Pet Insurance Customer Segmentation

Clustering pet insurance customers to drive marketing results

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.

The Marketing department at a leading pet insurance provider is seeking to better understand its customer base to prevent customer loss and drive additional company revenue. Data suggests that customers are most likely to cancel their policies around the 2-year point following enrollment. The goal of this project is to identify 3-4 customer segments (with some justification for each) that will enable the Marketing team to reduce customer shrinkage while improving outcomes related to targeted ads and/or direct-to-customer campaigns.

After evaluating different clustering methods, the best performing model resulted in 4 distinct clusters in the data representing 4 distinct customer segments. Two predominate driving factors were observed in the clustering results - the first being the number of policy years with claims (either 0, 1, or 2) and the second based on which specific policy years the customer claims take place (in the event of only having claims in a single year).

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.

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. And unsurprisingly, dogs also are much more likely to have claims in both of the first two policy years.

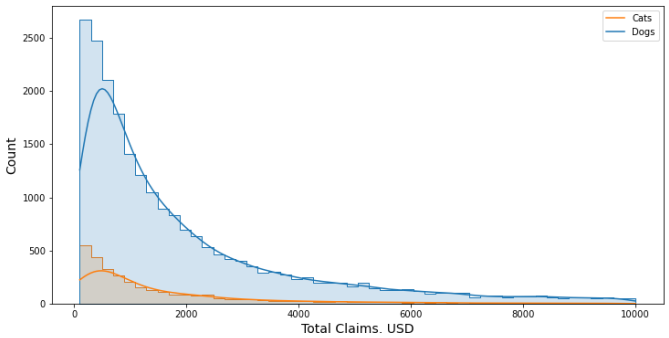

Figure 1. Dogs tend to have higher claims totals on average

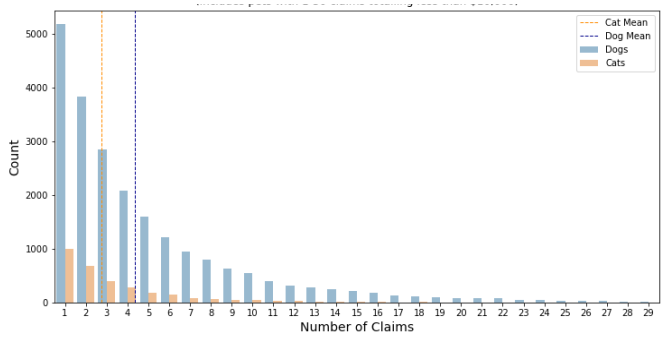

Figure 2. Dogs tend to have a higher number of claims over the first two policy years

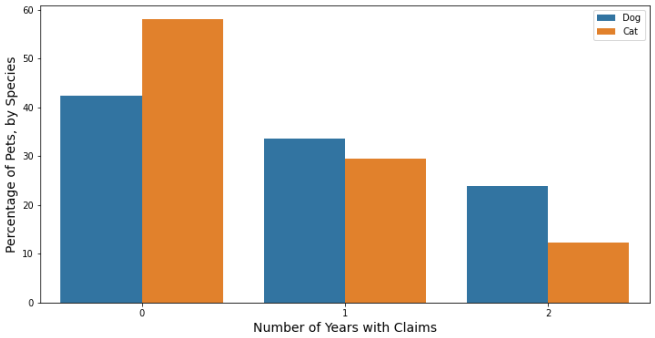

Figure 3. Cats are more likely to have no claims and Dogs are almost twice as likely to have claims in both policy years

Principal Components Analysis (PCA):

Exploratory data analysis was helpful in providing better context for the relationships in our dataset, but was inconclusive in terms of identifying clear customer groupings. PCA was utilized to better understand which features contribute the most to the variance in our data as a step toward identifying meaningful clusters of customers.

The result of PCA shows that nearly 85% of the variance in the customer data can be explained by looking at the first 3 PCA components.

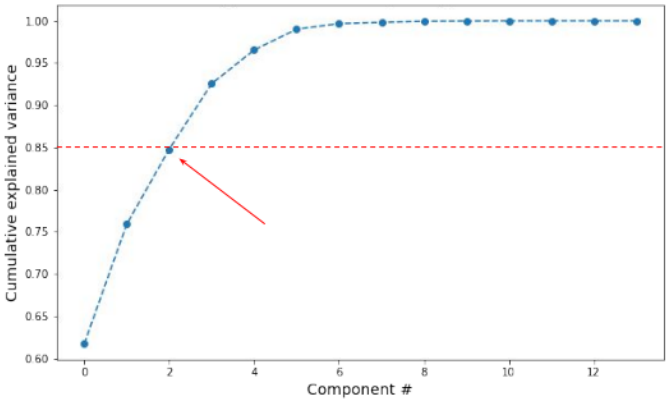

Figure 4. PCA result illustrating that nearly 85% of the variance is explained in the first 3 principle components

Based on the results above, we analyzed the first 3 components from PCA to identify which features contributed most to each component. Figure 5 shows the relative importance of each feature for each component.

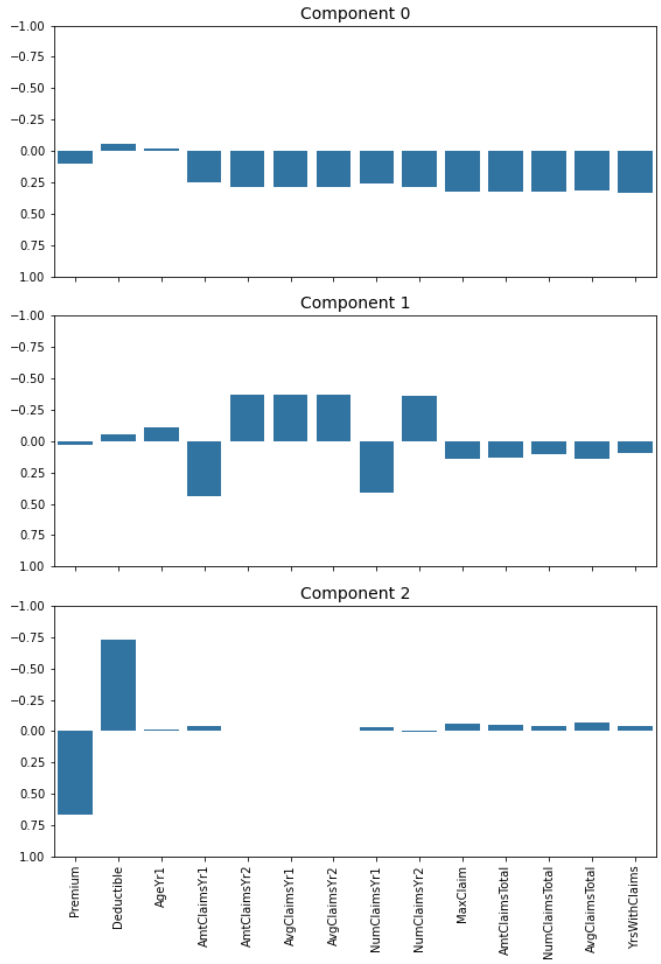

Figure 5. Relative feature importance for each of the first 3 PCA components

We observe that for component 0, all claims-related features have similar importance including features related to claims totals, claims in year 1 and claims in year 2. For component 1, we observe that claims-related

features are once again highest in terms of feature importance. However, for component 1, the claims-related features of highest importance are those focused on individual policy years only (i.e, Year 1 or Year 2) with much less contribution from the features related to claims totals for both policy years.

Finally, Component 2 is most heavily influenced by Premium and Deductible, with very little contribution from the other features in our dataset.

CLUSTERING RESULTS

To arrive at our final customer segmentation result, we evaluated two clustering algorithms - KMeans and DBSCAN. While both performed well in terms of arriving at distinct customer segments, we observed the best result by utilizing DBSCAN. For the purposes of this report, we will focus on the result from DBSCAN only. The full analysis is available upon request.

DBSCAN:

After incorporating our top 3 components from PCA into our final dataset and implementing DBSCAN, we arrived at 4 distinct customer segments. These segments can be observed in a scatterplot of component 0 vs. component 1 in figure 6.

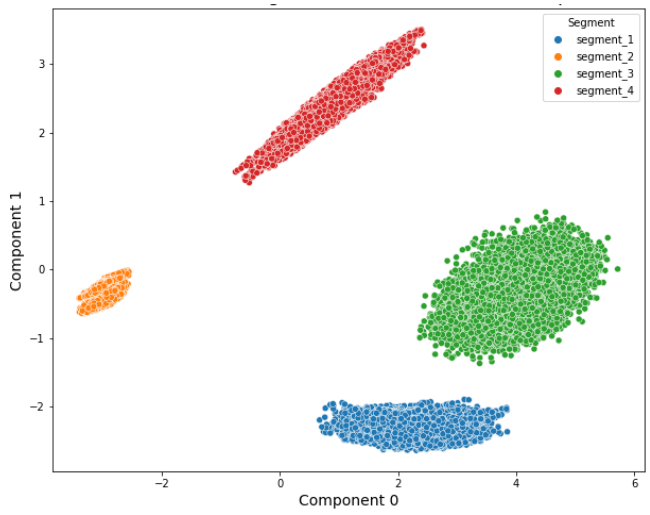

Figure 6. Customer segments identified using DBSCAN

Based on the clustering results obtained above, further analysis was performed to understand what is differentiating each customer segment from one another. The result of this analysis, displayed in figure 7, shows that the most critical factor in determining customer segments is the number of years with claims over the first two policy years.

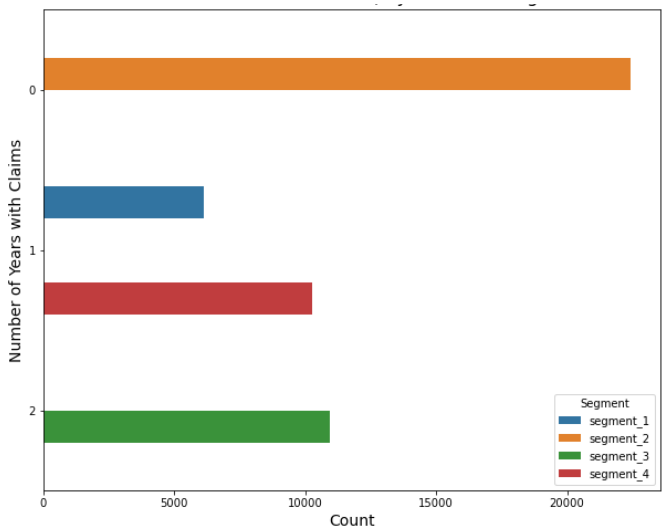

Figure 7. Number of years with claims by customer segment

The result above provides a clear basis for differentiating customers in segments 2 and 3. Segment 2 includes all customers with zero claims in the first 2 policy years and Segment 3 includes customers with claims in both of the first two policy years.

Segments 1 and 4 appear to be similar in the plot above as both have claims in only one of the first two policy years. Additional analysis indicated that the differentiating factor between segments 1 and 4 was the year in which the claims occurred (i.e., policy year 1 or policy year 2). This is clear when comparing the counts of the number of claims for years one and two for customer segments 1 and 4 (as in figures 8 and 9 below).

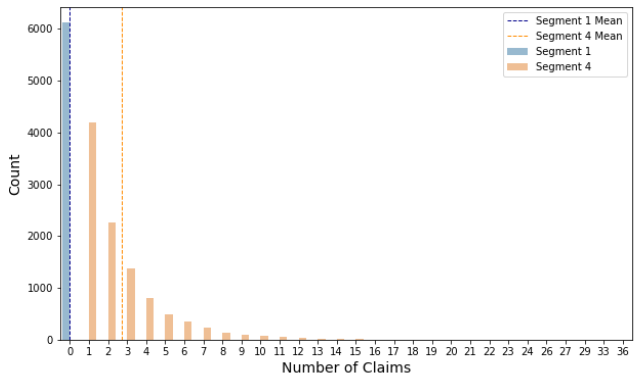

Figure 8. Count of claims in policy year one for customers in segments 1 and 4

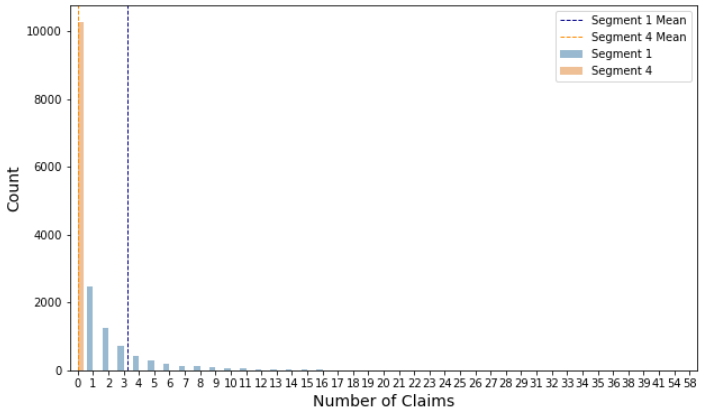

Figure 9. Count of claims in policy year two for customers in segments 1 and 4

Customer Segments:

Following the segment analysis above, we arrive at the following customer segments and associated descriptions.

Customer Segment	Claims in Year 1	Claims in Year 2
Segment 1	No	Yes
Segment 2	No	No
Segment 3	Yes	Yes
Segment 4	Yes	No

Table 1. Claims in policy year 1 and policy year 2 by customer segment

FUTURE RESEARCH

Although we observed a nice distinct clustering result, there may still be room for improvement depending on specific marketing goals and/or customer campaigns.

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

- **Obtain additional customer data** - The dataset for this project relies heavily on data related to customer claims. While this is helpful in identifying customer segments based on claims data, it's possible additional data could lead to a different clustering outcome and present new and interesting opportunities for marketing to existing customers or reducing customer churn.
- **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/