Predicting Pet Insurance Claims

Building a model to predict claims in the second policy year

EXECUTIVE SUMMARY

In 2020, the global pet insurance market is estimated to exceed 4 billion dollars [1]. And in the US alone, the market value is anticipated to be close to half this total amount ($1.6 Billion USD) with sustained year-over-year growth of nearly 15% for the foreseeable future [2]. With so much potential revenue at stake, the need for competitive policy pricing is as important as ever.

The idea behind pet insurance is simple and similar to the human health insurance market. When a pet insurance policy holder incurs veterinary expenses related to their enrolled pet, they can submit claims for reimbursement, and the insurance company reimburses eligible expenses. Given the stochastic nature of claims submissions however, policy pricing remains a challenge.

To price insurance products correctly, the insurance company needs to have a good idea of the amount their policyholders are likely to claim in future policy years. The goal of this project is to create a machine learning model to predict how much (in dollars) a given policy holder will claim for during their second policy year based on pet information and claims data from their first policy year. After evaluating a range of models, the best performing model resulted in claims predictions that were $400 more accurate than predictions made using a simple baseline model.

DATASETS

The underlying source data for the project consists of two files - *PetData.csv* and *ClaimData.csv* obtained from a large, national pet insurance provider. The PetData file contains data for 50000 unique pets who enrolled for policies during the 2018 calendar year. The pet data includes 8 features which provide information about the type of pet (e.g., species, breed, age) and the cost of the policy (i.e., premium and deductible). The claim data includes 4 features detailing insurance claims recorded over a 3-year period between 2018 and 2020 providing the claim date and amount. The two datasets are linked by a common feature, PetId, which can be used to understand the claims totals for each individual pet. Prediction.

DATA WRANGLING

Overall, the two datasets were relatively clean and the bulk of the data wrangling process consisted of data verification and determining how best to combine the pets data with the associated claims data. A few columns required some additional manipulation in preparation for exploratory data analysis.

The column presenting the greatest challenge was *Breed* due the large number of possible values and the high variability in terms of the number of pets in our data for each breed. Popular pet breeds (e.g., Labrador Retriever) would often cover 1000 or more pets in the data, while certain less common breeds (e.g., Selkirk Rex) were represented by a single pet.

Key Observations:

- **Pet Count** - Verified 50,000 unique pets (based on PetIds)
- **Species** - Data consists of two species of pets, cats and dogs (with dogs outnumbering cats 5 to 1)
- **Breed** - Observed 373 unique breeds in total (55 cat and 318 dog)
- **Age** - Pet ages range between 0 and 13
- **Premium** - Premiums fall into a wide range with a few outlier values close to $1000
- **Deductible** - Deductibles are fairly well distributed and appear to be stratified across a range of common values
- **Median Claims** - For cats and dogs, the median value for total number of claims and total amount of claims is 0
- **Outlier Claims** - Both species have some significant outliers in both categories (number and amount of claims)

DATA ANALYSIS

During exploratory data analysis, a number of observations stood out in relation to overall claims totals. In general, dog owners have more claims and higher total claims amounts than cat owners. As expected, this tends to translate to higher premiums.

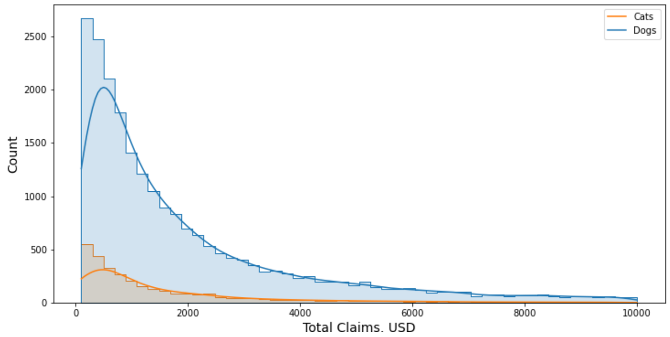

Figure 1. Dogs tend to have higher claims totals on average

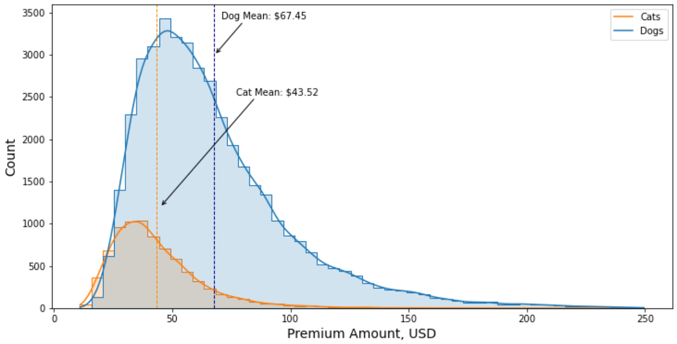

Figure 2. Dogs tend to have higher premiums on average

Looking at the data from the perspective of pet breeds, we see that as the average number of claims for a breed goes up, the average total claims amount goes up in almost a linear fashion. That said, one limitation of the data is that there is a significant imbalance in terms of the number of pets per breed. Generally, as the number of pets in a breed increases, the variability in claims (number and amount) goes down, moving the breed closer to the overall linear trend line.

Unsurprisingly, cat breeds tend to occupy the lower left quadrant of the plot indicating fewer claims and lower total claims on average when compared to dogs.

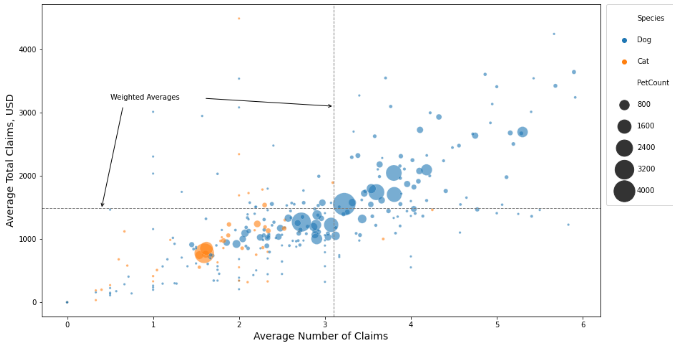

Figure 3. A linear trend line is visible in average number and amount of claims by species as the number of pets in the breed increases.

MODELING RESULTS

In this project, we evaluated data for 50,000 pet insurance customers with a goal of building a model to predict the total claims amount in the second policy year based on basic info about the pet (breed, age, species, etc.) and claims data (number of claims, amount of each claim, etc.) for the first policy year. We evaluated a variety of different models (listed below) and in the end, found the best performance using a Gradient Boosting Regressor with some parameter tuning.

Using our best model on our test data, we achieved a mean absolute error of roughly 638, well inline with our cross-validation scores from model tuning. While still somewhat high, this is down from an error slightly over 1020 using our baseline model of a simple dummy regressor. This represents an **improved accuracy of nearly $400 per customer**.

When considered in the context of a customer base of 65-75 million, these results are significant and could lead to substantial savings for the business in terms of improved pricing and risk models.

Model Evaluation Results:

Regressor	**Score (MAE)**
Dummy (Using Mean)	1020.41
Linear (Default)	930.46
Linear (with feature selection)	930.40
Lasso (Default)	930.28
Lasso (with simple tuning)	930.18
Gradient Boosting (Default)	932.25
Gradient Boosting (with n_iter and learning_rate tuning)	931.92
Gradient Boosting (with loss and alpha tuning)	677.75
Gradient Boosting (with tree-specific tuning)	673.82

Results with Test Data:

Regressor	**Score (MAE)**
Gradient Boosting (Using best estimator)	638.54

FUTURE RESEARCH

Although we observed a dramatic improvement in model performance using a gradient boosting regressor, our mean absolute error still leaves quite a bit of room for improvement.

The following are recommendations for potential next steps to refine or further expand upon the work done in this project:

- **Obtain a more balanced dataset** - The dataset for this project is highly imbalanced across a number of features (species, breed, age, etc.). While this imbalance likely reflects the population of pets, it presents challenges when building a model that predicts the claims amount for a specific pet. By starting with a more balanced dataset, it's possible the predictive accuracy could be improved overall.
- **Engineer additional features** - Feature engineering in this project was largely focused on relating pet age and breed to claims data. It's possible that additional feature engineering could be done to improve model performance. Suggestions include:
    - **Timing of claims** - As part of data wrangling, we rolled up our claims data into totals and averages per pet. But it stands to reason that the timing of when claims are submitted could be a powerful predictor of claims amounts in the second policy year. For example, a pet with $10,000 in claims in the first 3 months of year 1 may be less likely to have claims in year 2 when compared with a pet having an equal amount of claims in the last month of year 1.
    - **Additional Breed Data** - It is widely known that different pet breeds have different characteristics, but our limited dataset did not include any data specific to each breed. By including additional breed specific data in our analysis, it may be possible to engineer meaningful features to improve the predictive power of our model. Examples of this could be including an average weight or average lifespan per breed or engineering a feature that calculates the *risk index* for a pet given their age, breed and species.

REFERENCES

1. Pet Insurance Market Size, Share & Growth | Industry Report 2019-2028
2. https://www.ibisworld.com/industry-statistics/market-size/pet-insurance-united-states/