DATA SCIENCE REPORT

MULTIVAIATE CAR EVALUATION CLASSIFICATION

ON DATA PROVIDED BY THE UCI MACHINE LEARNING REPOSITORY

JOEL DOMINGO joeldomingo117@gmail.com

ABSTRACT

Using the dataset provided, we built a **classification model** using a **support vector machine** for producing an accurate evaluation of a car based on categorical features. Our successful model produced both an **accuracy score** and f1 score of **97%,** as well as precision and recall scores of the same value. Furthermore, we were able to adapt the model to have the ability to **classify pricing models** on listed car sales to assist in decision-making processes when purchasing the car. Using the principles based on this model, we gathered that this process can be comparable to the usefulness in decision-making processes when purchasing or allocating resources in a business context.

INTRODUCTION

Evaluating the class of a particular product is an important aspect of risk assessment. This is true for private buyers as well as business organizations, both of whom must be able to estimate the risk involved prior to allocating resources to purchase an item or applying for loans to do so. Techniques of data science and predictive analysis can be used to predict the true value of such products to make appropriate decisions. In this project, we used data which provides categorical and numerical features on cars and provides an evaluation class of that instance. Using this data, we will be creating a model which can be adapted for use on evaluation of an appropriate pricing model based on the categorical evaluation of the car.

Objective

The goal of this project is to use techniques of data science to estimate the appropriate valuation of car listings.

The Data

The car evaluation database is provided by the UCI Machine Learning repository and was derived from a simple hierarchical decision model originally developed for the demonstration of DEX, M. Bohanec, V. Rajkovic: Expert system for decision making. The model evaluates cars according to the following concept structure.

Feature	Description	Classes
Pricing		
Buying	Purchasing cost	Very high, high, med, low
Maintenance	Maintenance costs	Very high, high, med, low
Technical Characteristics		
Doors	Number of doors	2, 3, 4, 5 or more
Persons	Person capacity	2, 4, more
Lug boot space	Size of luggage boot	Small, med, big
Safety	Estimated safety of car	Low, med, high

Target Variable

Our target variable is the evaluation of an instance of a car, which is categorized into the 4 classes displayed below.

Abbreviation	Class
Unnacc	Unacceptable
Acc	Acceptable
Good	Good
vgood	Very Good

Data Statistics

- The data contains **1,728 records**, each with **6 features**.
- The data was developed in **1997**.
- Contains **multivariate** characteristics.

DISTRIBUTION AND BASELINE ACCURACY

Figure 1: Distribution of classes

Class	Instances	Normalized Percentage
Unacceptable	1210	0.70
Acceptable	384	0.22
Good	69	0.04
Very Good	65	0.04

Observations

Looking at the distribution of our classes, we see that majority of cars are classed as **unacceptable**, with the second highest class being **acceptable.** Overall, the data is quite unbalanced, and therefore our resulting model may have poor performance when forced to generalize on unseen observations.

However, looking at the normalized percentage shows us our baseline or target accuracy to beat. If the model were to evaluate all cars in the dataset to unacceptable, we will be predicting with an accuracy **of 70%.** Therefore, we want to be able to improve on this baseline accuracy. Our aim is to achieve a model accuracy and f1 score of **> 90%.**

FEATURE CORRELATION

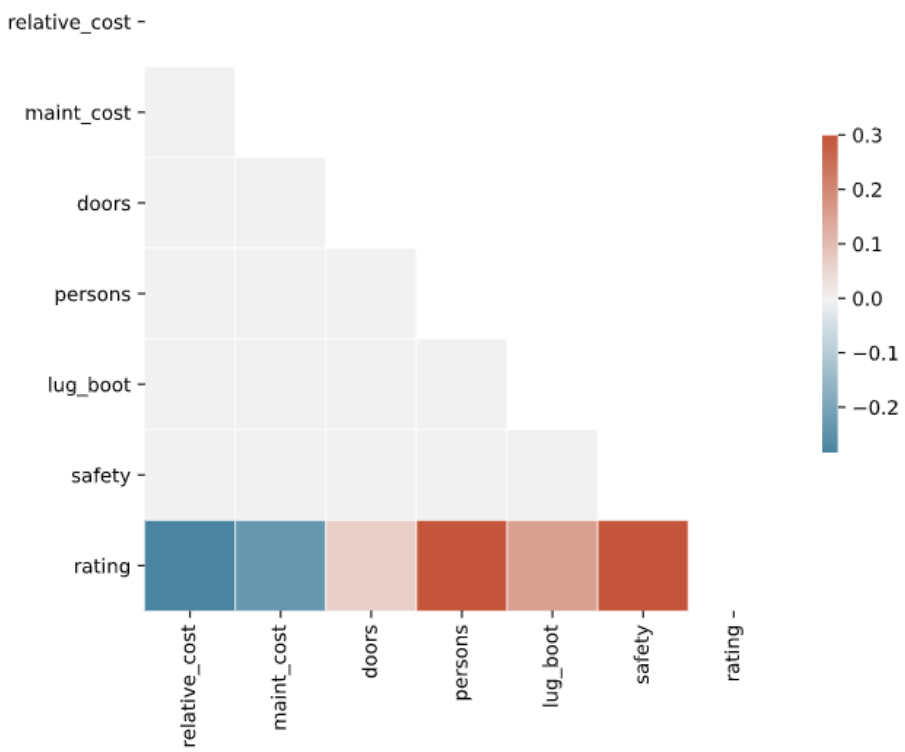


Figure 2: Correlation Heatmap of Features

Observations

- Rating is inversely correlated with relative cost. This means that higher cost means lower rating
- Same can be said for miniatous costs (higher cost means lower rating)
- More doors are slightly correlated with rating, but not as important
- Higher person capacity means a higher rating, which makes sense because you get more value out of the car
- Bigger boot space indicates a higher value
- Safety is always an important feature, so higher safety rating indicates better class rating for the car

MODELLING

To choose our model, we will be evaluating several different models based on the following features.

Evaluation Feature	Description
Accuracy	Number of correct predictions
Precision	Ratio of true positives to total positives
Recall (sensitivity)	Ratio of true positives to all observations in that class
F1-Score	Weighted Average of precision and recall. This takes both false positives and false negatives into account.

Triage of Models

The models which we will test are shown below. Out of these models, we will choose the best overall model based on our evaluation criteria.

Model	Description
Logistic Regression	Uses logistic function to model binary dependent variables.
Support Vector Machine	Linear algebraic method for separating n-dimensional data into classes.
K-Nearest Neighbors	Based on feature similarity or how closely out-of-sample features resemble the training set to determine how to classify a given data point.
Naïve Bayes	Based on Bayesian probability in which probability is a degree of belief. Uses prior distribution in addition to current samples to estimate future probability of a positive result.

Model Training Method

For our models, we will be splitting into training and test sets using a train-test-split of 20% and a random state of 17.

MODEL EVALUATION

Logistic Regression (multi-class = 'multinomial', solver = 'lbfgs', penalty = 'l2')

Number of mislabeled points out of a total 346 points: **56**

Model	Accuracy Score	Precision	Recall	F-Score
Logistic Regression	0.83	0.83	0.84	0.83

	precision	recall	f1-score	support
acc	0.69	0.61	0.64	76
good	0.75	0.33	0.46	18
unacc	0.88	0.96	0.92	243
vgood	1.00	0.44	0.62	9

Figure 3: Classification Report Logistic Regression

Observations

- Logistic regression works best when predicting **binary** categorical outcomes. Our data is multivariate.
- 0.83 is a considerable increase on our base score.
- Precision for the more median populated classes is low.
- Recall scores are all individually low, except for the highest populated class.

MODEL EVALUATION

Support Vector Classifier (C = 1, gamma = 1, probability = True)

Number of mislabeled points out of a total 346 points: **11**

Model	Accuracy Score	Precision	Recall	F-Score
SVC	0.97	0.97	0.97	0.97

	precision	recall	f1-score	support
acc	0.94	0.96	0.95	76
good	0.92	0.67	0.77	18
unacc	0.99	0.99	0.99	243
vgood	0.75	1.00	0.86	9
accuracy			0.97	346
macro avg	0.90	0.90	0.89	346
weighted avg	0.97	0.97	0.97	346

Figure 4: Classification Report SVC

Observations

- High accuracy
- High recall and precision scores
- Per class they are all relatively high
- Promising results

MODEL EVALUATION

Naïve Bayes

Number of mislabeled points out of a total 346 points: **106**

Model	Accuracy Score	Precision	Recall	F-Score
Naïve Bayes	0.7	0.75	0.69	0.67

	precision	recall	f1-score	support
acc	0.59	0.13	0.22	76
good	0.50	0.11	0.18	18
unacc	0.84	0.90	0.87	243
vgood	0.14	1.00	0.24	9
accuracy			0.69	346
macro avg	0.52	0.54	0.38	346
weighted avg	0.75	0.69	0.67	346

Figure 5: Classification Report Naive Bayes

Observations

- Relatively low accuracy, precision, and recall scores.
- Expected result. Naïve Bayes doesn't work well with highly correlated features (price and maintenance costs) and becomes heavily biased.
- Naïve Bayes is better for simple classification solutions.

MODEL EVALUATION

K-Nearest Neighbors (n_neighbors = 5)

For K-Nearest Neighbors, we need to decipher what our optimal number of K will be. Running a triage loop, we see that our number of K optimizes at around 5-7 neighbors.

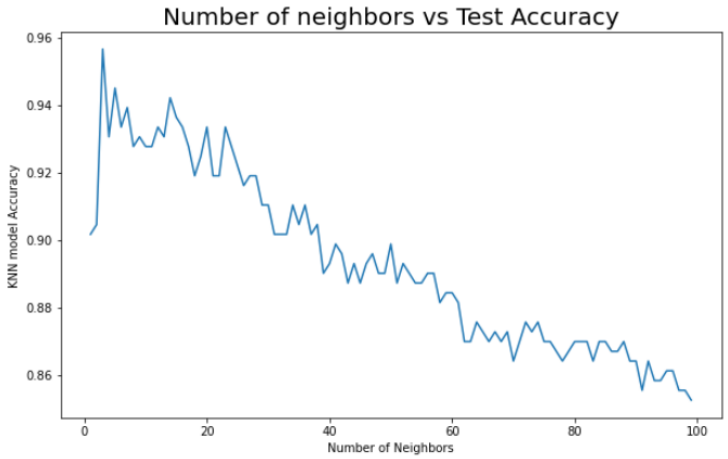


Figure 6: Number of K Optimization Chart

Number of mislabeled points out of a total 346 points: **19**

Model	Accuracy Score	Precision	Recall	F-Score
KNN	0.945	0.95	0.95	0.95

	precision	recall	f1-score	support
acc	0.85	0.91	0.88	76
good	0.93	0.78	0.85	18
unacc	0.98	0.97	0.98	243
vgood	0.80	0.89	0.84	9
accuracy			0.95	346
macro avg	0.89	0.89	0.89	346
weighted avg	0.95	0.95	0.95	346

Figure 7: Classification Report KNN

Observations

- High accuracy, precision, and recall scores with n_neighbors = 5.
- Stable across all classes.

MODEL EVALUATION - COMPARISON

Model Scores

Model	Accuracy Score	Precision	Recall	F-Score
Logistic Regression	0.83	0.83	0.84	0.83
SVC	0.97	0.97	0.97	0.97
Naïve Bayes	0.7	0.75	0.69	0.67
KNN	0.95	0.95	0.95	0.95

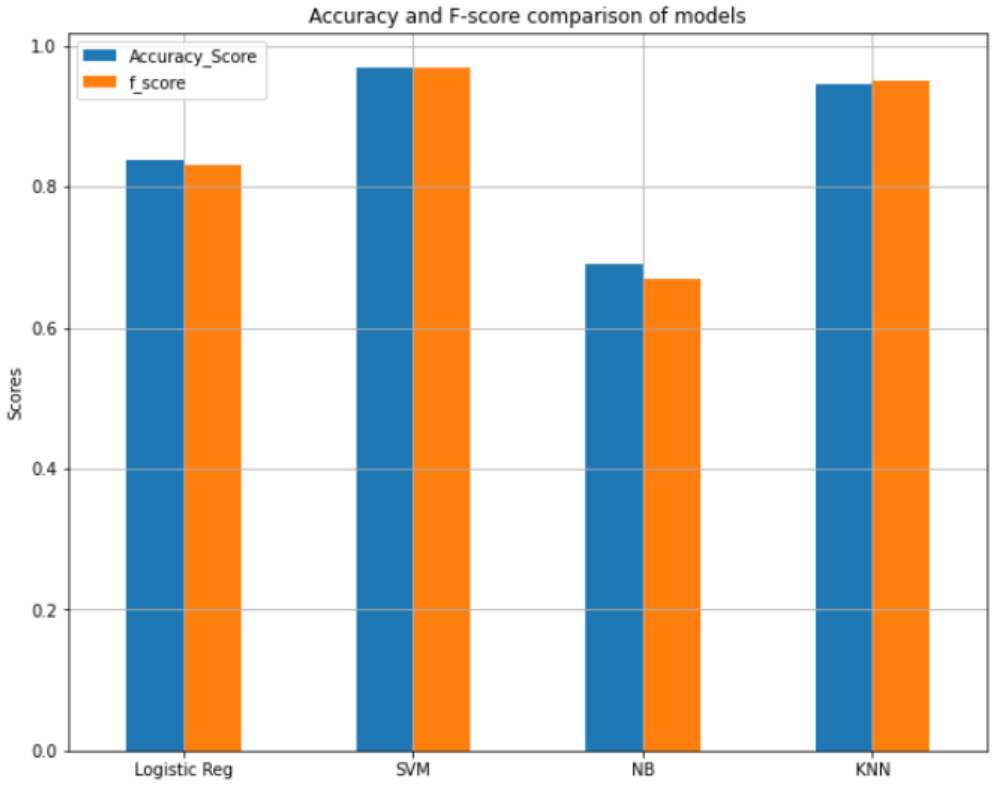


Figure 8: Comparison of Model Scores

Observations

- Based off our weighted scores, Support Vector Classifier and K-Nearest Neighbors are our best classifiers.
- We will be using SVC for our use case.

USE CASE – MAKING A CLASSIFICATION

Objective

For our use case, we will be making adapting our model to be used on evaluating the pricing model on an instance of a used car sales listing.

The Model

The model being used is the Support Vector Classifier, which achieved a weighted average across-the-board score of 0.97 on our training data.

The Data

- For our classification, we will be gathering data from the Australian used car listings hub www.carsales.com.au.
- We will be classifying listings of Subaru WRX models with model years later than 2014.

Criteria

Model features:

Feature	Description	Classes
Pricing		
Buying	Purchasing cost	Very high, high, med, low
Maintenance	Maintenance costs	Very high, high, med, low
Technical Characteristics		
Doors	Number of doors	2, 3, 4, 5 or more
Persons	Person capacity	2, 4, more
Lug boot space	Size of luggage boot	Small, med, big
Safety	Estimated safety of car	Low, med, high

To adapt our rating system to a price rating system, we will classify prices as low or high depending on the average price of the Subaru WRX.

Mean Price ($AUD)	Standard Deviation
38,000	10,000

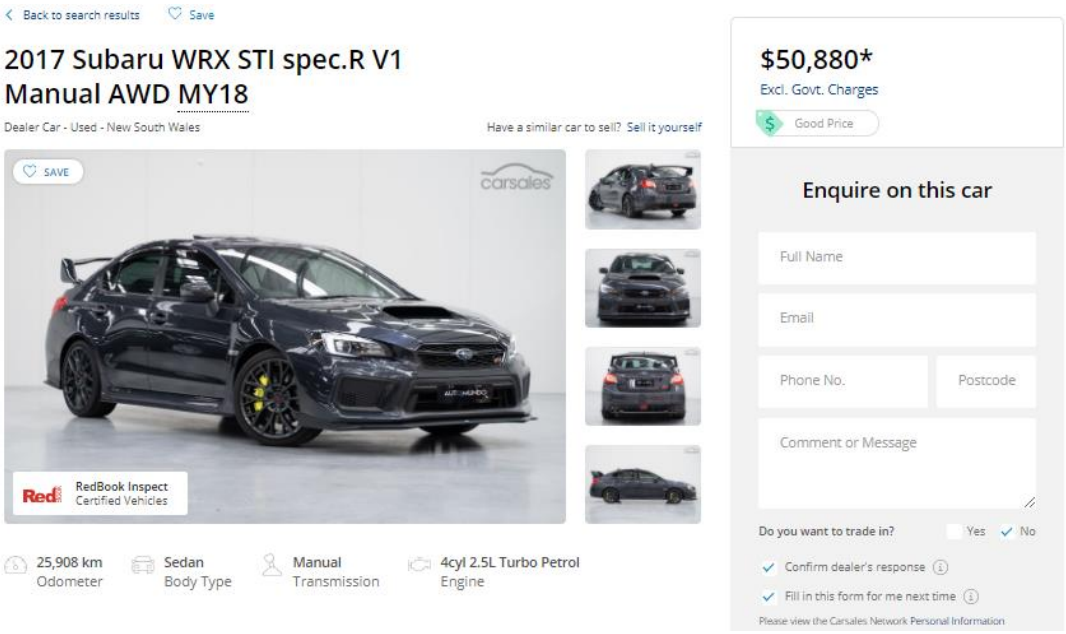
We base our input of the **relative_price** feature based on the following classifications:

Class	Standard Deviations	Price ($ AUD)
Low	<1 std	< 28,000
Med	Within 1 std	28,000 – 48,000
High	Between 1 and 1.4 std	48,000 – 52,000
Very high	>1.2 std	> 52,000

USE CASE – MAKING A CLASSIFICATION

The Listing

Our sale listing will be a 2018 Subaru WRX STI.



The listing has the following features:

Feature	Feature Value	Determined Class	Potential Values
Pricing			
Buying	$50,000	High	Very high, high, med, low
Maintenance	$1,500 - $2,000 p. a	Medium	Very high, high, med, low
Technical Characteristics			
Doors	4 doors	4	2, 3, 4, 5 or more
Persons	5-person capacity	More	2, 4, more
Lug boot space	Medium sized Boot	Med	Small, med, big
Safety	5-star ANCAP rating (maximum)	High	Low, med, high

Prediction:
Our SVC model has determined that this listing as **‘acceptable’**, considering the car’s features and relative price.

CONCLUSION

Based on the input data, we were able to develop several classification models which take in the features of a car listing and evaluate the acceptability of purchasing price considering its categorical variables. Out of these models, the support vector classifier provided with the best overall result based on a train-test-split of 20%. However, it is imperative to note that the dataset itself is outdated, submitted to the UCI machine learning repository in 1997. This does not mean the data is completely obsolete. The model which was used to classify the output variable within the data still remains relevant, but with quickly advancing car technologies, and numerous features being added every half-decade, it stands to reason that many more features can be added to make a more comprehensive dataset and classification model. With these changes, the model can be adapted to a number of use cases to assist decision-making processes regarding purchasing or allocation of business or personal resources.

Recommendations

- Use more current data.
- Develop a higher-dimensional dataset.
- Convert classification variables (such as low, medium, high) to standard or numerical class/range systems, such as the local ANCAP safety rating (Australia).

Sources

- **Dataset Source:** https://www.carsales.com.au/cars/details/2017-subaru-wrx-sti-spec-r-v1-manual-awd-my18/OAG-AD-19471214/?Cr=1
- **GitHub link to code:** https://github.com/jdomingo117/Projects/tree/main/Multivariate%20Car%20Evaluation%20Classification/Code
- https://archive.ics.uci.edu/ml/datasets/Car+Evaluation
- https://en.wikipedia.org/wiki/Precision_and_recall#F-measure
- https://holypython.com/nbc/naive-bayes-pros-cons/
- https://scikit-learn.org/
- https://nthu-datalab.github.io/ml/labs/06_Logistic-Regression_Metrics/06_Logistic-Regression_Metrics.html
- https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/