CAPSTONE PROJECT

MACHINE LEARNING FOUNDATIONS

By Dr. Sumudu Tennakoon

30/07/2022

—

STUDENT: NAMINDA JAYAWARDANA

Registration number: DSA_0299

1. Problem Statement

1.1. Used car industry in most of the counties is very popular. Selection for purchasing depends on the buyer's choice and there are common criteria for selection by most of the people. But there are some country specific criteria as well on final selection.

1.2. Under this exercise it is planning to automate the buyer's decision to identify the likelihood of buying based on identified criteria as input /feature variables. Under this project it is planned to use machine learning algorithm to automate the decision

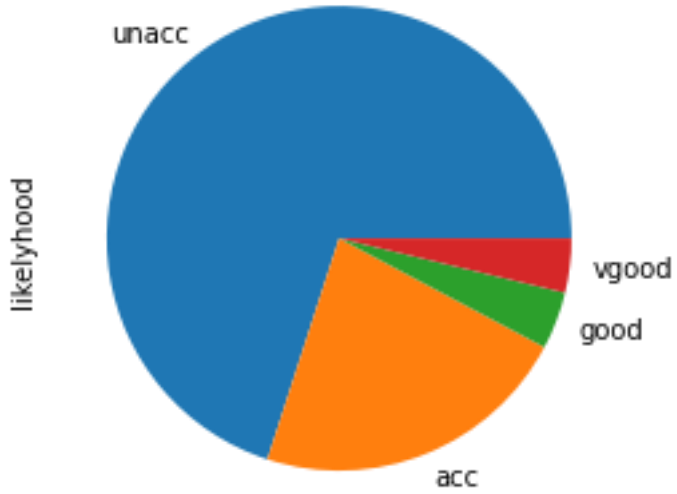
2. Data

2.1. Data source

Sample data for this exercise was collected from UCI data repository which was believed have common input variables for the analysis.
'https://archive.ics.uci.edu/ml/machine-learning-databases/car/car.data'

2.2. Exploratory Data analysis

Data set was not available with proper columns names, and it was renamed with meaningful column names.
1727 records were found in the data set and each record were with six variables namely price, maint, doors, persons, boot size, safety, and likelihood.
Each variable was with 2 ,3 or 4 unique values and data set was very clean, and no blanks were found. No Null values were found and no duplicates were found in the data set.
Following pie chart shows distribution of likelihood of buying.

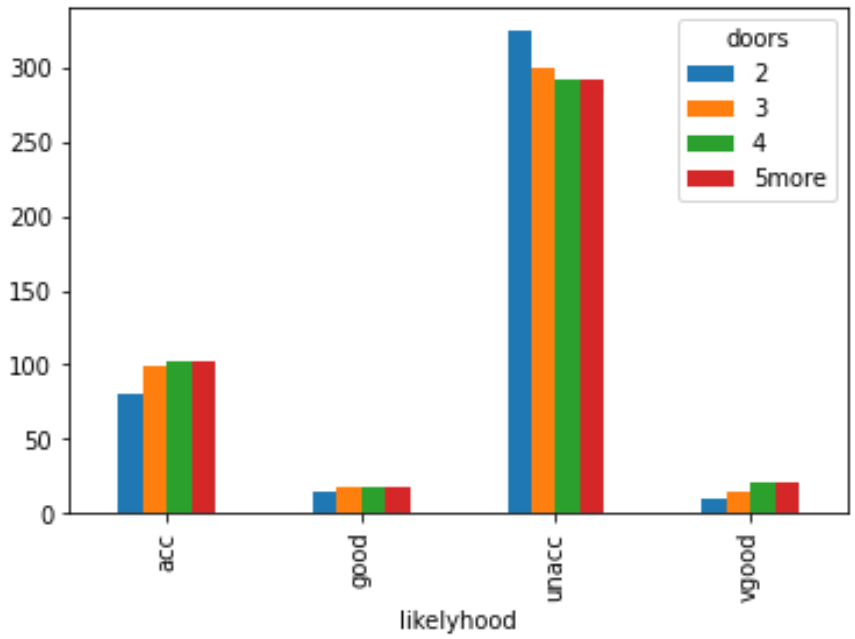


2.3. Data types

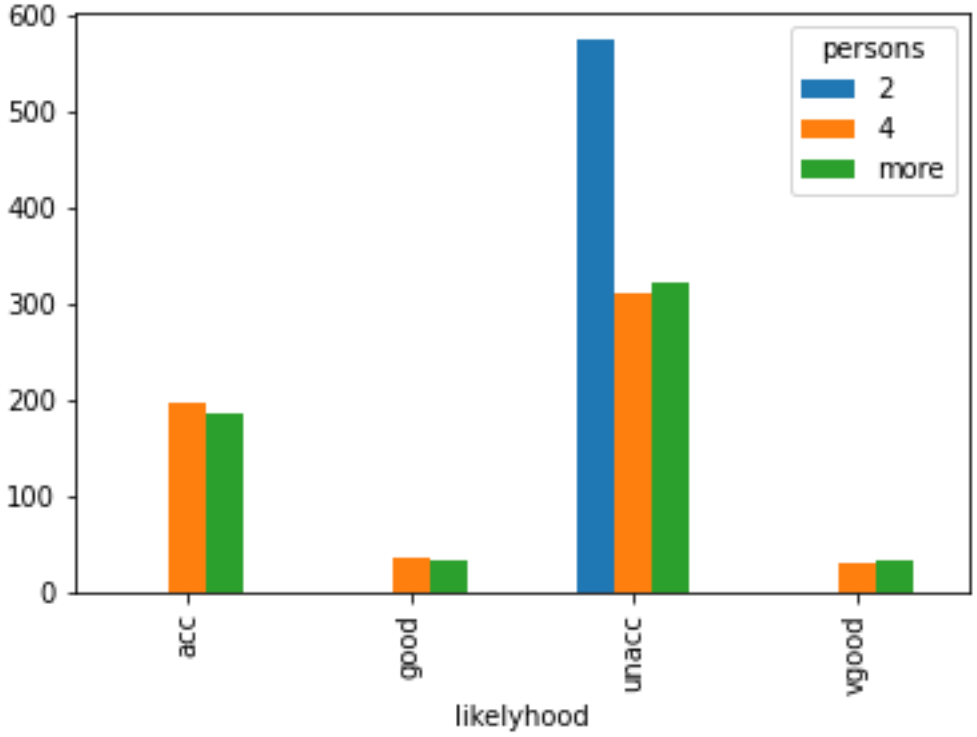
All are categorical variables as identified by each category of six variable. Each six variable was with 2,3 or 4 unique values.

2.4. feature analysis

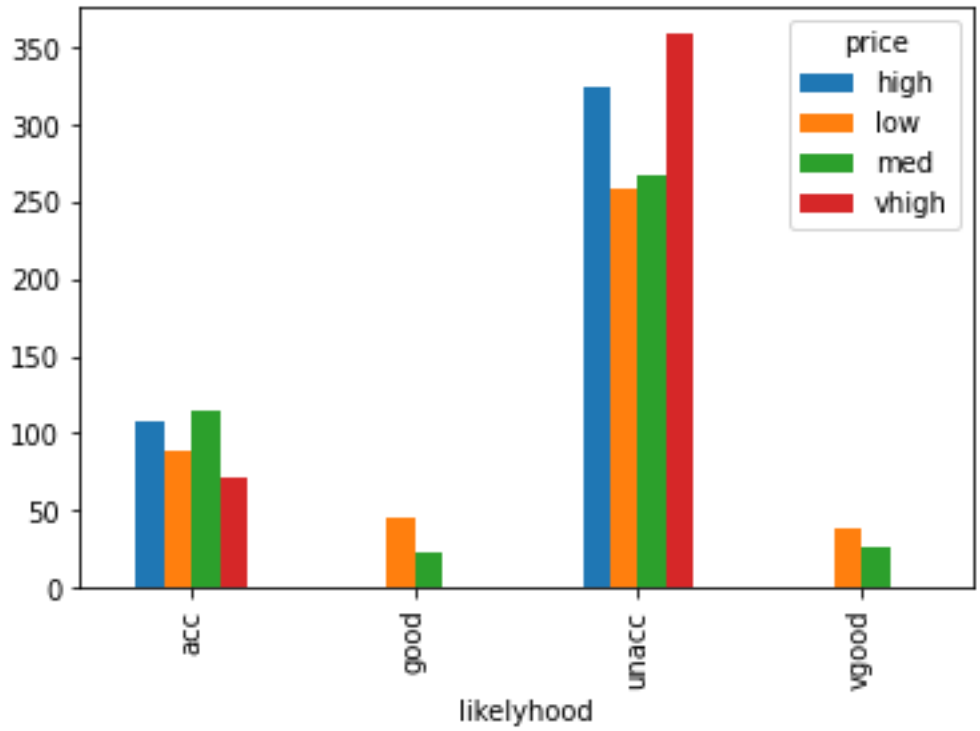
Number of doors was not a big dependency on buying decision and it was clear from the bar chart below. However, it was considered for model building.



2.4.1. It was observed that there is no likelihood to buy cars with 2 seats and it is clear from the following bar chart



2.4.2. It was observed that likelihood for buying is very less for cars with high price irrespective of the other parameters



3. Methodology

3.1. Feature identification: All features in the data set are categorical variables

3.2. Model selection

As it is categorical variables and results are also based on categories it was identified to use classification algorithms. Logistic regression, decision tree and random forest classifier was selected to model the data set

3.3. Define and X and y variables: Five variables were selected as feature variables (X variables) and one-hot encoding were used to encode them

3.4. Likelihood was selected as y variable and Probability of Output variables were also one-hot encoded to consider for model.

3.5. 70% : 30% Train: Test were used to split the data in model in training and testing

3.6. Performance of each model was evaluated for accuracy

4. Results

4.1. Model evaluation

Models were evaluate using Accuracy and F1-score matrices separately.

4.2. Evaluation comparison

Evaluated matrices were tabulated as follows for the comparison purposes. As per the data it is very clear that Model with random forest classifier shows the highest accuracy in predicting decision whether there is a likelihood for buying

	Model	accuracy	F1_score
0	Logistic Regression	0.897881	0.894841
1	Decision Tree Classifier	0.949904	0.951198
2	Random forest Classifier	0.953757	0.951797

5. Conclusion

1.1. Random forest classifier provides the best accuracy
1.2. Model object can be used in any API development to automate the buying decision and can be integrated to a web interface.

6. Discussion

1.1. In country specific situation there we may need some other parameters like service records history of car, number of owners who used the car, vehicle registration year and valid certificate etc
1.2. With the above filed also it may need to customize the model depend on the country