Predicting Pet Food Prices Using Text-based Product Descriptions

Team Tibbles & Bits: David Parks, Rosella Li, Lizhou Wang, Simran Kathuria
Client: Colgate Palmolive (Dr. Kli Pappas)
December 9, 2020

Executive Summary:

This capstone project's main objective is to predict the sales price of pet food products using their descriptions for our client Colgate Palmolive. Under the guidance of Dr. Kli Pappas, the team successfully identified the most influential text-based features utilizing the combination of Term Frequency - Inverse Document Frequency (TF-IDF) feature extraction and LASSO regression. The model shows that Colgate's pet food descriptions tend to indicate a lower price relative to the actual price Colgate charges for their pet food products due to vocabulary that is less trendy and more generally descriptive. Implementing our model would help Colgate take the lead in pet food product labeling by getting ahead of future trends while right-pricing future products.

Additionally, the team was asked to create a grouping model to discover how new proposed pet food descriptions fit among existing descriptions. The team accomplished this task with the open-source Doc2Vec document encoding library. New descriptions are compared with existing descriptions using cosine similarity to determine how closely related they are. This information is incredibly insightful for our client and helps them maintain a competitive edge while ensuring that any proposed pet food descriptions stay on-brand.

Problem Statement

This capstone project's main objective is to predict the sales price of new pet food products and provide related insights into the drivers of their price using the products' descriptions and other text-based attributes. Under the guidance of Dr. Kli Pappas, our client Colgate Palmolive expects the team to deliver a detailed analysis of the dataset, including an outline of the analysis strategy, a predictive model with the appropriate validation metrics, and a grouping algorithm pinpointing Colgate's fit in the existing market. This information will help Colgate identify deficiencies in their pet food product labeling while ensuring that future products stay on-brand.

Ethical Considerations

This project implements two separate models. The first model takes the text-based product descriptions, numerically vectorizes them, and regresses the resulting numerical matrix against a price vector to establish the relationship between a product's description and its price. The second model groups similar product descriptions using cosine similarity to ascertain summary statistics for the group. In terms of ethical considerations, each model poses distinct but potentially harmful consequences.

The first model relies on the assumption that a product's description is the primary determinant of its price. According to our client, pet food descriptions are typically drafted after the packaging has been designed (Dr. Pappas, via Zoom, Oct 12, 2020). The client asks the author to describe what comes to mind when they first see the packaging. This means that the product description is typically a subtle, text-based explanation of the graphics printed on the packaging. For example, Hill's Pet Nutrition, Colgate's pet food subsidiary, uses portraits of animals superimposed on a plain white background for their customer-facing packaging.

Conversely, Blue Buffalo, one of Hill's competitors, uses elaborate graphics on all sides of the product. It follows that Hill's product descriptions tend to be very matter-of-fact. In contrast, Blue Buffalo's descriptions typically involve an elaborate narrative written to invoke the customer's emotional attachment to their animal. As one may have guessed, Blue Buffalo's products sell at a higher price point. Given this situation, it is important to note that the predictive pricing model does not account for a product's visual appeal. Anyone using this model must acknowledge this limitation.

For the grouping model, one can imagine a nefarious actor applying this technique on targeted individuals to see if their writings match those of similar groups of interest. Take, for example, a government agency trying to identify political dissenters or protestors. The government could compare the social media posts of known opposition leaders with other similar posts to determine potential dissidents who could be preemptively placed on watchlists and possibly threatened, arrested, or killed based on their social media posts.

Although this project employs relatively simple techniques, they have wide-ranging ethical implications for both the client and society. However, if these models are applied as intended by the client, ethical consideration should be minimal.

Literature Review

To better understand consumer behaviors associated with pet food product purchasing, the team reviewed current literature concerning consumers' qualitative preferences for pet food product labeling. The search focused on analyzing consumer surveys that specifically targeted pet owners in North America and Europe. This information adds context and relevance to the project because it provides insight into the consumer's decision-making process, which is not captured in the dataset Colgate provided.

As early as 2011, author Timothy Bohrer reported that many of the same trends observed in human nutrition started to influence food labeling for dogs and cats. At that time, pet food manufacturers were beginning to experiment with phrases such as "organic," "locally and sustainably sourced ingredients," "grass fed and free range," "free of grains and fillers," "preservative and additive free," and "fresh" on their labels.[1] Specialty introductions included variations on themes evident in the consumer segment termed LOHAS (lifestyles of health and sustainability), made up of consumers committed to health, quality, environmental impact, and sustainability in their product choices.[2] Given that these trends have accelerated over the past decade, it is unsurprising that more recent research supports an increasing humanization of the pet food industry.

According to Schleicher et al, the growing trend of humanization and anthropomorphism of pets has spurred strong marketing messages and ingredient claims by pet food producers and conflicting claims of what is the best food for pets by consumers.[3] The authors found that pet owners with the highest anthropomorphism scores placed the most importance on health and nutrition, quality, freshness, and taste of pet food.[4] In fact, most pet owners surveyed in this study reported giving equal or more priority to buying healthy food for their pets than for themselves.[5] The article also states that consumers prefer lower-priced pet food, but value natural and organic ingredients ahead of price.[6] While it is encouraging that pet owners are trying to feed their pets the best nutrition possible, the results from this study suggest that pet owners who humanize their four-legged friends may believe some marketing strategies (e.g., gluten-free, grain-free, raw, holistic) which could result in feeding practices for which there are no scientific studies showing any health benefits to pets.[7]

A recent survey of pet owners in Italy agrees with the overall trends identified by Schleicher et al. Over the past decade, pet owners have increasingly demonstrated sensitivity toward their

[1] Bohrer, Timothy. "Pet food packaging: Evolution, revolution & innovation." *Paper, Film and Foil Converter* (2011): 2.

[2] Ibid, 1.

[3] Schleicher, Molly, Sean B. Cash, and Lisa M. Freeman. "Determinants of pet food purchasing decisions." *The Canadian Veterinary Journal* 60.6 (2019): 644.

[4] Ibid, 645.

[5] Ibid, 647.

[6] Ibid, 644.

[7] Ibid, 648.

companion animals, including an increase in the attention paid towards their nutrition.[8] According to the survey administered by Vinassa et al, the presence of natural ingredients was considered to be the most important indicator of pet food quality, whereas "characterized by a high price" was considered least important.[9] The authors of this study also discovered a demographic divide in consumer preferences split by age and education level. Respondents older than 65 attributed the least importance to the use of recyclable packaging and the presence of cruelty-free claims. They also attributed relatively little importance to label comprehension, location of the pet food production facilities, and the presence of natural ingredients. On the other hand, respondents under 35 placed most relevance on a high percentage of proteins in the pet food and to the presence of recyclable packaging. The cruelty-free and the grain-free claims received on average higher scores for the population aged 35 to 50.[10] Pet owners with a college degree gave positive scores to label comprehension, the location of production, the presence of natural ingredients, and grain-free and cruelty-free claims, whereas those without a college degree found them less relevant.[11]

Overall, these studies provide supporting evidence for the claim that consumer preferences for pet food labeling closely resemble their own. Interestingly, the data provided by Colgate appear to support the positive, qualitative relationship between today's human dietary trends and pet food pricing. Quantifying this relationship will help Colgate take the lead in pet food labeling by getting ahead of future trends while right-pricing future products.

Project Criteria

This project's primary goal is to build a predictive pricing model that best explains how specific words in the product's description relate to the overall price of the product. Thus explainability of the model is of the utmost importance. The second goal of this project is to create a model where the client can input a proposed product description and receive an output that groups the proposed description with existing descriptions to approximate where the new product fits with existing products. In this case, ease of use and simplicity are the desired criteria.

During the ideation phase, the team focused on regression models that best explain the correlations efficiently, rather than complex models that might be more accurate but are harder to interpret. We defined explainability to mean that the resulting model should, as clearly as possible, convey a direct connection between its coefficients and the predicted prices, and we chose root mean squared error as the primary performance metric so that the errors are on the same scale as the target variable.

For the second model, we defined ease of use to mean that the model should accept a text-based product description provided by the client and return output without the client needing to transform the input manually. We defined simplicity to mean that the model should take advantage of

[8] Vinassa, Marica, et al. "Profiling Italian cat and dog owners' perceptions of pet food quality traits." *BMC Veterinary Research* 16 (2020): 1.

[9] Ibid, 8.

[10] Ibid, 3.

[11] Ibid, 4.

established text vectorization packages. Cosine Similarity was selected as the metric to determine how well proposed product descriptions match existing descriptions.

Selected Solutions

To transform the pet food descriptions into a regressible numeric matrix, Term Frequency - Inverse Document Frequency (TF-IDF) was applied to the dataset. TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection, where the value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general.[12] This task was accomplished using Scikit-Learn's TfidfVectorizer feature extraction module.[13] After the initial transformation, the team manually removed any remaining stopwords not found in the default NLTK stopwords corpus[14] (see the Obstacles section for more discussion about implementing TF-IDF).

Next, the team engineered four categorical variables to capture some of the intrinsic features in the data. Categories were created for pet type (dog, cat), meal type (primary, treats), food type (wet, dry), and first word in the ingredients list (chicken, wheat, beef, corn, other). One numerical variable, ratio, was created by dividing the total package size by the individual unit size. Lastly, a Box-Cox transformation was performed on the original prices to ensure that the target variables were normally distributed.[15]

Using the transformed dataset, the team compared 19 different regression models using PyCaret, which is an open source, low-code Python wrapper around multiple existing machine learning libraries such as Scikit-Learn, XGBoost, and more.[16] Using the root mean squared error as the primary metric, this initial test showed that simple linear regression performed as well as or better than some more complicated regressors. Based on this result, the team instantiated, trained, and tested separate Linear, Ridge, LASSO, ElasticNet, and Random Forest regressors.

For the grouping model, the team utilized Doc2Vec to encode the pet food descriptions into numerical vectors. The team selected Doc2Vec because once the model is trained, the encodings remain constant; whereas with TF-IDF, the TF-IDF matrix must be recomputed every time a new description is added to the existing corpus. Using Doc2Vec also ensures that the values in the existing vectors remain constant when cosine similarity scores are calculated. This task was accomplished with Gensim's Doc2Vec module.[17]

[12] https://en.wikipedia.org/wiki/Tf-idf

[13] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html?highlight=tfidf#sklearn.feature_extraction.text.TfidfVectorizer

[14] https://www.nltk.org/

[15] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.boxcox.html

[16] https://pycaret.org/about/

[17] https://radimrehurek.com/gensim/index.html

To transform the pet food descriptions into static numerical vectors, the descriptions were tokenized into lists of words. Next, a vocabulary list was built using these lists of words and the model was trained using the entire corpus of pet food descriptions. Each word had to appear at least five times throughout the entire corpus to be considered by the model. The model was trained for 40 epochs and produced length 50 numerical vectors as recommended by the user guide.[18] The resulting vectors were compared with new draft descriptions using cosine similarity.

Results - Price Prediction Model

The hyperparameters of each selected regression model were optimized using a five-fold cross-validated grid search. After the initial grid search, the range of each hyperparameter was narrowed using the optimal parameters of the previous search as a guide. This process was repeated a third time, after which hypertuning was stopped to prevent overfitting of the final model. Each final model was re-instantiated with the optimal hyperparameters, retrained using the full training dataset, and tested with a holdout dataset. The LASSO and ElasticNet regression produced comparable results with the lowest root mean squared errors on the holdout dataset (see Figure 1). Since LASSO is the simpler of the two models, LASSO regression was selected for the final predictive pricing model.

Predictive Pricing Model Optimization Summary

Model Type	Grid Search Hyperparameters	Optimal Hyperparameters	Root Mean Squared Error
Linear Regression	Degree: [1, ..., 20] Intercept: [True, False] Normalize: [True, False]	Degree: 1 Intercept: True Normalize: False	2.0412
Ridge Regression	Alpha: [10^{-6}, …, 10^{4}] Intercept: [True, False]	Alpha: 100 Intercept: True	2.0397
LASSO Regression	Alpha: [10^{-6}, …, 10^{4}] Intercept: [True, False]	Alpha: 0.006428 Intercept: True	2.0368
Elastic Net Regression	Alpha: [10^{-6}, …, 10^{4}] Intercept: [True, False] L1 Ratio: [0.01, …, 0.99]	Alpha: 0.01 Intercept: True L1 Ratio: 0.60	2.0367
Random Forest Regressor	Max depth: [10, …, 100] Max features: [auto, sqrt] Min samples leaf: [3, ..., 50] N_estimators: [100, ..., 1000]	Max depth: 45 Max features: auto Min samples leaf: 3 N_estimators: 700	2.9028

Figure 1

[18] https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html

To find the value that each word contributes to the product's overall price, a LASSO regression model was instantiated using the optimal hyperparameters given above and trained on the entire dataset. The coefficients were then sorted in descending order to determine the most influential words by value. The top 20 largest and bottom 20 smallest coefficients were isolated, the results of which appear in figure 2 below.

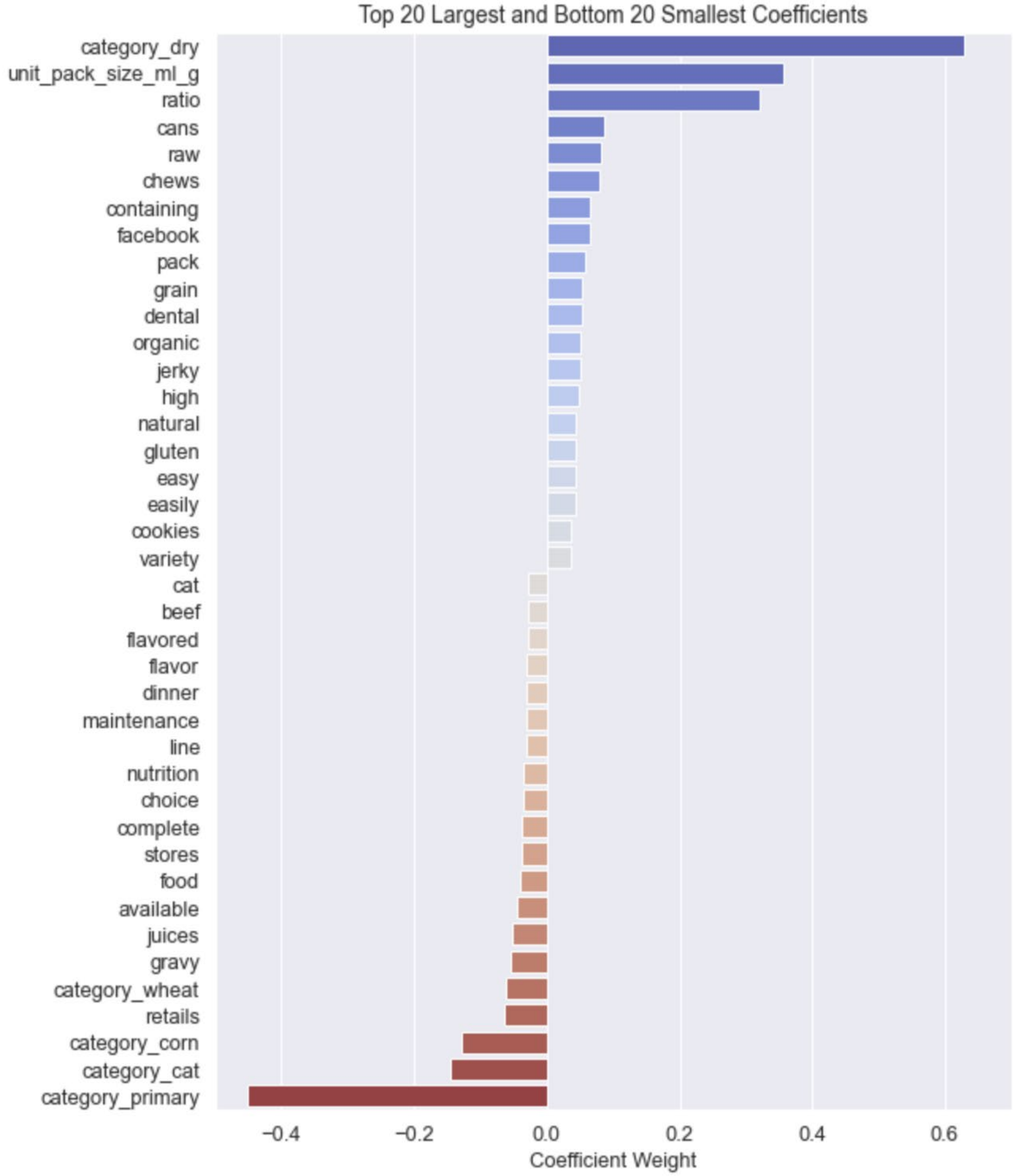

Figure 2

Starting with the top three and the bottom three coefficients, it is interesting to note that the engineered features had the highest and lowest weights, respectively. Dry foods are typically packaged in larger bags and boxes relative to wet foods in cans, so it seems reasonable that the dry category along with the unit size and ratio features have the largest positive influence on price (i.e. - paying more for more product). At the other end of the spectrum, foods in the primary category are discounted relative to the premium paid for treats. Cat foods and foods where corn is the first word in the ingredients list also incur a penalty.

Among the top 20 coefficients, one finds words such as "raw", "organic", and "natural," which are consistent with the preferences of pet owners found in the qualitative surveys discussed in the literature review. Words such as "gluten" and "grain" are also present. Based on the context in which these words typically appear in pet food product descriptions, one can assume that the absence of these ingredients is probably being promoted in these descriptions (i.e. - "gluten free" or "without grain"). The word "facebook" also appears as a positive coefficient and may be indicative of a relationship between a company's level of social media engagement and a consumer's willingness to pay more for a product to which they feel connected. However, additional research is needed to support this assessment.

As for the bottom 20 coefficients, one finds more generally descriptive words such as "complete", "maintenance", "nutrition", "dinner", "flavor", and "gravy" detract from the overall price and that having the word "wheat" as the first word in the ingredients list is the fifth lowest coefficient in the graph. It is also worth mentioning that cat foods in general (second lowest coefficient in the graph) and the mention of "cat" in particular both have a negative effect on price. The team assesses that the penalty for cat foods has more to do with the smaller unit size of cat foods relative to dog foods and recommends that this relationship be explored further to ensure that pet food manufacturers are not undervaluing these products.

To determine whether Colgate's pet food descriptions add value to their products, the team compared the predicted prices for Colgate's pet food products with their actual prices. Of the 131 products examined, the model overpredicted the price 63 times and underpredicted the price 68 times. However, summing the differences between the actual prices and the predicted prices produced a positive sum of 110.68, meaning that in aggregate, the prices predicted by the descriptions fall short of the actual prices. One interpretation of this result is that the consumer's perceived value of Colgate's pet food products comes from something other than the description. For example, the studies mentioned in the literature review found that consumers were willing to pay more for pet foods when they had been recommended by a veterinarian. Another possible interpretation is that Colgate's pet food descriptions "punch below their weight" because they include fewer trendier words, which have positive coefficients, and more generally descriptive words, which have negative coefficients. This last interpretation is supported by comparing the mean TF-IDF values for Colgate's pet food products with the mean TF-IDF values for all other brands.

The mean TF-IDF values for the words in the top 20 largest and bottom 20 smallest coefficients graphs (i.e. - excluding the categorical features) were computed for both Colgate and all other brands. Comparing the two outputs shows that other brands had higher mean TF-IDF values for 11 out of the 17 words with positive coefficients and lower mean TF-IDF values for 7 out of the 16 words with negative coefficients. This result indicates that Colgate's pet food descriptions do in fact have a tendency to include fewer trendier words with positive coefficients and about the same amount of generally descriptive words with negative coefficients. The team recommends that Colgate consider expanding its current pet food description vocabulary to accommodate this finding.

The full price prediction model was also tested on certain subsets of the data based on the following categories: pet type (dog, cat), meal type (primary, treats), food type (wet, dry), and brand. Of the subset types tested, the meal type treats had the highest root mean squared error, 2.1368, and the food type dry had the lowest root mean squared error, 1.9229. Of all the tested brands, the 10 brands with the lowest root mean squared error are featured in Figure 3 below. Based on these results, the team recommends that those employing this model in the future explore the idea of training and employing the model on subsets of the data to achieve better accuracy.

Brands with the Lowest Root Mean Squared Error	
Brands	Root Mean Squared Error
Dad's Special Mix	1.0001
Nylabone Edibles	1.0014
Sophistacat	1.0018
Gourmet Flavor Dog Biscuits	1.0026
Cloud Star Muttos	1.0030
Rocky Mountain Chocolate Factory	1.0039
Purina One Smartblend	1.0051
Excel	1.0057
Snausages Fortune Snookies	1.0059
PetSafe Busy Buddy	1.0059

Figure 3

Results - Grouping Model

For the grouping model, the team instantiated and trained a Doc2Vec model on the full corpus of pet food product descriptions. The model was trained for 40 epochs and produced a 7525 x 50 size matrix where each row is a vectorized numerical representation of each description. The resulting vectors were compared with a new fictional pet food product description (see below) using cosine similarity to demonstrate the models functionality and determine where the new description fits within the existing corpus.

Fictional Test Description

> *Tibbles & Bits brand organic dog food is naturally gluten and grain-free, and provides your pet with the best variety of raw ingredients. Follow us on Facebook to see all of our cookies, chews, and jerky.*

The results of the test show that the fictional description most closely matches Exclusively Pet's Exclusively Dog Cookies Best Buddy Bones Cheese Flavor Training Treats with a cosine similarity score of 0.57. The product description and the resulting summary statistics for the top 10 closest matching products are featured below.

> *Exclusively Dog Cookies Best Buddy Bones Cheese Flavor Training Treats are made with natural, kosher ingredients and are said to be a healthy alternative to traditional dog treats. The treats are free of animal parts, by-products and fillers and can be used as a treat, reward or snack. This product has been kosher certified and retails in a 5.5-oz. resealable package. The following flavors are also available: Chicken; Peanut Butter; and Beef & Liver.*

Summary Statistics for the Top 10 Closest Matching Products	
Mean Price	$7.02
Median Price	$7.24
Std Dev	$4.87
Min Price	$0.62
Max Price	$12.00

Figure 4

The client can use these results to see the vocabulary of the closest matching description and determine if the desired price point is in the range of the summary statistics.

In earlier iterations of the project, the team attempted to build a grouping model by first applying Principal Component Analysis (PCA) to the transformed dataset to reduce it to two dimensions. A plot of the two principal components did not reveal compellingly distinct clusters (see Figure 5 below). The team also attempted K-means clustering, but the results were even less compelling. These efforts were ultimately abandoned.

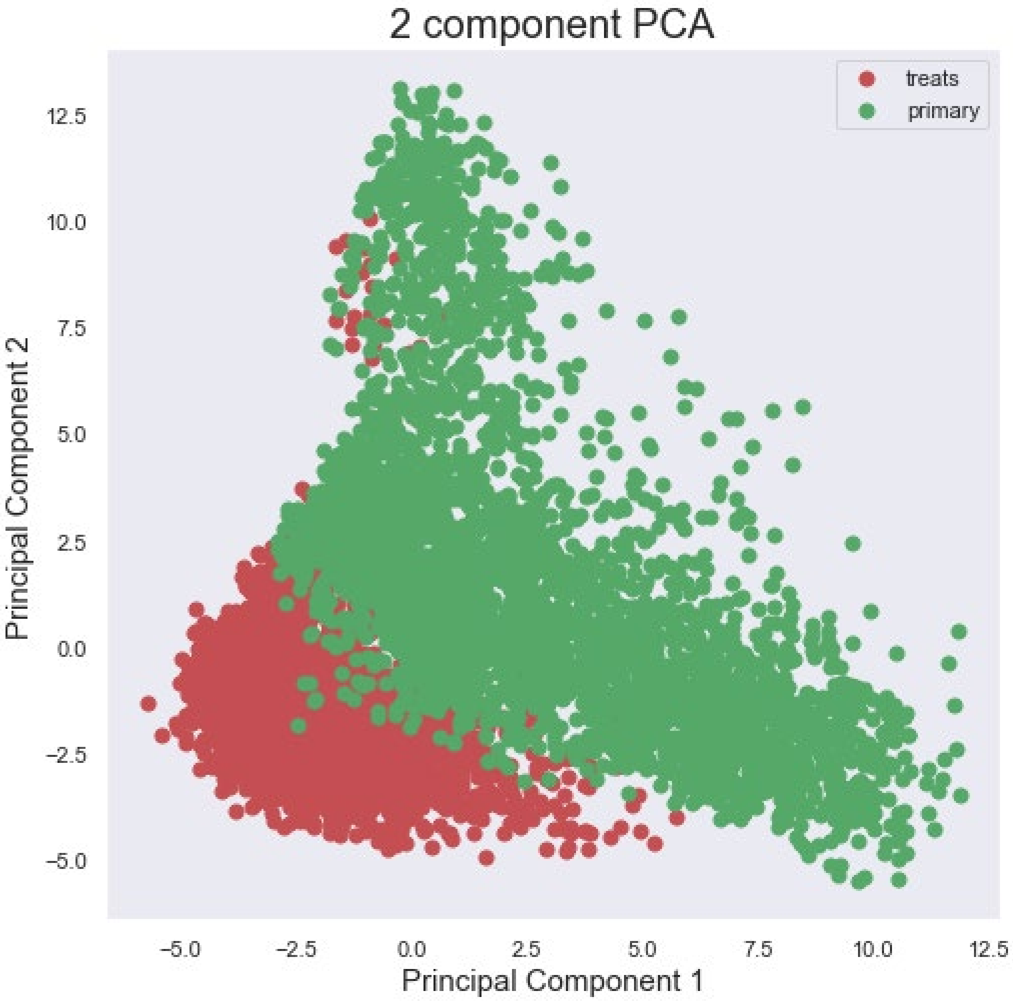

Figure 5

Obstacles

The biggest challenge of this project was overcoming instability issues in the feature matrix generated by the TF-IDF feature extraction method. The unconstrained transformation created a sparse matrix with 7,822 rows and 7,367 columns. Dividing the data into training and testing sets using an 80/20 ratio created a sparse, underdetermined training matrix with 6,257 rows and 7,381 columns (11 categorical and 3 numerical variables were added to the final dataframe). Using this training matrix, the regression models produced root mean squared errors of magnitude 10^9 or larger, which were grossly unacceptable given that the largest price in the dataset is $32.99 USD.

Because the client specified that they wanted to know the value of each coefficient in the model (i.e. - how much value a word adds to the overall price), the team was unable to implement traditional dimension reduction methods. Instead, words that appeared only once throughout the entire corpus of pet food descriptions were removed from the TF-IDF feature matrix. This resulted in a reduction of over 2,400 words, and hence columns, from the training data. Using the resulting training matrix in the regression models reduced the root mean squared errors to a magnitude of 10^6, which were still unacceptable.

Noting that there is a *max_features* hyperparameter within Scikit-Learn's TfidfVectorizer that retains the top *max_features*[19], the team created a search function to determine the "breaking point" for the regression models as the number of input columns increased. Running this function, the team found that any TF-IDF feature matrix with *max_features >= 409* caused major instability. Therefore, the TF-IDF feature matrix was limited to 400 maximum features. This change resulted in root mean squared errors that were less than 10^1 in magnitude.

Roles

Our team comprises four members: David Parks, Rosella Li, Simran Kathuria, and Lizhou Wang. All four members used their strengths and overcame their weaknesses to contribute to this capstone project productively and efficiently. Each member had the unique opportunity to take on leadership roles, whether through presenting, meeting with the client, or writing coding for the project.

David took on the project manager's role. With his prior real-world experience, he had a great understanding of how to create a collaborative work environment where the workload was appropriately distributed. David took charge of cleaning the dataset, visualizing several plots of the models, kept track of iterative model development, and made sure that our tech memos accurately recounted our efforts. He conducted the initial testing of the predictive pricing model by comparing multiple machine learning algorithms. David also delivered part of the midterm presentation.

Rosella took the initiative to transform our dataset using the TF-IDF feature extraction method, getting the ball rolling on our project. She consistently lent a hand to any team member when needed. She scheduled meetings with our mentor to make sure that we stayed on the same page. She and Lizhou attempted the grouping algorithms using PCA and K-Means clustering, identifying where the Colgate brand stands in the market. Rosella also delivered part of the midterm presentation and part of the tools and techniques presentation.

Simran fostered a collaborative environment by consistently engaging and encouraging each team member to check their progress and develop the project at each stage. She took on the role of working with the Doc2Vec transformation as well as performing the cosine similarity. Ensuring we as a team have a healthy relationship with our client has been one of her strengths. Simran also communicated with the client at all stages, keeping the client's requests and recommendations a priority. She broke the ice and delivered a strong elevator pitch for our capstone project.

[19] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Lizhou served as a resource for coding and model building. He was always willing to lend a helping hand to any team member. Lizhou assisted Rosella with the TF-IDF model and assisted Simran with researching the Doc2Vec transformation, and trained and tested the grouping model. Lizhou also delivered part of the tools and techniques presentation and the live demonstration.

Having diverse and empathetic team members allowed our team to grow and learn from one another while thriving with positive attitudes toward success.

Conclusions and Future Work

The results from our project show that Colgate's pet food descriptions tend to indicate a lower price relative to the actual price Colgate charges for their pet food products due to vocabulary that is less trendy and more generally descriptive. Given that Colgate’s pet food brand, Hill’s Science Diet, is among the top pet food brands recommended by veterinarians, Colgate’s price point is most likely supported by consumers who are willing to pay more for pet foods that are recommended by a veterinarian, which was noted in the studies discussed in the literature review.

To elevate the estimated value of their pet food product descriptions, the team recommends that Colgate consider expanding its current pet food description vocabulary to reflect current trends. The team also recommends that Colgate explore the possibility of increasing both its social media engagement and the promotion of this engagement in its product descriptions. As noted in the results, the presence of the word “Facebook” in the pet food description added value to the overall price. However, additional research is needed to support this assessment.

Lastly, Colgate may want to consider a redesign of its product packaging. Hill's Pet Nutrition uses portraits of animals superimposed on plain white backgrounds for their customer-facing packaging, which tends to be less visually appealing than some newer brands based on the teams first-hand experience looking at pet foods during the initial learning phase of the project. With deference to the growing trend of humanization and anthropomorphism of pets discussed in the literature review, Colgate may also want to consider using less “sciency” language and more emotion evoking language in its descriptions. As one team member mentioned anecdotally, “No one wants a bowl of science for breakfast.”

For future research, the team recommends that those employing the price prediction model explore the idea of training and operating the model on subsets of the data to achieve better accuracy. Rudimentary testing of the model on specific subsets of the data showed improved performance in some cases.

Ultimately, the results of this project will help Colgate take the lead in pet food product labeling by getting ahead of future trends while right-pricing future products. This information is incredibly insightful for our client and will help them maintain a competitive edge while ensuring that any proposed pet food descriptions meet Colgate’s high standards while staying on-brand.