ACST4062/8032

Actuarial Data Analysis

Assignment One

This assignment is worth 20% of your final grade.

This assignment is out of 50 marks.

This assignment must be submitted by 3pm on Monday, 20 September 2021. If it is handed in later than this time, your mark for this assignment will be 0%.

Assignments are to be submitted using Turnitin on the course Wattle site. Submitted assignments must include the cover sheet provided in this document. Please keep a copy of the assignment for your records.

The ANU is using Turnitin to enhance student citation and referencing techniques, and to assess assignment submissions as a component of the University's approach to managing Academic Integrity. University policies on plagiarism will be strictly enforced.

Background

You are an actuary working in the analytic team of Providence Insurance, a general insurance company. Providence provides a range of insurance services for customers, such as motor insurance, home & contents insurance, and others.

Recently a hailstorm hit Cranberry City, causing damage to the properties in the region. Your team has been tasked to estimate the claim cost related to this event. You are given two specific tasks:

Data Quality Assessment and Exploratory Data Analysis [30 marks]

Providence maintains a centralized database which contains information of the insured properties. The dataset `internal` contains a batch of property data which has not yet been integrated to the database. The dataset `independent` contains the valuation of the properties conducted by a third-party real estate company. The descriptions of the datasets are given in the Appendix. These two datasets have to be combined and cleaned before being integrated into the centralized database.

Your first task is to combine the two datasets `internal` and `independent`, assess the quality of the data and make any necessary adjustments. Then, conduct an exploratory data analysis on the combined and cleaned dataset. You have also been told to pay more attention to dwelling size and the valuation of the properties as these variables are known to be useful in estimating claim amount. Note, you are not required to show/integrate the combined and cleaned dataset.

Instead, write a short report that:

- Outlines the key actions and considerations that you have taken to prepare the cleaned and combined dataset; and
- Summarises the key insights you obtain from your exploration analysis.

Use a combination of words, graphs and/or tables to explain your findings.

You should provide reasoning to your analysis choices whenever appropriate. For example, you should provide explanation for omitting data and/or adjusting the values of the data. You are not required to include any R code in the body of your report but in the appendix of your report (see further instruction at the end of the assignment). However, you may include snippets of R code in the report if it helps with your description.

Rubrics:

- 8 marks for cleaning, adjustment (if any) and exploratory analysis on valuation.
- 8 marks for cleaning, adjustment (if any) and exploratory analysis on dwelling size.
- 8 marks for cleaning, adjustment (if any) and exploratory analysis on all other variables.
- 6 marks for report presentation and communication.

Predictive Model [20 marks]

Following the hailstorm, the Claim Processing (CP) team has been gathering data to investigate the damage done to the insured properties in the region. The dataset `damaged` contains about 50% of all insured properties in the region and whether the property was damaged in the hailstorm.

The CP team would like to have a model that predicts the probability of a claim given the location of the property. They have built themselves a simple model for this purpose. The performance results of their model are given to you as the R object `cpt_results`.

You have been asked to build a predictive model based on the `damaged` dataset using a k-nearest neighbour model for the CP team. Compare the performance of your model with the CP team's current model and determine which model offers better predictive performance.

The CP team is also keen to re-use the model to predict damage caused by other weather events in the future.

Write a short report that:

- Describes the key steps of building a KNN model to predict whether a property is damaged during the hailstorm; and
- Discuss which model (your KNN model or the CP team's model) offers better predictive performance; and
- Discuss the viability of using the exact model you built to predict damaged caused by other weather events in the future.

Use a combination of words, graphs and/or tables to explain your findings.

You should provide reasoning to your modelling choices wherever appropriate. For example, you should justify the choice of variables to be included in the model, and the k hyperparameter. You are not required to include any R code in the body of your report but in the appendix of your report (see further instruction at the end of the assignment). However, you may include snippets of R code in the report if it helps with your description.

Rubrics:

- 10 marks for building a predictive model using k-nearest neighbour.
- 5 marks for model comparison.
- 5 marks for discussing the viability using the model for other weather events.

Appendix - Data description

Variable	Unit	Description
id		Property ID
bathrooms	integer	The number of bathrooms in the property. 0 means missing.
bedrooms	integer	The number of bedrooms in the property. 0 means missing.
dwellingSize	m^2	This refers to the total indoor area of the property.
type		House - a freestanding house. Townhouse - a house with shared walls. Unit - Apartment or low-density housing with shared facility such as common area and garage. Other - a catch-all category that includes all other unspecified property type and missing values.

Table 1: Variables for dataset `internal`

Variable	Unit	Description
id		Property ID
valuation	character	Property valuation by a third-party real estate company. Sometimes it is a range instead of a single value.

Table 2: Variables for dataset `independent`

Variable	Unit	Description
id		Property ID (de-identified)
lng		The longitude coordinate of the property.
lat		The latitude coordinate of the property.
damage	Y/N	Whether the property is damaged by the hailstorm.
cladding		The main material used to construct the outside walls of the property.
roofing		The main material used to construct the roof of the property.

Table 3: Variables for dataset `damaged`

Variable	Unit	Description
prob_estimate	continuous	The predicted probability of the property being damaged by the hailstorm.
truth	Y/D	Whether the property is damaged by the hailstorm.
lng		The longitude coordinate of the property.
lat		The latitude coordinate of the property.

Table 4: Variables for dataset `cpt_results`

Further instructions:

Due date: Monday 20 September 2021, 3pm. Late submissions will receive a mark of 0%.

Cover page: Please include a cover sheet in your submission. The cover sheet is available on Wattle and it does not count towards the page limit.

Page limit: Please submit your assignment in a word or PDF document not more than 12 pages. You should think about how to best display your answers, working and assumptions within this limit and marks will be awarded for presentation and communication. You may (and should!) include relevant tables and graphs in your document. Please ensure you explain your analysis and justify any modelling choices.

Appendix: Please submit the relevant R code for your exploratory data analysis and model building. The appendix does not count towards the page limit. The appendix will not be marked but might be checked to clarify a response in the report if necessary.

Reference: You do not need to reference any other material to complete this assignment but if you do, please ensure you properly reference your work. You must adhere to appropriate practices regarding referencing the work of others in any work that you do: Accepted academic practice for referencing sources that you use in presentations and assignments can be found via the link on the Wattle site.

End of assignment