Case Study 1: Predictive Data Mining
Identifying Car Quality: “Is the purchased car a Kick”

Due date: 7th April, 2019
Weighting: 25%

Introduction

This assignment is intended to allow you to display your knowledge and understanding of predictive data mining. In this assignment, you will use SAS Enterprise Miner or Python, in particular decision tree, regression and neural network to display your technical competence gained from the practicals. It is also an opportunity for you to display the knowledge that you have gained from lectures and your readings and to show the relation between theory and practical.

The purpose of this assignment is to give you (1) an understanding that various methods can be applied to a data set and (2) the benefits of applying data analytics techniques to a data domain.

Instructions

1. The assignment is due on 7th April. It is a firm deadline.

2. You should submit the assignment via Blackboard Assignment.

3. The assignment (data mining results) **will be marked in the practical class**. Each group member will be asked specific questions about the case study in **week 7** practical lab. A 15% marks (out of 25 marks) will be assigned to you on the individual performance.

4. This is a group assignment. It is your responsibility to form a team of 3 members and you should do so preferably by week 3. Groups are to be ARRANGED and MANAGED by you. As in real life, the performance of individuals in the team shall be judged by the performance of the team together, so choose your partners carefully.

5. To ensure that everyone agrees as to their responsibilities in the team and how you will work together, we have asked that you complete a Team Contract. This should be done before the team is registered. You can find the team agreement template and guidance under the Assessment Item 1 link. You should update the file with your current group marks distribution with the final submission.

6. Once the team is formed, complete the team contact and register the team on Blackboard. Choose “Tools” from the left side of panel. Select the “Groups” tool and choose one of the IFN645 groups to register. This should be done by week 3.

7. Of course, the work you (group) hand in must be your own; no collaboration or borrowing from other groups is permitted. We will use the usual methods of detection of any plagiarism.

8. All the datasets required for this assignment can be found in the provided file named as **casestudy1-data.zip**.

9. A report should be submitted via online submission answering each question of the case study. There is no need of including introduction, summary, conclusion or references in the report. The report should just include responses to the questions set in the case-study. Some answers may require screen shots. Use them as needed, but you may include your own table detailing those results. While you may like to go into extreme detail about, you will not have the space to do so. Rather, write down the important points and attach the important screen dumps, to show that you have thought the matter through. The report is expected to be about 15-20 pages long. Remember to include the final diagram of the project showing all nodes connected in your diagram.

10. Name the case-study report as **casestudy1.doc**. The word file should include a cover page with Student ID number and full name (as in QUT-Virtual) for all students, along with the group name. Combine this file with your **team contract**, and name the compressed file as **casestudy1.zip.** Submit this file on **Blackboard (under the Assignment 1 link)**.

11. Read the Assessment Policies on Blackboard or QUT Website.

Case Study Scenario

A common challenge faced by a car dealership in purchasing a used car at an auto auction is to determine 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 the dealer is not able to address, issues with getting the vehicle title from the seller, or some other unforeseen problem. Kick cars can be very costly to dealers after transportation cost, throw-away repair work, and market losses in reselling the vehicle.

A dealership manager will 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.

You have been hired as a data analyst consultant by this management. Your task is to (1) predict if the car purchased at the Auction is a Kick (bad buy); and (2) inform decision makers the (characteristics of) potential kick cars.

Case Study Dataset

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

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 time of purchase

MMRAcquisitionAuctionCleanPrice	Acquisition price for this vehicle in the above Average condition at time of purchase
MMRAcquisitionRetailAveragePrice	Acquisition price for this vehicle in the retail market in average condition at time of purchase
MMRAcquisitonRetailCleanPrice	Acquisition price for this vehicle in the retail market in above average condition at time of purchase
MMRCurrentAuctionAveragePrice	Acquisition price for this vehicle in average condition as of current day
MMRCurrentAuctionCleanPrice	Acquisition price for this vehicle in above condition as of current day
MMRCurrentRetailAveragePrice	Acquisition price for this vehicle on the retail market in average condition as of 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

Table 1: List of Variables

Case Study Tasks

Your task is to build various predictive models such as decision tree, regression model, neural network and ensemble model on this data set and compare them. Name all the models meaningfully.

Set up a new project for this task with **DMProj1** as the project name or the Python file with **KICK** as the data source. Set **Kick** as the diagram (if using SAS). Include various models in this diagram or in this source file. Name all the models meaningfully.

Answer the followings (add screen shots as appropriate).

Task 1. Data Selection and Distribution.

1. What is the proportion of cars who can be classified as a “kick”?
2. Did you have to fix any data quality problems? Detail them.
3. Can you identify any clear patterns by initial exploration of the data using histogram or box plot?
4. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
5. What distribution scheme did you use? What data partitioning allocation did you set? Explain your selection.

Task 2. Predictive Modeling Using Decision Trees

1. Python: Build a decision tree using the default setting.
   SAS: Build a maximal tree (using the automated manner).
   Answer the followings:

   a. What is the classification accuracy on training and test datasets?

   b. What is the size of tree (i.e. number of nodes)?

   c. How many leaves are in the tree that is selected based on the validation data set?

   d. Which variable is used for the first split? What are the competing splits for this first split?

   e. What are the 5 important variables in building the tree?

   f. Report if you see any evidence of model overfitting.

   g. Did changing the default setting (i.e., only focus on changing the setting of the number of splits to create a node) help improving the model? Answer the above questions on the best performing tree.

2. Python: Build another decision tree tuned with GridSearchCV
   SAS: Build an optimal decision tree.
   Answer the followings:

   a. What is the classification accuracy on training and test datasets?

b. What is the size of tree (i.e. number of nodes)? Is the size different from the maximal tree or the tree in the previous step? Why?

c. How many leaves are in the tree that is selected based on the validation data set?

d. Which variable is used for the first split? What are the competing splits for this first split?

e. What are the 5 important variables in building the tree?

f. Report if you see any evidence of model overfitting.

g. What are the parameters used? Explain your choices.

3. What is the significant difference do you see between these two decision tree models (steps 2.1 & 2.2)? How do they compare performance-wise? Explain why those changes may have happened.
4. From the better model, can you identify which cars could potential be “kicks”? Can you provide some descriptive summary of those cars?

Task 3. Predictive Modeling Using Regression

1. In preparation for regression, is any imputation of missing values needed for this data set? List the variables that needed this.

Python:

2. Apply transformation method(s) to the variable(s) that need it. List the variables that needed it
3. Build a regression model using the default regression method with all inputs. Once you done it, build another one and tune it using GridSearchCV. Answer the followings:

h. Name the regression function used.

i. How much was the difference in performance of two models build, default and optimal?

j. Show the set parameters for the best model. What are the parameters used? Explain your decision. What are the optimal parameters?

k. Report which variables are included in the regression model.

l. Report the top-5 important variables (in the order) in the model.

m. What is classification accuracy on training and test datasets?

n. Report any sign of overfitting.

4. Build another regression model using the subset of inputs selected by RFE and selection by model method. Answer the followings:

    a. Report which variables are included in the regression model.

    b. Report the top-5 important variables (in the order) in the model.

    c. What are the parameters used? Explain your choices. What are the optimal parameters? Which regression function is being used?

    d. Report any sign of overfitting.

    e. What is classification accuracy on training and test datasets?

    f. Did it improve/worsen the performance? Explain why those changes may have happened.

SAS:

2. Build a regression model using the default regression method with the pre-processed data set. Answer the followings:

    a. Name the regression function used.

    b. Analyse the outcome and see whether the performance can be improved by using the selected variables (i.e. the subset of inputs selected either by stepwise or forward method).

    c. Choose the best model and report the followings:

        i. Which input selection method performed the best on this dataset or default was the best method?

        ii. What is classification accuracy/RMSE on training and test datasets on the best model?

        iii. How much was the difference in performance of various models build with each distinct method?

        iv. Show the set parameters for the best model.

        v. Which variables are included in this regression model?

        vi. Provide the top-5 important variables in this model? Did the top-5 variables differ in other models? List them too.

        vii. Report any sign of overfitting.

3. See whether you can further improve the performance by applying transformation to regularize input distributions. Report the variables that required transformation. What transformation function did you use and why?

4. Choose the best model in previous step (step 3.2: SAS) to apply transformation of variables. Does it improve the performance?

Python & SAS:

5. Using the best regression model, which cars could potential be "kicks"? Can you provide some descriptive summary of those cars?

Task 4. Predictive Modeling Using Neural Networks

1. Build a Neural Network model using the default setting. Answer the following:
    a. What is the network architecture?
    b. How many iterations are needed to train this network?
    c. Do you see any sign of over-fitting?
    d. Did the training process converge and resulted in the best model?
    e. What is classification accuracy on training and test datasets?

2. Python: Refine this network by tuning it with GridSearchCV.
SAS: Would change in architecture help here? Build another Neural Network model by changing the number of hidden nodes of the network.
Report the trained model.
    a. What is the network architecture?
    b. How many iterations are needed to train this network?
    c. Sign of overfitting?
    d. Did the training process converge and resulted in the best model?
    e. What is classification accuracy on training and test datasets? Is there any improvement in the outcome?

3. Would feature selection help here?
Python: Build another Neural Network model with inputs selected from RFE with regression (use the best model generated in Task 3) and selection with decision tree (use the best model from Task 2).
SAS: Build another Neural Network model with feature selection with regression (use the best model generated in Task 3) and selection with decision tree (use the best model from Task 2)..
Answer the following:

a. Did feature selection help here? Any change in the network architecture? What inputs are being used as the network input?

b. What is classification accuracy on training and test datasets? Is there any improvement in the outcome?

c. How many iterations are now needed to train this network?

d. Do you see any sign of over-fitting?

e. Did the training process converge and resulted in the best model?

4. Using the comparison methods, which of the models (i.e one with selected variables and another with all variables) appears to be better?

From the better model, can you identify cars those could potential be "kicks"? Can you provide some descriptive summary of those cars?

Is it easy to comprehend the performance of the best neural network model for decision making?

Task 5. Generating an Ensemble Model and Comparing Models

1. Generate an ensemble model to include the best regression model, best decision tree model, and best neural network model.

   a. Does the Ensemble model outperform the underlying models? Resonate your answer.

2. Use the comparison methods (or the comparison node) to compare the best decision tree model, the best regression model, the best neural network model and the ensemble model.

   a. Discuss the findings led by (a) ROC Chart (and Index); (b) Score Ranking (or Accuracy Score); (c) Fit Statistics; (or Classification report) and (4) Output.

   b. Do all the models agree on the cars characteristics? How do they vary?

Task 6. Final Remarks: Decision Making

1. Finally, based on all models and analysis, is there a particular model you will use in decision making? Justify your choice.

2. Can you summarise positives and negatives of each predictive modelling method based on this analysis?

3. How the outcome of this study can be used by decision makers?

Marks Distribution

In data analytics, there is hardly ever a single solution. The solution depends upon various setting such as input variables role and measurements, training size and the selected method parameters. You may find that your project partner may have different solution as yours. Your group should decide on a single project that you would like to be marked. Submit the report discussing the final project components.

We would mark your data mining project in the Week 7 practical class to explore your understanding of the data mining concept. You should be prepared to show your final diagrams and results panels to your marker. The marker will ask each individual student questions and will assign individual mark (~15%).

Assignment Components	Marks
Data Pre-processing	3.5
Decision Tree Models	5
Regression Models	6
Neural Network Models	5
Comparison: Predictive Models	2
Final Remarks: Decision Making	2.5
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 with /without submission and demonstrates the ability to run the program and add some components. Questions were poorly answered.	1-5
Has demonstrated a task with a working model having a data source, and diagram with substantial but incorrect implementation of at least one of the three components . Questions were poorly answered.	6-9
Has implemented models for all three tasks (three data mining algorithms) with at least one being substantially correct. Shows some understanding of concepts with some success applying knowledge in basic questions	10-13
Has implemented models for all three tasks with pre-processing and ensemble modelling. Two of the three tasks are fundamentally correct, with substantially correct work flow diagrams which may contain minor errors. Response to questions shows fundamental understanding of terms and concepts.	14-17
Has fundamentally correct implementation of all six tasks i.e. correct allocations of targets where available, rejections of variables according to instructions, models and their comparison/ assessment. Includes a demonstration of competent application of tools. Almost all questions have been reasonably answered. Demonstrate a strong understanding of the methods and terms including predictive mining, partitioning, imputation, comparison node, ensemble, misclassification, average squared error, sensitivity, specificity, lift, ROC chart, lift chart, support and confidence during written analyses. Some minor errors are allowed. Written application is required to be of reasonable standard.	18-20
Has implemented all of the requirements above with very few errors. A strong focus on application on creative application of tools, and evaluation and interpretation of results is evident.	21-23
All of the criteria above are met; extensive model generation and analysis has been conducted to produce quality outcomes and have applied principles learnt in lectures to enhance the results.	24-25