Advanced (Business) Data Analysis

ASSIGNMENT 1

Summary

This assignment is an individual assignment. The aim is to provide experience in the steps involved with creating, evaluating, improving classification models, and finally presenting and interpreting the model in a business report. You are strongly encouraged to commence this assignment after its release, and you should progress thoughtfully through the steps. Hasty decisions made early in the design process may result in much more work later. Feel free to discuss concepts and ideas with peers but remember your submission must be your work. Be careful not to allow anyone to copy your work.

Specification

The Australian real estate market is one of the most dynamic and competitive markets in the world, with a wide range of properties available to buyers and sellers alike. However, with such a large and complex market, it can be difficult to accurately predict how long a property will stay on the market before it is sold.

You have been given access to a rich dataset of property sales data from the previous year, which includes information on recently sold properties and their sales history. Using this data, your task is to develop a classification model that can help classify the "days on market" for a given property based on a range of factors, including its location, features, size, and other relevant details.

Your classification model will be a powerful tool for real estate professionals, helping them to better understand the factors that influence a property's time on the market and make more informed decisions about how to sell it effectively. By analyzing the data and identifying the key variables that affect a property's days on market, you can help real estate professionals optimize their pricing strategies, marketing efforts, and other key aspects of the sales process. The analysis will also be helpful for the seller themselves to help them better plan and prepare for the sales process by understanding a realistic asking price to developing a targeted marketing campaign.

There are many factors to consider when classifying days on market for a property. These might include the property's size and amenities, as well as its location, the local real estate market conditions, and a host of other external factors. By carefully analyzing all these variables and developing a robust classification model, you will help real estate professionals stay ahead of the curve and make better, more informed decisions about how to sell properties in this highly competitive market.

Dataset

The property data provided consists of 28,509 recently sold properties and their sales history/details. The provided dataset contains 17 inputs between the dates February 2022 to February 2023. The classification goal is to predict if the property has a high or low 'days on market' using the days_on_market variable.

There are 3 types of input variables and only 1 target variable:

A) INPUT - Location data:

1. address (type: categorical)
2. suburb (type: categorical)
3. state (type: categorical)
4. postcode (type: categorical)

B) INPUT - Related to property details:

1. number_of_beds: The number of bedrooms on the property (type: numeric)
2. number_of_baths: The number of bathrooms on the property (type: numeric)
3. number_of_parks: The number of parking spots on the property (type: numeric)
4. property_size: The size of the property in square meters (type: numeric)
5. listed_date: The date the property was listed for sale (**type: date**)
6. listed_price: The price the property was listed for (**type: date**)
7. sold_date: The date the property was officially sold (type: numeric)
8. sold_price: The price the property was officially sold for (type: numeric)
9. property_classification: The type of property (House/Unit/Land) (type: categorical)
10. property_sub_classification: The sub-type of property (type: categorical)

C) INPUT - Related to suburb details:

11. average_days_on_market: The average days in market that a property is on sale for in a specific suburb compared to similar properties (type: numerical)
12. average_median_price: The average median property price in a specific suburb compared to similar properties (type: numerical)

E) OUTPUT - Desired Target:

- days_on_market: Indicates whether a property has a high or low number of days on market (type: binary)

Deliverables

You should submit a written report via TurnItIn and a Jupyter Notebook file.

Your reports should include the following parts:

- Introduction, data exploration and clustering
- Model building: you can to develop your initial model with a classification technique like NN. Then try alternative approaches, such as DT, NB and KNN.
- Model evaluation: evaluate your initial model and alternative model using various approaches.
- Conclusion and summary: Include those results that are most significant for your strategy development and recommendations and justify them.

You may decide to have four main sections in your report, including *a) introduction, b) model building, c) model evaluation, and d) conclusion & summary*. Alternatively, you may decide to combine part b) and part c) and have three sections, including *1) introduction, 2) model building and evaluation, 3) conclusion and summary*. Both approaches are accepted.

It is up to you to decide what proportion of your report goes to each part. You may include tables, charts, or tables of your analysis and models. At the end of your analysis, your Jupyter Notebook should be uploaded along with your report.

The consistency of your Jupyter Notebook will be checked with the results in your report. You **do not need** to provide the **screenshots** of your Jupyter Notebook, as the marker can observe them from your file. Consider the following points for designing your process:

- You need to create only one Notebook with as many cells and outputs that are needed.
- You should not modify "A1_Data.xlsx" file before importing it in your Notebook.
- All of your analysis should be done after importing "A1_Data.xlsx" in Jupyter Notebook, not Excel, or any other analytical tool.

Formatting and professionalism

The project report is to be written to a professional standard. This requires a formal writing style – do not use dot points - and adopt a professional tone. Given the report's nature, you may choose to write this report in the first person. The report must be consistent with the University's policies on academic integrity, plagiarism, and consequences as noted below. The report should be typed (in Times Roman 12-point font or larger, single-spaced) and the Word Count should be 1500 words (+/- 10%) in total length. The Word Count excludes the title page, tables, footnotes and references (if required). The word limit must be observed or the assessment will be affected as noted in the Rubric. No appendices are to be provided.

Submission

Submissions are to be done via the two following links in the Assignment 1 folder:

- Written report to be submitted through the TurnItIn link.
- The Jupyter Notebook to be submitted by the Assignment Submission link.

Acceptable submission formats for the written report are Microsoft Word and PDF formats, and a Jupyter Notebook file. The files must be named in the format of StudentID.pdf (or a. docx or .doc extension). If your ID is 41724593, the name of your files would be 41724593.pdf or 41724593.ipynp. The written assignment file should not be zipped.