BUS5PA Predictive Analytics – Semester 2, 2022

Assignment 2: Building and Evaluating Predictive Models using SAS Enterprise Miner

Release Date: 2nd September 2022
Due Date: 20th September 2022
Weight: 40%
Format of Submission: A report (electronic form) and electronic submissions of SAS projects

Objective:

a) Demonstrate knowledge of building different types of predictive models using SAS Enterprise Miner
b) Demonstrate skill and knowledge in applying predictive models in a real-life predictive analytics task
c) Relate theoretical knowledge of predictive models and best practices to application scenarios

PART A:

1. Exploratory analytics– COVID-19 outcomes…… (20%)

On December 31, 2019, the World Health Organization (**WHO**) was informed of an outbreak of "pneumonia of unknown cause" detected in Wuhan City, Hubei Province, China. Early on, many of the patients in Wuhan, China, reportedly had some association with Wuhan South China Seafood Market. The virus causing the outbreak was quickly determined to be a novel coronavirus (now known as Covid-19). Gene sequencing further determined it to be the new coronavirus, a betacoronavirus, related to the Middle Eastern Respiratory Syndrome virus (MERS-CoV) and the Severe Acute Respiratory Syndrome virus (SARSCoV). However, the mortality and transmissibility of Covid-19 are still unknown, and likely to vary from those of the prior referenced coronaviruses.

For this task, we have preprocessed and transformed a publicly available Covid-19 dataset[1] and provided a description of the selected variables in figure 1. We have also generated a simple decision tree model using the data set for your reference (figure 2). You are expected to carry out an exploratory analysis and predictive modelling features in SAS Enterprise Miner to explore and understand the data and build a predictive model targeting the outcome (will patient recover/or not). Carry out the following tasks:

Use SAS Enterprise Miner to explore the data and build predictive models. Use the decision tree provided as a guide to check and compare your models.

Use the reference articles (provided – or similar) to obtain background knowledge from analysis carried out on corona patients.

[1] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Prepare a report (2 pages with screen shots/diagrams) to discuss the outcomes of the predictive models and data exploration. Does the predictive model represent and confirm the Covid-19 analysis provided in external reports you referenced? Are there any key aspects in recovery prediction which are not being identified by your models? How do you explain/interpret using your knowledge of predictive model development?

Table 1: Covid 19 data set – variables and description

Variable	Description
ID	ID of the patient
age	Age
sex	Sex
city	City
province	Province
country	Country
latitude	Geo location (latitude)
longitude	Geo location (longitude)
has_fever	Shows symptoms of fever (yes/no)
has_cough	Shows symptoms of cough (yes/no)
has_respiratory_problems	Shows symptoms of respiratory problems (yes/no)
lives_in_Wuhan	Lives in Wuhan
reported_exposure	Reported exposure
chronic_disease	If the patient has a chronic disease
outcome	**Outcome of treatment (Deceased/ Recovered) - Target**
days_to_confirm_covid	Average number of days from symptoms to confirm COVID

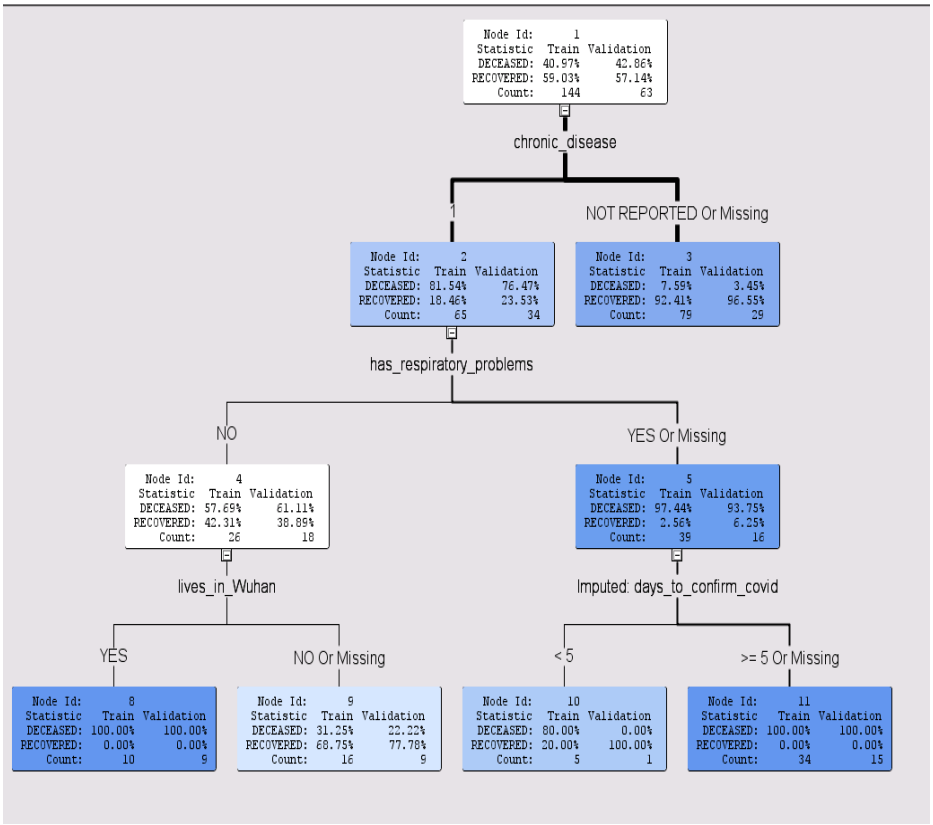

Figure 1: SAS Enterprise Miner decision tree model for selected covid19 data set described in Table 1

PART B:

Business Case – Predictive Model for Vehicle Price Prediction

Beta (Pvt) Ltd is an Australian online car sales platform for providing effective car buying and selling services. In order to help boost sales transactions, the management of beta is in the process of building a car price estimation system to help second-hand car sellers to sell their cars at the best price.

Beta management is very keen to trial predictive modeling for this task and has gathered a historic car sales dataset from a publicly available data repository. The dataset contains 23 attributes of previously sold cars. The attributes include the selling price of cars, age, kilometers driven, fuel type, automatic or manual, number of doors, etc. The list of attributes and their descriptions are given below.

Variable	Description
no_of_airbags	Available air bags for protection
fuel_average_distance_km	Average distance that can be travelled with a full tank of fuel
engine_size	Engine capacity
year	First sold year
fuel_consumption	Fuel consumption of the vehicle (km/l)
fuel_type	Fuel type
doors	How many doors
gears	How many gears in the gearbox
engine_location	Location of the engine
body_configuration	Luggage space
cylinders	Number of cylinders in the engine
valves_ports_per_cyl	Number of valves per cylinder
price	**Price sold in the second-hand market (Target)**
rim_material	Rim design material
ancap rating	Safety rating indicating the level of safety provided in a event of a crash
seat_capacity	Seat capacity
odometer_km	Total distance travelled by the vehicle
fuel_capacity	Total fuel capacity in the fuel tank
gear_type	Type of the gear
body_style	Vehicle body style
transmission	Vehicle gearbox type
title	Vehicle name
emission_standard	European emission standards define the acceptable limits for exhaust emissions of vehicles sold (https://en.wikipedia.org/wiki/European_emission_standards)

The management of Beta.com Ltd. is considering you as an external consulting group to outsource the task to develop a reliable predictive model to predict the selling price of the cars, using the aforementioned historic dataset. Beta has provided you with a sample data sets of Toyota and BMW cars to build separate

price-prediction models. They also wish to compare and contrast the attributes that are useful for price prediction in Toyota and BMW models.

1. Setting up the project and exploratory analysis (10%)

a. Create a new project and create both Toyota (**toyata_train**) and BMW (**bmw_train**) data sources. Set price as Target.

b. Carry out a data exploration by using a **StatExplore** Node. Explain your findings with regards to Toyota and BMW datasets.

c. Create a **Data Partition** with 70% of the data for training and 30% for validation.

2. Decision tree-based modeling and analysis (20%)

Carry out the following modeling tasks for both Toyota and BMW datasets separately.

a. Create **two Decision Tree** models. Use **two-way** and **three-way** splits to create the two separate decision tree models.

For each decision tree,

I. How many leaves are in the optimal tree?

II. Which variable was used for the first split?

III. What were the competing splits for this first split?

b. Which of the decision tree models appears to be better? Justify your answer.

c. Refer to the selected decision tree model in part (b) and

I. Identify leaf nodes which have good predictive performance (two leaf nodes) and poor predictive performance (two leaf nodes).

II. Provide justifications for your selections

III. Write down the rules for the pathways leading up to each selected leaf node.

3. Regression based modeling and analysis (20%)

Carry out the following modeling tasks for both Toyota and BMW datasets separately.

a. In preparation for regression, is any missing values imputation needed? If yes, should you do this imputation before generating the decision tree models?

Why or why not?

b. Use an **Impute** node connected to **Data Partition** node to handle missing values. Which variables have been imputed?

c. Conduct **data exploration** to select the best variables for the model with **Variable Clustering** node. Explain your findings.

d. Create a **Regression** model using the set of variables you identified as suitable in part c. You can choose the stepwise selection and use validation error as the selection criterion.

e. Run the **Regression** node and view the results.

I. Which variables are included in the final model? Explain what this means to the vehicle sales organization (very briefly).

II. What is the validation ASE? What does this mean in a predictive model?

4. Model Comparison and Scoring (30%)

a. Use the model comparison (separately for Toyota and BMW predictive models) to compare and contrast the results from the decision trees and regression based analysis. Describe and justify how you ascertained the better model.

b. Compare and contrast the best model selection for Toyota and BMW separately. Would it have been sufficient to use only one modeling technique (decision tree or regression)? Provide justifications for your answer.

c. Which variables were used in the predictive models to determine the price of Toyota vehicles and BMW vehicles? Discuss further comparing the feature importance of two models.

d. Use **toyota_score and bmw_score** data sets to score the best model for Toyota and BMW respectively (remember to change the role of the data sets to Score). Explain the output using plots.

(Hint: You may use screenshots from your Enterprise Miner project in the report. Answer for this part should not exceed 4 pages).