the data with different models. The most influencing factors are being discussed in the project. Making an informed selection requires assessing an automobile's condition before buying. It takes a lot of effort and labor to manually distinguish an automobile in good or acceptable condition from one in poor shape [6]. In this project, I will evaluate several automotive qualities on a physical level and then help or advise a user in making decisions based on those attributes.

There are 3 research questions. In which I used 2 types of models to answer them, and a few exploratory analyses are used. I chose decision tree and random forest models to design the required graphs.

II. Dataset

The dataset I chose is car evaluation analysis which is taken from UCI machine learning repository [7]. The Car Evaluation Database was created using a straightforward hierarchical decision model that was first created by M. Bohanec and V. Rajkovic to demonstrate DEX: Expert decision-making system. This dataset contains 7 variables, and 1728 instances.
The 7 variables namely

1. **Buying price:** It is named as 'buying' in the csv file. The attributes it has are Very high as vhigh, high as high, Low as low and medium as med.
2. **Maintenance level:** It is named as maint in the csv file. The attributes it has are Very high as vhigh, high as high, low as low and medium as med.
3. **Number of doors:** It is named as doors in the csv file. The attributes it has are 2, 3, 4, 5 or more.
4. **No of persons capacity:** It is named as persons in the csv file. The attributes it has are 2, 4, and more.
5. **Luggage boot space:** It is named lug_boot. The attributes are small, med and big.
6. **Safety level:** It is named as safety in csv file. The attributes are high, med and low.
7. **Class:** This is the main variable which shows the acceptability of a car. It is named as class in csv file. The attributes are unacceptable as unacc, acceptable as acc, very good as vgood, and good as good.

Reading the data information and then investigating the key aspects is the first step in data exploratory analysis. Gaining an understanding of the number of variables, cases, attribute datatypes, and possible values for each attribute is crucial.

III. Exploratory analysis

For the exploratory analysis, I compared target variable to the remaining 6 variables. By this I can know the change in patterns of acceptance of cars based on each variable.

The comparison of all 6 variables with class variable is shown below.

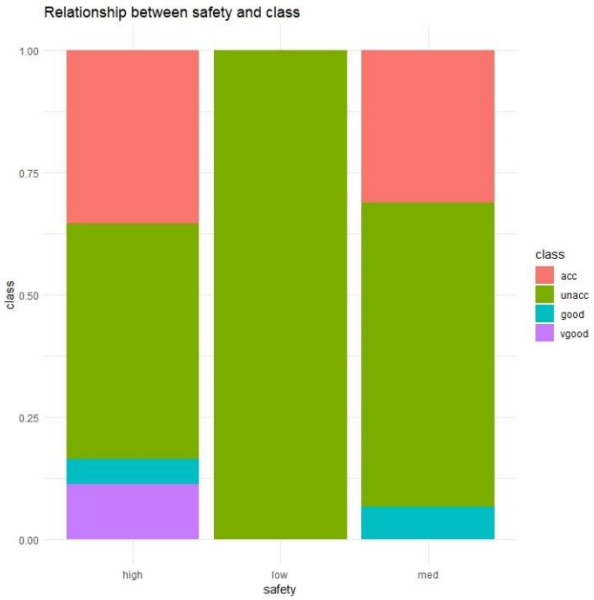

Figure 1: Relation between safety and class

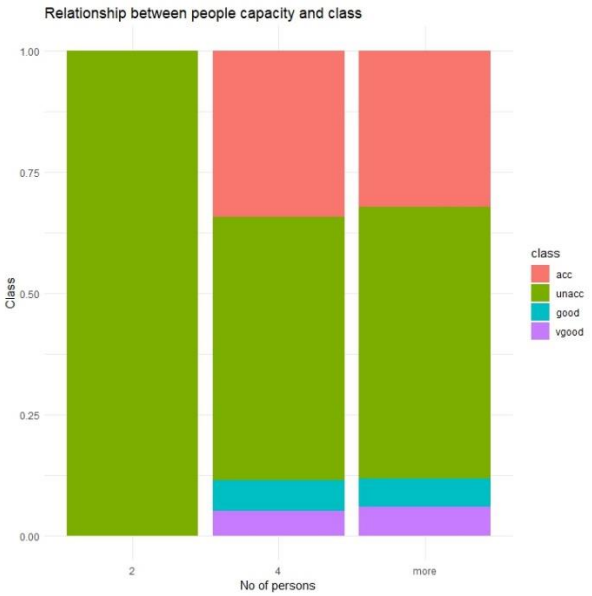

Figure 2: Relation between people capacity and class

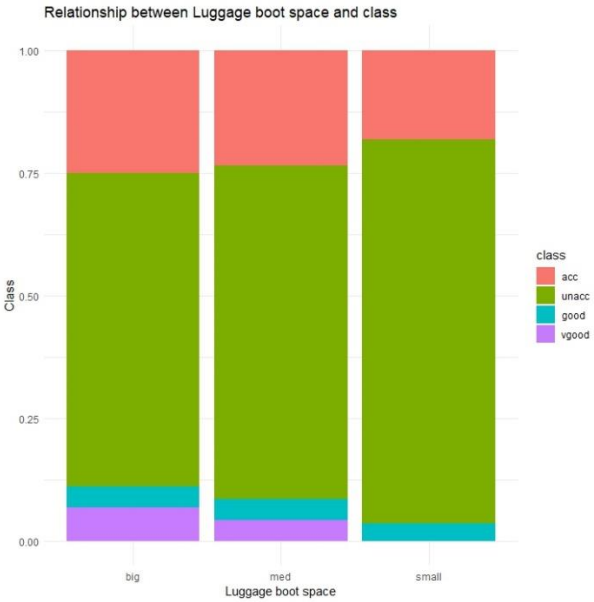

Figure 3: Relation between Luggage boot space and class

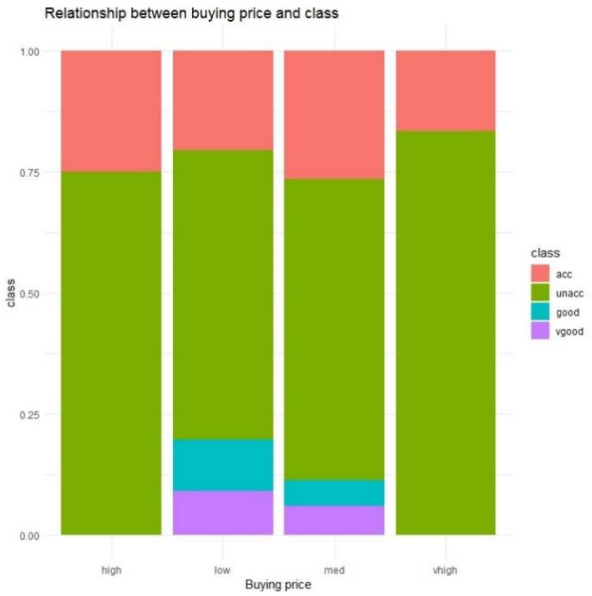

Figure 4: Relation between Buying price and class

Figure 1 shows the relation between safety of a car and class variable. From that graph, I can see that a car with low safety is unacceptable for buying. Only a few cars with high safety have very good acceptance. There is no single car with medium safety has very good acceptance. Both medium and high safety cars have same amount of acceptable class. Which means high level safety is preferred.

Figure 2 shows the relation between the no of people capacity and class variable. In which cars with 2 people capacity are unacceptable to buy. Cars with 4 and more people capacity have a higher acceptance rate.

Figure 3 shows the relation between luggage boot space and class. Cars with small boot space are mostly unacceptable to buy but in some cases like low price and high safety it is acceptable and good to buy them. Small boot space doesn't have very good acceptance. Only cars with big and medium boot space have very good acceptance.

Figure 4 shows the relation between the buying price of a car and class variable. Cars with very high and high buying prices are unacceptable in most of the cases and sometimes acceptable to buy. Only cars with low and medium prices have very good acceptance.

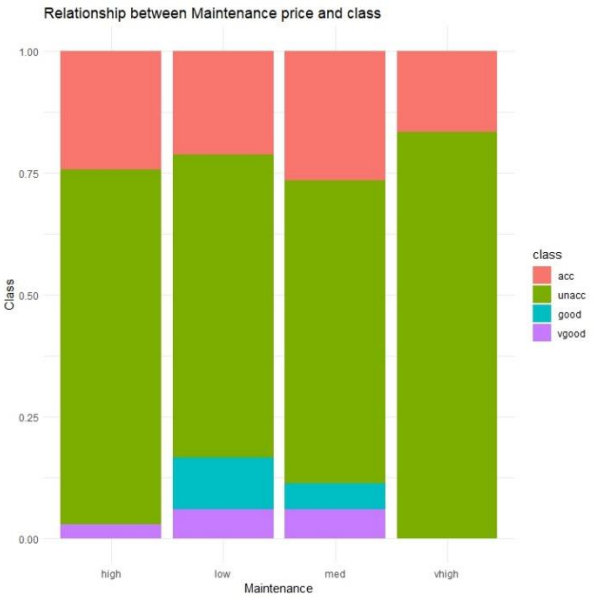

Figure 5: Relation between maintenance price and class

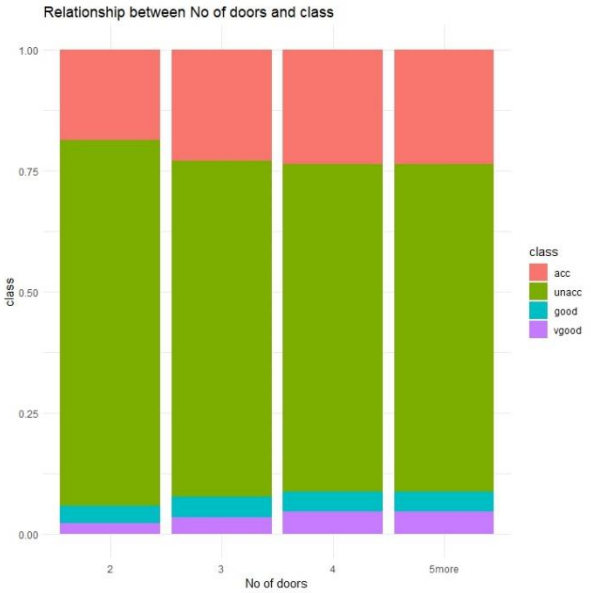

Figure 6: Relation between No of doors and class

Figure 5 shows the relation between maintenance price and class variable. In this comparison, cars with very high maintenance have no single car with very good acceptance. High maintenance cars have very a smaller number of cars with very good acceptance. Low and medium maintenance cars have almost same amount of very good acceptance. But the low maintenance cars have a greater number of good acceptances.

Figure 6 shows the relation between no of doors and class variable. The acceptance rate doesn't vary much based on the number of doors. Cars with 2 doors have less acceptance compared to cars with 3 or more doors.

IV. Research questions

1. What are the most influential features affecting the evaluation level (class variable) of a car?
2. Investigate the relationship between the price of a car and its safety level. Do higher-priced cars generally have better safety ratings?
3. Analyze how the size of the luggage boot (small, med, big) affects the buying preferences. Are the cars with larger boot spaces more likely to be in certain buying categories?

V. Data analysis: methods used, software used and results.

I used R studio for data analysis and plotting graphs [8]. R is a free software environment for which offers a collection of powerful tools for statistical computing, data analysis and visualization. R studios plot high quality graphs, and it is an open source.
I used many libraries. The details of the libraries are mentioned below.

1. **GGPLOT2**: I used ggplot2 library from r language to plot the exploratory figures [2]. Ggplot2 is a well-liked R package for producing sophisticated data visualizations. Its foundation is the Grammar of Graphics, which enables you to construct intricate plots using a multi-layered method. The key features of this library are declarative approach,

layered graphics, grammar of graphics, wide range of plots, easy customizations, extensive documentation and community.

2. **RandomForest**: I used RandomForest library to make prediction of the most influencing variable [3]. For a variety of machine learning tasks, the random forest library is an effective tool, especially for classification and regression. It is renowned for its dependability, accuracy, and adaptability. The key features are random feature selection, feature importance, high accuracy and robustness, handles high-dimensional data, less susceptible to overfitting and provides feature importance.
3. **Caret**: I used caret library in plotting decision graph [4]. The R package known as the "caret package," which stands for "Classification and Regression Training," is an effective tool for expediting the creation of predictive models. The key features are data splitting, pre-processing, feature selection, model tuning, variable importance, model comparison.
4. **DPLYR**: I used dplyr library in data manipulation [5]. Dplyr facilitates the creation of clear and understandable code by offering a collection of "verbs" that stand in for typical data manipulations. Dplyr is a member of the tidyverse, a group of R packages with a common design ethos. The key features are chaining verbs, lazy evaluation, integration with other tidyverse packages, supports data frame and other data types.
5. **Rpart:** I used rpart to build decision tree [9]. Rpart is a robust R package for creating regression and classification trees. It uses a technique called recursive partitioning, which divides data recursively into progressively homogeneous subsets according to features. The key features are predicting target variables, identifying important features, visualizing data relationship.
6. **Rpart.plot:** I used this library to plot the decision tree [10]. An R package called rpart.plot offers improved plotting capabilities for models created using the rpart package. It adds more features to rpart's basic plotting capabilities and allows for more customization and tree visualization options. The key features are multiple plot types, customization options, enhanced annotations, integration with other packages which means it works seamlessly with other packages like ggplot2 and more.

Result of research question 1:

The question is to identify the most influential feature or variable in the dataset. I used two models to identify the most influential variable. The first one is the decision tree.

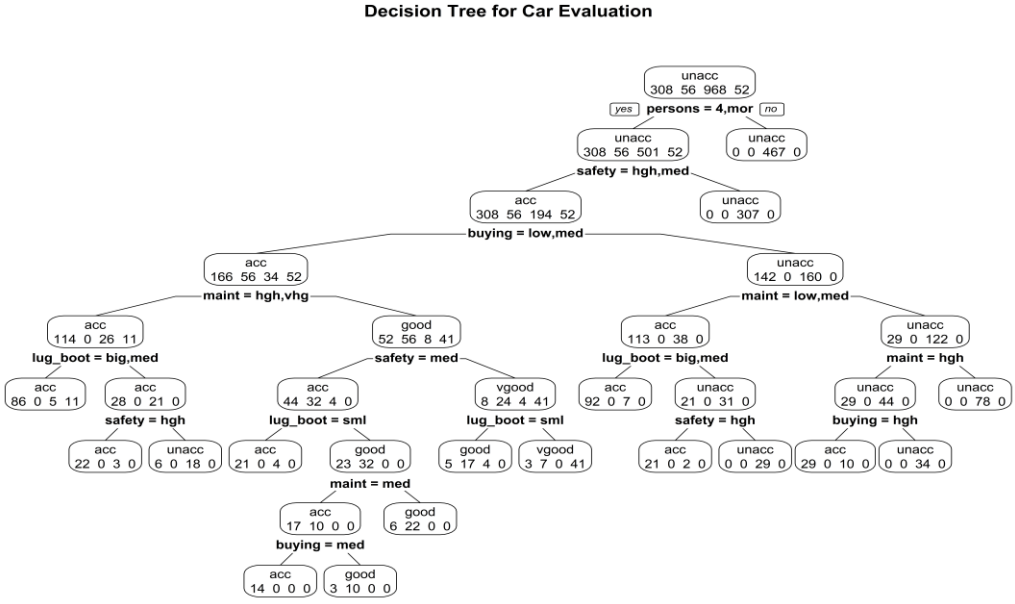

Figure 7: Decision Tree

A decision tree is a supervised learning algorithm in R that can be applied to tasks involving regression and classification. It is a tree-like structure that divides the data into progressively smaller subsets through a sequence of binary splits. The objective is to identify splits that maximize the separation between the various classes or target values. Each split is based on a single feature. The decision tree contains a root node, a leaf node and branches as decision nodes.

I have plot figure 8 as the improvement to figure 7 by adding color pattern for easy identification.

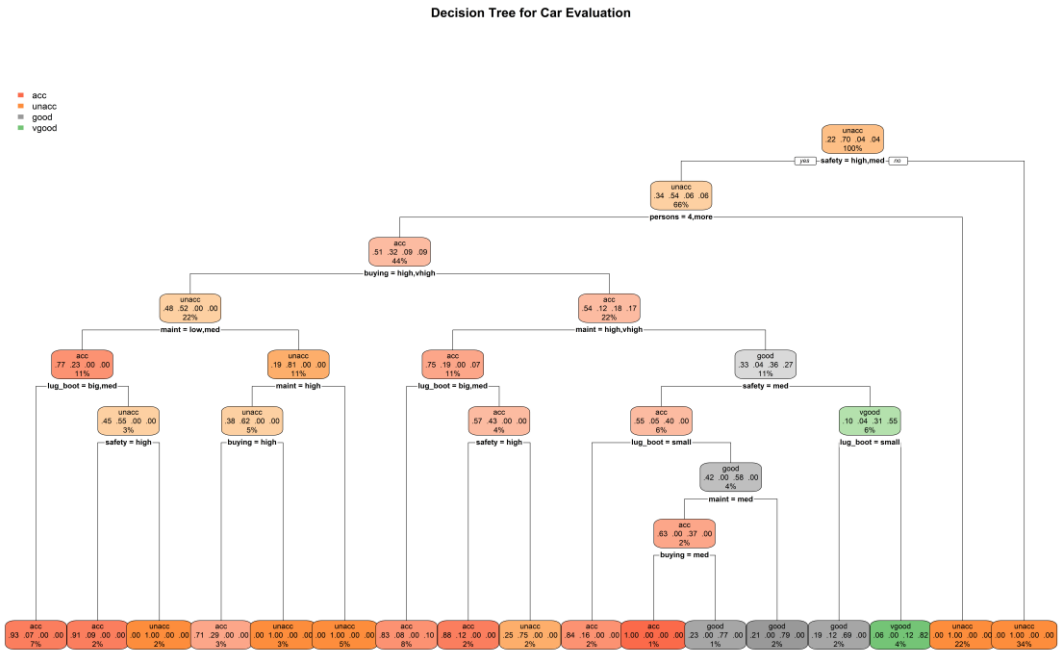

Figure 8: Decision tree with enhanced colors

The variable at root node is the most influential variable. From figure 7 and figure 8, the root node is safety, and the 2nd root node is persons. Which means, the most influential variables are persons and safety.

To get an accurate result I decided to plot the graph using other models. I chose random forest because it has high accuracy.

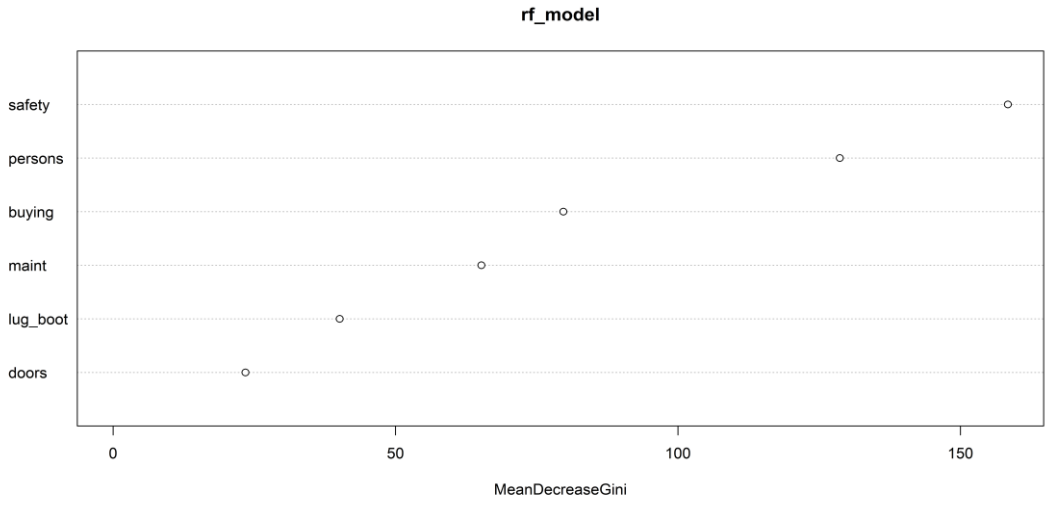

Figure 9: Random Forest Model

The second model I chose is randomForest. Based on the idea of ensemble learning, Random Forest is a potent machine learning algorithm. In comparison to individual trees, it achieves better generalization and performance by combining multiple decision trees. RandomForest is more accurate than decision tree. RandomForest combines multiple trees to give one result, which is the main reason for its accuracy.

Variable	Mean Decrease Gini
Safety	158.41239
Persons	128.65132
Buying Price	79.71534
Maintenance	65.19808
Luggage boot space	40.11610
No of Doors	23.43571

Table 1: Mean Decrease Gini index of all variables

From figure 9, the rf model shows that safety and persons are the most influential variables. The degree to which each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest is indicated by the mean decrease in the Gini coefficient. Comparing both the models, the top 2 influential variables are safety and persons.

The partial dependence of all the variables is plotted as follows.

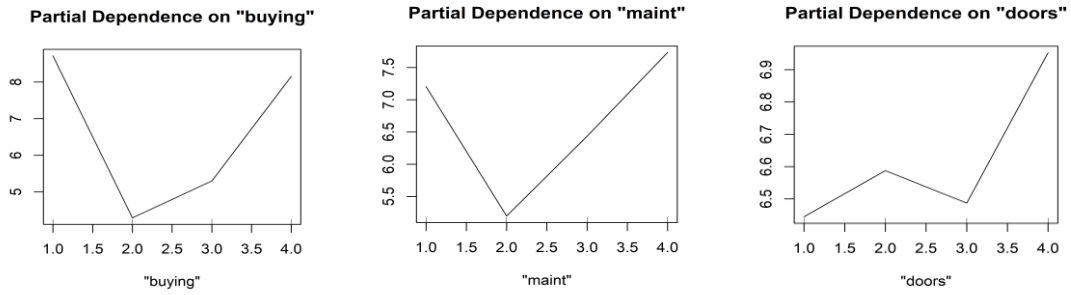

Figure 10: Partial dependence on buying, maintenance and no of doors

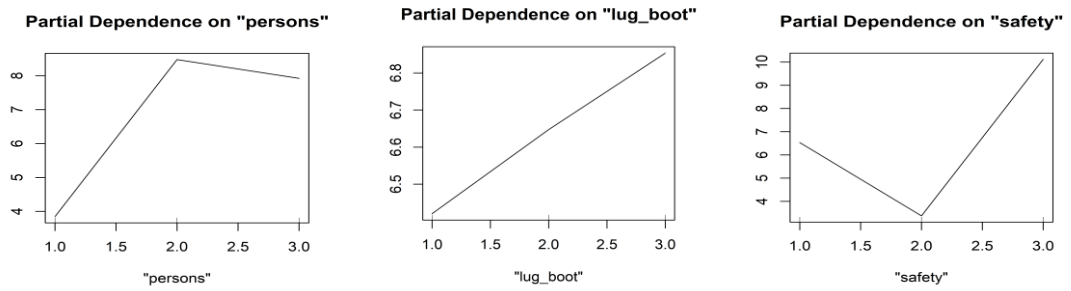

Figure 11: Partial dependence on persons, boot space and safety

The figures 10 and 11 shows the partial dependence of all the individual variables, which shows the dependence factor on class variable, from the graphs the gini index is decided.

Result of Research Question 2:
The 2nd question is to find a relation or pattern between safety and the buying price of the car. The below graphs are my findings for it.

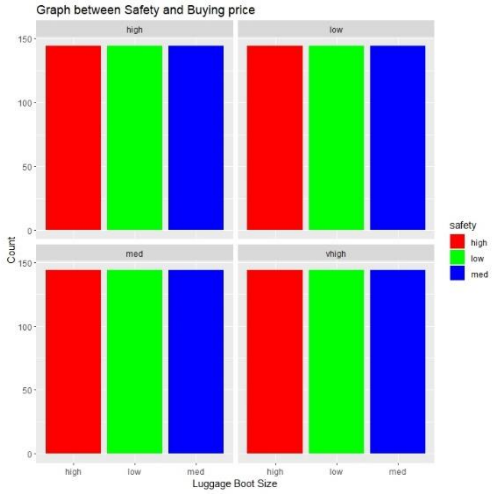

Figure 12: Relation between safety and buying price.

From figure 12, I see that all the types of variables in buying price have the same number of cars with all kinds of safety features. As this is derived from hierarchical decision model, all types of buying prices have all kinds of safety levels. Which means, some cars with very high buying have low safety as well as high safety too. So, comparing both the buying price and safety gives the equal size of graph. But when both buying price and safety are compared with class variable, the results will change. From figure 1 and 4, cars with very good acceptance rate have high safety and low buying price.

Result of Research question 3:

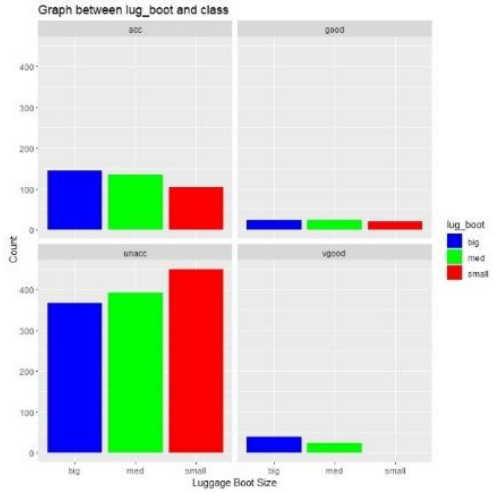

Figure 13: Relation between class and luggage boot space

From figure 12, I see that cars with very good acceptance (buying preference) have only big and medium luggage boot space. Cars with good acceptance (buying preference) have all three sizes of luggage boot space but they are less in number. Many cars with small luggage boot space are under unacceptance category. From this analysis I can say that boot space has an impact on buying preference. Cars with less boot space are not preferred to buy and cars with big are more preferred to buy as compared to medium size boot space.

VI. Conclusion, future analysis and challenges

From the graphs I plotted, I can get a rough idea on how the acceptance changes over all variables. And I found the most influential variables which are safety and no of persons. The decision tree and randomForest are used to predict the influential variable. I can use the model to determine whether a new car with a given feature will sell in good numbers, or customers can use it to

determine whether a specific car is suitable to purchase or not. All the graphs are plotted using r studio [8].

The analysis can be further developed with real time data. The data is old, and it is derived from a hierarchical decision model. So, new dataset with real time responses will give more detailed analysis. The same analysis can be performed using other machine learning techniques to predict the most influential variable.

There are a few challenges I faced during the analysis of the data. As the dataset is categorical and is evenly distributed, it is very hard to find patterns between them.

VII. References

[1] R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[2] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[3] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

[4] Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

[5] Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A Grammar of Data Man ipulation_. R package version 1.1.3, https://CRAN.R-project.org/package=dplyr

[6] Masum, M., PhD. (2023, November 22). Car evaluation analysis using Decision Tree Classifier. *Medium*. https://towardsdatascience.com/car-evaluation-analysis-using-decision-tree-classifier-61a8ff12bf6f

[7] Bohanec,Marko. (1997). Car Evaluation. UCI Machine Learning Repository. https://doi.org/10.24432/C5JP48.

[8] R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[9] Therneau T, Atkinson B (2022). _rpart: Recursive Partitioning and Regression Trees_. R package version 4.1.19, https://CRAN.R-project.org/package=rpart

[10] Milborrow S (2022). _rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'_. R package version 3.1.1, https://CRAN.R-project.org/package=rpart.plot

Appendix

Code For Decision Tree:

```
# Load necessary libraries
library(rpart)
library(rpart.plot)
library(caret)

# Load your dataset
car_data <- read.csv("car_evaluation.csv")
car_data$class <- factor(car_data$class, levels
= c("acc", "unacc", "good", "vgood"))

# Split the data into training and testing sets
set.seed(123)  # For reproducibility
train_indices <- createDataPartition(car_data$
class, p = 0.8, list = FALSE)
train_data <- car_data[train_indices, ]
test_data <- car_data[-train_indices, ]

# Train the decision tree model on the training
data
tree_model <- rpart(class ~ buying + maint + d
oors + persons + lug_boot + safety, data = trai
n_data, method = "class")

# Make predictions on the test data
predictions <- predict(tree_model, newdata = t
est_data, type = "class")

# Get unique levels from both predictions and
actuals
all_levels <- union(levels(predictions), levels(t
est_data$class))

# Convert predictions and actuals to factors wi
th the same levels
predictions <- factor(predictions, levels = all_l
evels)
actuals <- factor(test_data$class, levels = all_l
evels)

# Evaluate the model
conf_matrix <- confusionMatrix(predictions, a
ctuals)
accuracy <- conf_matrix$overall["Accuracy"]

# Print the evaluation metrics
cat("Accuracy: ", round(accuracy, 4), "\n")

# Visualize the confusion matrix
print(conf_matrix)

# Print the decision tree
print(tree_model)


# Plot the decision tree
#+ fig.width=12, fig.height=8, fig.dpi=1500
prp(tree_model, type = 2, extra = 1, branch = 0
.5, main = "Decision Tree for Car Evaluation")

#Plot the decision tree with improved colors
library(rpart.plot)
#+ fig.width=12, fig.height=8, fig.dpi=1500
rpart.plot(tree_model, main = "Decision Tree f
or Car Evaluation", tweak=1.1)

# Summary statistics for buying price and mai
ntenance by class
summary_stats <- aggregate(cbind(buying, ma
int) ~ class, data = car_data, FUN = function(x
) c(mean = mean(x), median = median(x), sd =
sd(x)))
print(summary_stats)

# Remove rows with missing values
car_data_cleaned <- na.omit(car_data)

# Convert 'buying' and 'maint' to numeric (assu
ming they are categorical variables)
car_data_cleaned$buying <- as.numeric(as.fact
or(car_data_cleaned$buying))
car_data_cleaned$maint <- as.numeric(as.facto
r(car_data_cleaned$maint))
# Convert 'class' to numeric for plotting
car_data_cleaned$class_numeric <- as.numeri
c(as.factor(car_data_cleaned$class))

# Scatter plot for buying vs. maintenance
plot(car_data_cleaned$buying, car_data_clean
ed$maint, col = car_data_cleaned$class_nume
ric, pch = 16, main = "Scatter Plot of Buying v
s. Maintenance", xlab = "Buying Price", ylab =
"Maintenance")


citation("rpart")
citation("rpart.plot")
```

Code For RandomForest:

```
# Install and load required packages
library(randomForest)
library(caret)
```

```
# Load your dataset
car_data <- read.csv("car_evaluation.csv")
summary(car_data)
# Remove rows with missing values in 'buying' or 'maint'
car_data <- na.omit(car_data)

# Convert 'class' to a factor
car_data$class <- as.factor(car_data$class)

# Split the data into training and testing sets
set.seed(123)  # For reproducibility
train_indices <- createDataPartition(car_data$class, p = 0.8, list = FALSE)
train_data <- car_data[train_indices, ]
test_data <- car_data[-train_indices, ]

# Train the random forest model
rf_model <- randomForest(class ~ buying + maint + doors + persons + lug_boot + safety, data = train_data, ntree = 500)
#+ fig.width=12, fig.height=6, fig.dpi=900
print(rf_model)
# Make predictions on the test data
rf_predictions <- predict(rf_model, newdata = test_data)

# Evaluate the random forest model
rf_conf_matrix <- confusionMatrix(rf_predictions, test_data$class)
rf_accuracy <- rf_conf_matrix$overall["Accuracy"]


# Print the evaluation metrics for the random forest model
cat("Random Forest Model Metrics:\n")
cat("Accuracy: ", round(rf_accuracy, 4), "\n")

# Visualize the confusion matrix for the random forest model
print(rf_conf_matrix)

# Plot variable importance
varImpPlot(rf_model)
print(rf_model$importance)

influential_predictor <- rf_model$variable.importance[which.min(rf_model$cptable[, "xerror"]), ]
cat("Most influential predictor:", influential_predictor, "\n")
```

Code For Research Question 2 & 3:

```
train_data$buying <- factor(train_data$buying, ordered = TRUE)
summary(train_data$buying)
#+ fig.width=4, fig.height=4, fig.dpi=900
partialPlot(rf_model, train_data, "buying")

train_data$maint <- factor(train_data$maint, ordered = TRUE)
summary(train_data$maint)
#+ fig.width=4, fig.height=4, fig.dpi=900
partialPlot(rf_model, train_data, "maint")

train_data$doors <- factor(train_data$doors, ordered = TRUE)
summary(train_data$doors)
#+ fig.width=4, fig.height=4, fig.dpi=900
partialPlot(rf_model, train_data, "doors")

train_data$persons <- factor(train_data$persons, ordered = TRUE)
summary(train_data$persons)
#+ fig.width=4, fig.height=4, fig.dpi=900
partialPlot(rf_model, train_data, "persons")

train_data$lug_boot <- factor(train_data$lug_boot, ordered = TRUE)
summary(train_data$lug_boot)
#+ fig.width=4, fig.height=4, fig.dpi=900
partialPlot(rf_model, train_data, "lug_boot")

train_data$safety <- factor(train_data$safety, ordered = TRUE)
summary(train_data$safety)
#+ fig.width=4, fig.height=4, fig.dpi=900
partialPlot(rf_model, train_data, "safety")

train_data$class <- factor(train_data$class, ordered = TRUE)
summary(train_data$class)
#+ fig.width=4, fig.height=4, fig.dpi=900
partialPlot(rf_model, train_data, "class")

citation()
citation("randomForest")
citation("caret")
```

Code For Research Question 2 & 3:

```
path_to_csv <- "car_evaluation.csv"

# Read the CSV file into a data frame
df <- read.csv(path_to_csv)

# Load the required libraries
library(ggplot2)
library(dplyr)
```

```
# Calculate the overall probability of 'very goo
d' class
probability_very_good <- mean(df$class == "v
good")

# Print the probability
cat("Probability of being classified as 'very go
od':", probability_very_good, "\n")

# Create a data frame for counts
count_df <- df %>%
  group_by(lug_boot, buying) %>%
  summarise(Freq = n())

# Calculate proportions
count_df <- count_df %>%
  group_by(lug_boot) %>%
  mutate(Proportion = Freq / sum(Freq))

# Print the resulting data frame
print(count_df)

df$lug_boot <- as.factor(df$lug_boot)
df$class <- as.factor(df$class)

# Plot the graph
ggplot(df, aes(x = lug_boot, fill = lug_boot)) +
  geom_bar() +
  facet_wrap(~ class) +
  scale_fill_manual(values = c("small" = "red",
"med" = "green", "big" = "blue")) +
  labs(title = "Graph between lug_boot and clas
s",
       x = "Luggage Boot Size",
       y = "Count")

ggplot(df, aes(x = safety, fill = safety)) +
  geom_bar() +
  facet_wrap(~ buying) +
  scale_fill_manual(values = c("high" = "red",
"low" = "green", "med" = "blue")) +
  labs(title = "Graph between Safety and Buyin
g price",
       x = "Luggage Boot Size",
       y = "Count")
```

Code For Exploratory Analysis:

```
# Read the CSV file into a data frame
df <- read.csv("car_evaluation.csv")

# Load the required libraries
library(ggplot2)
df$class <- factor(df$class, levels = c("acc", "u
nacc", "good", "vgood"))

# Plot for Safety and Class with the specified o
rder
ggplot(df, aes(x = safety, fill = class)) +
  geom_bar(position = "fill") +
  labs(x = "safety", y = "class" ) +
  ggtitle("Relationship between safety and clas
s") +
  theme_minimal()

ggplot(df,aes(x = persons, fill = class)) +
  geom_bar(position="fill")+
  labs(x = "No of persons", y = "Class" ) +
  ggtitle("Relationship between people capacit
y and class")+
  theme_minimal()

ggplot(df,aes(x = buying, fill = class)) +
  geom_bar(position="fill")+
  labs(x = "Buying price", y = "class" ) +
  ggtitle("Relationship between buying price an
d class")+
  theme_minimal()

ggplot(df,aes(x = maint, fill = class)) +
  geom_bar(position="fill")+
  labs(x = "Maintenance", y = "Class" ) +
  ggtitle("Relationship between Maintenance p
rice and class")+
  theme_minimal()

ggplot(df,aes(x = doors, fill = class)) +
  geom_bar(position="fill")+
  labs(x = "No of doors", y = "class" ) +
  ggtitle("Relationship between No of doors an
d class")+
  theme_minimal()

ggplot(df,aes(x = lug_boot, fill = class)) +
  geom_bar(position="fill")+
  labs(x = "Luggage boot space", y = "Class" )
+
  ggtitle("Relationship between Luggage boot s
pace and class")+
  theme_minimal()

citation()
```