Case Study 1: Classical Supervised Machine Learning
Predicting Kicks in Auto Auction
Due date: 11:59 PM, 5th September 2025
Weighting: 20%

Introduction

This assignment will allow you to display your knowledge and understanding of classical supervised machine learning. In this assignment, you will use classification algorithms implemented in Python to display your technical competence gained from the theoretical and applied practices. You should primarily use the codes and libraries introduced in the subject. If you use a major library outside the subject scope, explain why it was needed.

Instructions

1. This unit treats Academic integrity and plagiarism with the utmost seriousness. You are **NOT allowed to collaborate with other groups or seek external assistance for assessment tasks such as those generated by an AI, obtained from another person, or sourced from the internet**. In all cases, the submission is not an original piece of work created by the student. Any violation of this policy may result in severe consequences. Read the Assessment Policies on Canvas or the QUT Website for further details.

2. The dataset required for this assignment is available on Canvas with **kick.csv**.

3. The assignment will be **evaluated based on two components**. The first component, **worth 85%, will be assessed based on the report** submitted by the group on Canvas. The second component, worth **15%**, will be assessed individually based on an **online quiz in Week 8**. Note that no extensions will be allowed on the quiz part.

4. The deadline for submitting the **assignment report is Friday, 5 September 2025**, at 11:59 p.m. Each group is required to submit only one final report.

   The report must include responses to all questions specified in the tasks and should be adequately formatted to ensure easy navigation through the answers. You do not need to include an introduction, summary, conclusion, or references in the report. Some answers may require screenshots, which should be included in a table. While it may be tempting to provide excessive detail, it is crucial to focus on essential points and attach relevant screenshots to demonstrate thorough consideration of the matter. The report is anticipated to be around 20 pages long.

   The Jupyter notebook outlining all the code must also be submitted. However, the primary document for marking will be the group report, and the code will only be used in case of uncertainty. Therefore, the report must be self-explanatory and not rely on the Jupyter Notebook attachment.

   Name the final report **assessment1-report.doc** or **assessment1-report.pdf**. The Word or pdf file should include the cover page (same as the template provided) with the Student ID number and full name (as in QUT-DW) for all students, along with the group number. Also, submit the **team agreement and Jupyter notebook**. Submit these three files on Canvas.

The submission process involves accessing the "Assignments" section on Canvas and submitting the report under the Assignment 1 section by selecting the "Assignment 1 Submission" link.

5. For the report submission, you must work in **a team of three members**. The team must be registered **on Canvas by Week 3**. You are to choose an available group designated for Assignment 1 in Canvas.

   You are to manage the group. As in real life, the performance of the individuals in the team is judged by the performance of the team together, so choose your partners carefully. Once you form the group, group movements are NOT allowed.

   To ensure that everyone agrees to their responsibilities in the team, we have asked the team to complete a **Team Agreement form**. You can find the team agreement template in the Assessment module under the Assignment 1 section. Once the team is formed, complete the team agreement form and register the team on Canvas. It should clearly be stated if there is an unequal distribution of marks between the team members. This should be agreed upon by all members, as shown by their signatures.

Case Study Dataset

A common challenge faced by a car dealership in purchasing a used car at an auto auction is determining the risk that the vehicle might have serious issues that will prevent it from being sold to customers. The auto community calls these unfortunate purchases "kicks".

Kicked cars often result when there are tampered odometers, mechanical issues that the dealer is unable to address, issues with obtaining the vehicle title from the seller, or some other unforeseen problem. Kick cars can be very costly to dealers, considering transportation costs, throw-away repair work, and market losses incurred from reselling the vehicle.

A dealership manager would like to determine which cars have a higher risk of being kick. This will provide real value to the dealership, trying to provide the best inventory selection possible to their customers. They have been collecting the data for many years and have also manually labelled the data.

The data set kick.csv contains over 40,000 observations and 31 variables. The variables in the data set are listed in Table 1.

Table 1: List of Variables

Name	**Description**
PurchaseID	Purchase Identification Number
PurchaseTimestamp	Purchase Timestamp
PurchaseDate	Purchase Date
Auction	Auction company
VehYear	Year Vehicle is made
Make	Vehicle’s make

Color	Color of the car
Transmission	Auto or manual Transmission
WheelTypeID	Wheel type Identification Number
WheelType	Wheel type
VehOdo	Vehicle’s Odometer reading
Nationality	Nationality of the vehicle
Size	Size of the vehicle
TopThreeAmericanName	If the vehicle is from one of the top three American manufacturers.
MMRAcquisitionAuctionAveragePrice	Acquisition price for this vehicle in average condition at the time of purchase
MMRAcquisitionAuctionCleanPrice	Acquisition price for this vehicle in the above Average condition at the time of purchase
MMRAcquisitionRetailAveragePrice	Acquisition price for this vehicle in the retail market in average condition at the time of purchase
MMRAcquisitonRetailCleanPrice	Acquisition price for this vehicle in the retail market in above average condition at the time of purchase
MMRCurrentAuctionAveragePrice	Acquisition price for this vehicle in average condition as of current day
MMRCurrentAuctionCleanPrice	Acquisition price for this vehicle in the above condition as of the current day
MMRCurrentRetailAveragePrice	Acquisition price for this vehicle on the retail market in average condition as of the current day
MMRCurrentRetailCleanPrice	Acquisition price for this vehicle on the retail market in above average condition as of current day
MMRCurrentRetailRatio	Ratio of MMRCurrentRetailAveragePrice and MMRCurrentRetailCleanPrice
PRIMEUNIT	Level of demand with respect to a standard purchase
AUCGUART	The risk that can be run with the vehicle, meaning how much guarantee the seller is willing to give
VNST	Geographic region
VehBCost	Acquisition cost paid for the vehicle at time of purchase
IsOnlineSale	1 = Sale done online, 0 = No
Warranty cost	Warranty price (term = 36month and millage = 36K)
ForSale	Whether is car is available for sale
IsBadBuy	1 = Yes, 0 = No

Case Study Tasks

Your task is to build various classification models, including decision trees, regression functions, and neural networks, on this data set to predict if the car purchased at the Auction is a Kick (bad buy) and compare them. Results inferred by these models should inform decision-makers of the (characteristics of) potential kick cars. This information can be utilised in multiple ways to assist various stakeholders. The specific tasks for each machine learning process are listed below.

Task 1. Data Preparation for Modelling. (3.5 marks)

1. The dataset may include irrelevant and redundant variables to the underlying ML task. What variables did you include in the modelling, and what were their roles and measurement level set? Justify your choice.
2. Did you have to fix any data quality problems, including data imputation? Detail them.
3. Report the proportion of values of the target variable for the dataset before and after the pre-processing.

Task 2. Decision Tree Modeling (4.5 marks)

1. Build a decision tree using the default setting. Examine the tree results and answer the following:
    a. What parameters have been used to build the tree? Detail them.
    b. What data split was used to create training and test datasets?
    c. What is the classification accuracy on the training and test datasets?
    d. What is the size of the tree (number of nodes and rules)?
    e. Which variable is used for the first split? What are the variables that are used for the second split?
    f. What are the five important variables in building the tree?
    g. Report if you see any evidence of model overfitting.
2. Build another decision tree tuned with GridSearchCV. Examine the tree results.
    a. What are the optimal parameters for this decision tree? Explain your choice of hyperparameters to search, and the chosen search range(s)?
    b. What is the classification accuracy on the training and test datasets?
    c. What is the size of the chosen tree (number of nodes and rules)?
    d. Which variable is used for the first split? What are the variables that are used for the second split?
    e. What are the five important variables in building the tree?
    f. Report if you see any evidence of model overfitting.
3. What is the significant difference between these two decision tree models – default (Task 2.1) and using GridSearchCV (Task 2.2)? How do they compare performance-wise? Produce the ROC curve for both DTs. Explain why those changes may have happened.
4. From the better model, can you provide the characteristics of cars most likely to be ‘kicks’? If it is hard to comprehend, discuss why.

Task 3. Regression Modeling (5 marks)

1. Describe what additional processing was required on this dataset to be used in regression modelling. List the variables that need further processing and provide details of the processing.
2. Build a regression model tuned with GridSearchCV. Answer the following:
    a. Name the Regression function used.
    b. What are the optimal parameters for this regression model? Explain your choice of hyperparameters to search, and the chosen search range(s)?
    c. Report the variables that are included in the regression model.
    d. Report the top-5 important variables (in order) in the model.
    e. What is the classification accuracy on training and test datasets?
    f. Report any sign of overfitting.
3. Build another regression model on the reduced variables set. To minimise variables, either perform dimensionality reduction with Recursive Feature Elimination or select a subset of inputs found significant by the decision tree (use the best decision tree model under Task 2). Tune the model with

GridSearchCV to find the best parameter setting. Answer the following:

   a. Was dimensionality reduction helpful in identifying a good feature set for building the accurate model? Report the feature selection method used.
   b. Report the variables that are included in the regression model.
   c. What is the classification accuracy on the training and test datasets?
   d. Report any sign of overfitting.

4. Produce the ROC curve for both regression models. Using the best regression model, can you provide the characteristics of cars most likely to be 'kicks'? If it is hard to comprehend, discuss why.

Task 4. Predictive Modeling Using Neural Networks (5 marks)

1. Describe what additional processing was required on this dataset to be used for neural network modelling.
2. Build a Neural Network model tuned with GridSearchCV. Answer the following:
   a. Explain the parameters in building this model, e.g., network architecture, iterations, activation function, etc. Explain your choice of hyperparameters to search, and the chosen search range(s)?
   b. What is the classification accuracy of the training and test datasets?
   c. Did the training process converge and result in the best model?
   d. Do you see any sign of over-fitting?
3. Build another Neural Network model with the reduced feature set. Perform dimensionality reduction by selecting variables with a decision tree (use the best decision tree model under Task 2). Tune the model with GridSearchCV to find the best parameter settings. Answer the following:
   a. Did feature selection favour the outcome? Report the changes in the network architecture. What inputs are being used as the network input?
   b. What is the classification accuracy on the training and test datasets?
   c. How many iterations are now needed to train this network?
   d. Do you see any sign of over-fitting? Did the training process converge and result in the best model?
4. Produce the ROC curve for both NNs. Using the best NN model, can you provide the characteristics of cars most likely to be 'kicks'? If it is hard to comprehend, discuss why.

Task 5. Final remarks: The decision making (2 marks)

1. Finally, based on all models and analysis, is there a model you will use in decision-making i.e. the best-performing model? Justify your choice. Draw an ROC chart and an Accuracy Table to support your findings.
2. Based on this analysis, can you summarise the positives and negatives of each predictive modelling method?

Marks Distribution (20 marks)

The assessment will be marked in two parts. The first part (85%), the group work, will be assessed via the final report you submit on Canvas. Note that in machine learning, there is hardly ever a single solution. The solution depends upon various settings, such as the role and measurements of input variables, training size, underlying algorithm, and the selected algorithm parameters. Your project partner may have a different solution than yours. Your group should decide on a project you would like to be marked. **Submit a single report per group** after discussing the final project components. If the marker finds it challenging to navigate the submission to find the answers for the case study tasks, a penalty of up to three marks will be applied.

The second part (15%), the individual work, will be assessed via **the online quiz set in Week 8**. This will test your understanding of machine learning concepts.

Assignment Components	Marks (20)		
Data Pre-processing	3.5	Variable Roles, Data Types & Measurements, Feature Selection	1.5
		Data Quality Enhancements	1.5
		Target Variable Distribution before & after	0.5
Decision Tree Models	4.5	Model 1	1.5
		Fine-Tuned Model	1.5
		Models Comparison	0.5
		Characteristics Summary	1
Regression Models	5.0	Pre-processing	0.5
		Model 1 + Finetuning	1.5
		Model 2 (Feature Selection) + Finetuning	2
		Models Comparison & Characteristics Summary	1
Neural Network Models	5.0	Pre-processing	0.5
		Model 1 + Finetuning	1.5
		Model 2 (Feature Selection) + Finetuning	2
		Models Comparison & Characteristics Summary	1
Predictive Mining Final Remarks: Comparison	2	Models Comparison	1
		Models Summary	1
Report Presentation	-3		
Team Agreement	-1		

Assignment 1 Criteria Sheet:

Criteria	Comments and scoring
Non-submission of all components/ evidence of plagiarism	0
Has demonstrated a task with a working model and the ability to run the program and add some components. Questions were poorly answered.	1-5
Has demonstrated a task with a working model and results with the substantial but incorrect implementation of at least one of three predictive models. Questions were poorly answered.	6-9
Has implemented models for all three tasks (three machine learning algorithms), with at least one being substantially correct. Shows some understanding of concepts with success in applying knowledge in fundamental questions.	10-12
Has implemented models for all three algorithms: Three of the six models are fundamentally correct, with substantially correct code that may contain minor errors. Response to questions shows a fundamental understanding of terms and concepts.	13-14
Has the fundamentally correct implementation of all six models, i.e. correct allocations of a target, rejections of variables according to instructions, running all models and comparing them. Almost all questions have been reasonably answered. Demonstrate a strong understanding of the methods and terms, including partitioning, imputation, modelling, overfitting, model interpretation, misclassification, F1score, macro average, weighted average, ROC chart during written analyses. Some minor errors are allowed. The written application is required to be of a reasonable standard.	15-16
Has implemented all of the requirements above with very few errors. A strong focus on the creative application of tools, evaluation, and interpretation of results is evident.	17-18
All criteria above are met; extensive model generation and analysis have been conducted to produce exceptional outcomes and have applied principles learnt in lectures to enhance the results. An outstanding presentation of the report and team agreement is included.	19-20