Machine Learning Assignment Spring 2023

Store Performance Classification

Project Objective: Develop classification models that will identify whether stores will perform well or poorly.

Context:

The owner of a chain of over 100 shops in the UK has asked you to build a machine learning model capable of classifying stores as having either good or bad performance based on details about the store. For each store, the company has recorded data on performance (simply 'good' or 'bad') and details such as the size of the store and the number of staff it employs. You have been provided with a dataset and a data dictionary that describes the data.

Your Task:

You must develop logistic regression, decision tree and neural network models that will identify whether stores will perform well or poorly. You can use Orange, Python, R, or any data mining package of your choice. The data for the assignment is in a file storedata.csv, which you can download from the same place you found this document. The data dictionary is given at the end of this document. You must follow the correct methodology to use the data to build and test your models.

What to Submit:

You must submit a **single page** infographic poster showing the results of your analysis. Create the poster using a word processor (like Word) or a presentation package (like PowerPoint). Set the page size to A3 and use a 12pt font. You can choose the layout, but you must include:

1. Put your student number (not your name) at the top of the poster
2. A list of the steps you took to carry out the project, including details of the train / validate / test split that you chose;
3. A table showing which variables you used and whether your model treats them as numeric (continuous or discrete) or categorical. Explain one consequence of your choices;
4. A single example of how you used a histogram to detect an error in the data and what you did to fix that error. State the data cleaning operation you carried out;
5. A table showing the different models and hyper parameters you trained, along with the correct metric for each; Add a sentence on how you chose the hyper parameter values.
6. A justification of the choice of the final model and a confusion matrix showing its results

on the correct data split. Add a comment on what the confusion matrix tells us about how useful the model will be.

7. One or more references that are used to justify decisions that you made and cite any materials that you used (such as illustrations).

Important: Make sure your poster has the sections listed above. Missing sections will cost you a lot of marks.

Save your poster as a PDF and submit it on canvas. Check the canvas page for the deadline and other submission rules.

Marks:

Marks will be allocated as follows:

Mark Range	**Requirements / Achievement**
0 – 39 (clear fail)	Many sections missing, no models built, little work completed
40 – 49 (marginal fail)	Most sections completed, but with methodological errors and poor or no interpretation
50 – 59 (pass)	All sections completed with correct methodology, but with poor or no interpretation of results
60 – 69 (merit)	All sections completed with correct methodology, good interpretation of results, good justification of decisions
70 – 100 (distinction)	All sections completed with correct methodology, excellent interpretation of results, good justification of decisions with references, excellent presentation, excellent insight.

MOST IMPORTANT OF ALL!!

Read this part very carefully. You must work on this assignment completely on your own. Do not ask classmates for advice and do not share your answers with anybody. Do not copy any work from the internet. If you do look up something online, make sure you include a reference to it. Marks will not be given to work that is clearly not your own.

Data Dictionary:

Variable	Description
Town	Name of town where the store is located
Store ID	Unique ID for the store
Manager name	Name (fictional) of the store manager
Staff numbers	Number of staff employed
Floor Space and Window Space	Floor area and window area of the store
Car park (yes or no)	Does the store have a car park?
Demographic score	How well does the demographic of the local population match the ideal demographic of the store? Higher is better
Location (Shopping Centre, High Street, Retail Park)	Type of location for the store
40 min, 30 min, 20 min and 10 min drive time population size	The population within each of the different drive time distances
Store age	The age of the store in years
Clearance space in store	How much of the store's floor space is given over to discounted clearance stock
Competition number	How many competing stores are near the store
Competition score	A measure of quality of the competing stores. High = high competition
Performance	Good or bad. An assessment of how well the store is performing