University of South Australia

Assignment 2

Building a decision tree predictive model

Due Date: 11PM Sun 15 Oct, 2023
Word limit: ~2,000 words (excluding spaces, tables, and references)

Submission

The assignment should be completed in groups or individually and submitted through learnonline. Please make sure that you complete all the parts of the assignment.

Assignment Requirement

This assignment is the continuation of the project we started in Assignment 1. In that sense, it is critical to address feedback you received on your previous assignment as you go through Assignment 2.

In this assignment, our objective is to build upon the work done in Assignment 1 by developing a decision tree model to predict healthcare fraud. The primary addition in this assignment is the development of a predictive decision tree model using the dataset we have previously explored. This model will leverage the features and insights we have derived from our earlier data exploration and feature engineering efforts. This assignment will have the following parts:

1. **Introduction** (2 marks). In this part you will briefly discuss the goal of the assignment, focusing on addressing the comments you received in the previous assignment.
2. **Related Work** (3 marks). Revisit the papers you reviewed in Assignment 1, incorporating the feedback you received. Discuss any new insights or understandings gained from this process, and how they might influence your approach to the project.
3. **Data exploration** (5 marks). Using R and RStudio for descriptive statistics and visualizations, in **approximately 600-700 words**, in this part you will:
    a. Explain the features you decided to extract. This can be exactly the same set of features you extracted in Assignment 1. However, you may decide to revise previously decided feature set and you should address comments raised in the feedback you received.
    b. Provide descriptive statistics (using tables and figures) of your feature set. This **may include** value distributions, skewness, histograms, bar plots, box plots, and other details that you find relevant.
    c. Along with the univariate analysis (such as value distribution), you are also expected to investigate the correlation between variables that you find important (this does not have to include all the features). A useful tool here would be correlation plots or correlograms. When running correlation analysis, please make sure that included variables are actually suitable for this kind of analysis.

    This section should also be structured around the content you provided in Assignment 1, taking into the account the feedback you received. The section should be updated to reflect any changes you made in features selection.

4. **Building decision tree models** using R (10 marks). In Practical we used `rpart` to fit a decision tree model. However, you are welcome to use any R library to do so. Whatever the method/library you decide to work with, in this part you should:
   a. Explain data partitioning. What is the ratio of data you decided to keep for training vs. testing and why?
   b. Explore various parameters, such as the maximum depth (maxdepth in rpart), the minimum number of samples a node must have before it can split (minsplit in rpart) or complexity parameter (cp in rpart). In so doing, you are expected to explain what those parameters represent as well as to fit several models (3-4 would be sufficient) for different values of those parameters.
   c. Provide a tree summary or plot. Those plots can be rather large and not very clear. In that case, please make sure you include at least a part of your tree.
   d. Report various performance indicators for each of the models you fit. Those metrics may be accuracy, precision, recall, F-score, kappa, AUC, and other performance indicators that you may want to include. It should be fairly straightforward to find an R function to extract those metrics. If not, please focus on the metrics you can calculate from the confusion matrix.
5. **Compare the models** you built in the previous part (5 marks). The focus should be on describing performance indicators and selecting the best performing model. For the model that yields the best performance, you should explain the main splitting attributes and discuss the variables that are most predictive of the outcome.

While there is a word limit assigned to this assignment, your submission may be longer (within reason).

The report should include a cover and table of content.

Marking Criteria

High Distinction

In meeting this level, you will address all the parts of the assignment, demonstrating clear understanding of the topics covered in the course, also taking into the account feedback received on the previous submission. Data inspection will be supported with relevant, quality, figures and accompanied explanation of observed trends. All the figures and tables will be labelled. Clear description of the decision tree building would be provided. Finally, you will identify relevant metrics for model comparison, explaining what each of the metrics mean and how you interpret them in this context. This level requires clear and coherent writing, with the concise narrative. Overall, in meeting this level, you will demonstrate a comprehensive knowledge of the concepts through your descriptions, explanations, and discussions of the content.

Distinction

In meeting this level, you will address all the parts of the assignment, including the feedback you received on the previous assignment. Data inspection includes basic plots (e.g., histograms, correlation matrix) that are supported with relevant, quality, figures and accompanied explanation of observed trends. All the figures and tables will be labelled. Clear description of the decision tree building would be provided. Basic metrics (e.g., accuracy or recall) for model comparison are being used for model comparison. Writing is clear and the narrative is coherent across the report. Overall, in meeting this level you will demonstrate a well-considered knowledge of the concepts through your descriptions and explanations.

Credit

In meeting this level, you will address at least two parts of the assignment, where initial inspection would be a mandatory part, addressing parts of the feedback you received on your previous assessment. Data

inspection includes basic plots (e.g., histograms) that are supported with relevant figures and accompanied explanation of observed trends. Overall, in meeting this level, you will demonstrate a sound knowledge of concepts through your descriptions and explanations.

Pass

In meeting this level, you will address at least two parts of the assignment. In doing so, you will demonstrate knowledge of the concepts through your descriptions.

Academic integrity

You are expected to reference and cite all resources mentioned using a selected referencing convention (e.g., UniSA Harvard, or APA).

Extensions

Extensions for assignments are available under the following conditions

- permanent or temporary disability, or
- compassionate grounds

In all cases, documentary evidence (e.g. medical certificate, road accident report, obituary) must be presented to the Course Coordinator. A medical certificate produced on or after the due date will not be accepted unless you are hospitalized.

If you apply for extension within 24 hours before the deadline, you must see the course coordinator in person unless you are in an emergency like being admitted in a hospital.

Late Penalties

Unless you have an extension, late submission will incur a penalty of 30% deduction per day (or part of it) of lateness.