School of Risk and Actuarial Studies

ACTL4305/ACTL5305
Actuarial Data Science Applications

Datathon Assignment 2024

This is a real industry project partnered with Fetch Pet Insurance (not just another assignment)!

Background

Pet insurance is booming in Australia, with lots more people getting pets and looking for cover for unexpected and rising vet costs. In Australia, 70% households have a cat or a dog, 11m pets in total. Australians spend $33bn a year, including $10bn on vets, health, and insurance. Gen Y & Z are now the now largest pet parent group, with ~80% of them having pets. They see them as their first child or fur-baby 🐶🍼, taking more interest in them and spending more on their health and well-being.

Everyday, there's more we can do for our pets - with medical advances, treatments for complex orthopaedic procedures, advanced diagnostics, and cancer are common, but they can cost up to $20k.

Opportunity

You've identified a chance to tap into this market by offering insurance tailored to owner demographics, breed risks, geographic factors, and any other data you can get your hands on.

However, you're not alone—**40 other startups** are entering the same market, all aiming to get a slice of the pie!

As one of the new pet insurance startups launching in Australia, you'll need to create a unique brand and establish a market-leading, competitive pricing model.

Price too high, and you'll lose customers to competitors with more attractive rates.

Price too low, and while you may capture a larger market share, you risk losing money on each sale. Success in this competitive environment hinges on your ability to design a pricing model that attracts the right customers at the right price while maintaining profitability.

You've got **6 weeks** before your funding runs out to launch your brand, product and your pricing model into the market.

Key Considerations

- **Regulated market**: You must build a model with **high explainability** to ensure it meets regulatory requirements. While the final pricing model must be highly interpretable, feature engineering and exploratory data analysis can initially be conducted using less interpretable models.
- **Correlation vs. Causality**: Many variables may be correlated but not necessarily causal. Use your judgment to determine the factors that directly influence risk to avoid noise in your model.
- For this model, please **focus on technical pricing (pure premium)** only. Exclude expenses and tax assumptions.

Approaches to win in the market

- **Getting more data:** Identify additional factors or data sources that can enhance the accuracy of your pricing. You have access to data purchased from the market, but you noticed that you could take free data from the ABS to enrich your insights. You can use any publicly available information for your model.
- **Data Manipulation & Feature Engineering**: Extract more insights from existing data through creative feature engineering and data manipulation. For example pet date of birth is not very useful but pet age may be an important indicator.
- **Model specifications:** Common practices in the industry include building separate models for **claim frequency** and **claim severity**. Others prefer to

combine these into a single model for total costs. Additionally, some models treat **small** and **large claims** separately to capture differences in risk. Others categorise different types of claim costs (e.g. **condition category**) into different additive pricing models.
How you approach splitting or combining your data into homogenous risk groups influences how well your pricing model performs.

Progress Milestone Deliverables

Your co-founders are getting antsy. It's a few weeks from launch, and they haven't seen anything from the pricing model yet. They ask you to share some of your early findings to let them know everything's under control. You promise to deliver:

1. **Brand & Logo**: A document with a unique brand name and logo attached (png or jpg) that resonates with your target market. Explain why you chose these and why you believe it will appeal to your target market.
2. **Expected top 3 pricing factors:** From the existing datasets, show in a 1-page summary the **top 3 factors** that you think the market is currently using to price pet insurance. Include your hypothesis on why these will be important factors and show whether your hypothesis is likely to be true with any plots or analysis that you've done.
3. **Creative 3 pricing factors:** Using your street smarts, develop another **3 creative factors** that you believe will differentiate yourself from the market. You can do this by bringing any publicly accessible data, or combining or transforming any of your existing data set. Show this in a separate 1-page summary including your hypothesis, and any plots or analysis.

Please refer to the Assignment outline and marking guide document for detailed submission instructions and due dates.

Final Deliverables

You need to deliver a comprehensive pricing model to the Chief Actuary & Co-founder Fei. To ensure your model is prioritised and implemented, you must convince her that it is sound, logical, and well-founded. Fei has very little time, and will only take a meeting when a succinctly written report outlining your findings is sent to her beforehand.

1. **Reproducible Codes**: Reproducible codes for this pricing challenge with instructions on how to use them (such as RMarkdown files or R scripts with

a readme file, including datasets used).

2. **Pricing Output:** You will be provided with a set of ~10,000 prospective customers. Use your pricing model to output a premium for each of the customers, rounded to the nearest cent. Note that a “Sample_price_output_file.csv” is included in the Insurance Dataset (download link available on Moodle).
3. **Final Report**: Provide an executive summary with a report (max 8 pages) illustrating the problem solving process. This should include findings and decisions made from exploratory data analysis, modelling, comparison, evaluation and interpretation.
4. **Video Presentation:** Prepare a 5-minute video presentation for the Chief Actuary and Co-founder, Fei. In this presentation, detail the **unique** aspects of your group's model and explain how they differentiate your approach from competitors, providing a competitive advantage in the market. You may choose who should participate in the video, and there are no requirements regarding the number of participants.

Please refer to the Assignment outline and marking guide document for detailed submission instructions and due dates.

Market Dynamics

You have been given a sampled set of ~10,000 prospective customer profiles which are representative segments of 1 million participants in the market. You and 40 other startups will be competing for these customers with your business. These customers are price elastic, and your price relative to the market will heavily influence the number of customers that choose to insure with you.

For each customer, the price you provide will be ranked against the prices that other competitors provide in the market. Customers will be allocated such that:

1. The most expensive insurer will receive no customers in that segment
2. If you are ranked *nth* in the market in ascending order of price, you will receive (1 - *n*/40*)* of the customers (i.e. The 10th cheapest out of 40 insurers wins (1 - 10/40) = 3/4 of customers with that profile).

Your success is measured by the number of customers you have in your portfolio, but potential investors will be scared off if your portfolio is not profitable.

Claims data will be allocated to your customers, and based on your loss ratio (Claim cost / Premiums earned) of your portfolio, a weighting will be taken off your total customer score.

Loss Ratio	Multiplier
0 ~ 100%	1
100% ~ 120%	0.8
120% ~ 150%	0.5
150% +	0.3

(Premium earned refers to the portion of an insurance premium that applies to the expired portion of a policy. It represents the amount of premium the insurer has "earned" by providing coverage for the specified time period. As the policy period progresses, the insurer earns the premium gradually. For example, imagine a customer pays $1,200 upfront for a 12-month insurance policy. This premium covers the entire year of protection. After 6 months, half of the policy term has passed, so the insurer has "earned" half of the premium. In this case, the earned premium after 6 months would be $600. The remaining $600 is considered unearned premium because the insurer hasn't provided coverage for the rest of the year yet.)

Data Availability

To kickstart your pricing model, you’ve invested part of your life savings into purchasing claims data from existing insurers. This data will help you understand the underlying claims patterns and guide your pricing strategy. Both the insurance data (subject to confidentiality agreement) and ABS datasets can be downloaded via Moodle.

UNSW_claim_data:

Each row in the claims data represents a single claim, detailing the treatment start date, condition category, coverage status, and the associated financial amounts (paid and total claim).

Column	Description
claim_start_date	Date when the treatment related to the claim started.
claim_status	Status of the claim, whether it is fully covered, partially covered, or covered with exclusions.

condition_category	The category of the condition being claimed (e.g., Behavioural Issues).
claim_id	Unique identifier for each claim.
tenure	Number of calendar months between the inception of the policy and the claim.
vet_id	Identifier for the veterinarian associated with the claim.
account_id	Identifier for the account (policyholder).
exposure_id	Identifier for the exposure (specific insured risk under the policy).
claim_paid	The amount that has been paid out for the claim.
total_claim_amount	The total amount of the claim, including any amounts not covered by the policy.

UNSW_earned_data:

This file includes details about the policy tenure for each exposure or insured risk. Each row represents a specific exposure with data during that month and the corresponding exposure units.

Column	**Description**
UW_Date	Month that the exposure's policy was active from.
exposure_id	Identifier for the exposure (specific insured risk under the policy).
tenure	Number of calendar months between the inception of the policy and the current policy period
earned_units	Exposure units related to the covered risk over the time period.

This dataset also contains detailed information about the insured pets, their owners, and the corresponding policy details. It includes factors such as the pet's age, gender, source, and desexing status, as well as owner information and policy-specific data. This data can be used to understand the factors affecting the insurance risk and the likelihood of claims.

Column	**Description**
pet_gender	Gender of the pet (e.g., male, female).
pet_dob	Date of birth of the pet.

pet_de_sexed	Whether the pet is desexed (True/False).
pet_de_sexed_age	Age at which the pet was desexed (if applicable).
pet_source	Source of the pet (e.g., breeder, rescue).
pet_is_switcher	Indicates if the pet's policy was switched from another insurer.
nb_policy_first_inception_date	Date when the insurance policy was first incepted.
pet_age_months	Pet’s age in months at time of policy inception.
nb_contribution	This is the percentage amount of any valid claim the policy covers after the excess is applied.
nb_excess	Excess is the fixed amount paid as part of a customer’s contribution to a claim. Excesses apply once per condition for the same specific cause.
nb_address_type_adj	Type of address (e.g. House, Apartment).
nb_street	Street address of the policyholder.
nb_suburb	Suburb of the policyholder.
nb_postcode	Postcode of the policyholder’s address.
nb_state	State of the policyholder’s address (e.g., NSW, QLD).
person_dob	Date of birth of the policyholder.
nb_contribution_excess	Contribution % and excess of the policy.
pet_age_years	Pet’s age in years.
owner_age_years	Age of the policyholder in years.
nb_number_of_breeds	Number of breeds for multi-breed pets.
nb_average_breed_size	Average size of the pet’s breed(s).
nb_breed_type	Whether the pet is a: - Pure breed - Designer breed (New breeds) - Cross (Pet is mixed with multiple pure breeds known to the customer) - Unnamed Cross (Pet is mixed with multiple pure breeds not known to the customer)
nb_breed_trait	Genetic makeup of this dog’s breed, if the dog has a clear lineage. For example Golden retrievers, labradors

	historically share the same lineage and are descended from retriever type dogs. Pugs and bull dogs share the same lineage from brachycephalic dogs. Crosses have less clear lineages, and their traits are generally mixed.
nb_breed_name_unique	Customers can select (between 1~8) pet breeds for their pet, based on existing pet breeds. This field captures the breed of the very first selection.
nb_breed_name_unique_concat	Customers can select (between 1~8) pet breeds for their pet, based on existing pet breeds. If multiple breeds are selected, the names of each breed is concatenated.
is_multi_pet_plan	Whether the policy covers multiple pets.
lead_date_day	The day when the lead (potential customer) was generated.
quote_date	Date when the insurance quote was generated.
quote_time_group	Time of day when the quote was generated (e.g., Morning, Evening).
account_id	Unique identifier for the account (policyholder).
exposure_id	Unique identifier for the exposure (insured risk under the policy).

External datasets:

You are provided with the following external datasets downloaded from the Australian Bureau of Statistics (ABS) websites for your analysis (the dataset files are available on Moodle). You may also consider other external datasets that may be relevant or useful.

Postcode_SA_mapping:

This dataset provides a mapping key that links postcode-level data to different statistical area (SA) levels, facilitating geographic analysis across various regions.

ABS_SA_data:

Each dataset contains statistics mapped to different levels of statistical areas (SA).

- **Economy and Industry**: Monitors business activity, industry-specific employment, and economic performance, including data on business turnover, entries, exits, and real estate transactions.
- **Education & Employment**: Tracks educational qualifications and workforce participation across sectors, offering insights into skill levels and employment patterns.
- **Family & Community**: Provides statistics on family structures, community engagement, and household compositions, highlighting social dynamics and trends.
- **Income**: Examines income distribution across regions and industries, shedding light on economic conditions and disparities within the population.
- **Land and Environment**: Covers land use, agricultural production, and environmental resources, focusing on sustainability and natural resource management.
- **Persons Born Overseas**: Reports on migration patterns, population demographics, and the influence of the overseas-born population on the economy and community integration.