PET PRODUCT TAXONOMY FOR E-COMMERCE

Team Six Dogs Ruyue Meng, Tiancheng Sun, Diantian Fu, Sheng Qian, Leona Liu, Zijue Li

Link to code https://drive.google.com/drive/folders/1RaLfxQWaakeK7iPwlKCqMv1WiEq-XL-y?usp=sharing

TABLE OF CONTENT

EXECUTIVE SUMMARY **1**
- Project Background 1
- Problem Statement 2
- Our Solution 2
- Dataset Collection and Summary 2
- Baseline Model Selection 3
- Model Development 3

EXPERIMENTAL RESULT **6**
- Model Evaluation 6

DISCUSSION **8**

CONCLUSION, FUTURE WORK, AND LESSON LEARNED **12**
- Constraints & Challenges 12
- What We Learned and Would've Done Differently 12
- How To Improve and Expand 12

REFERENCES AND APPENDICES **14**

EXECUTIVE SUMMARY

Project Background

As information technology advances, e-commerce is quickly becoming a significant part of the retail market. E-commerce is now taking 19.1% of the total retail sales in the entire United States. Amazon is now the largest e-commerce platform in the US, hosting 6.3 million merchants on its site. With the help of NLP, we can use the enormous amount of product data from e-commerce to benefit the e-commerce merchants, customers, and the platform itself.

Problem Statement

Given a set of pet products data from Amazon including product titles, reviews, and related features, we intend to develop a "smart" classification model, that can make accurate hierarchical predictions of the category, and sub-categories of the pet product through the product's title, and buyers real-time reviews. This set of classification models will lay out the foundation for new features on the websites to boost operation efficiency, enable a future analysis to generate business insights, and so much more.

Our Solution

Aside from ordinary procedures, such as data cleaning, processing, EDA, and interpretation, we implemented three-part procedures to tackle the core problem of our project. A multi-classification model was developed and explained in the following section of the report.

What's More

After the final model selection, we generated product-level labels using the entire Amazon pet product dataset. We performed a qualitative analysis of the prediction results and identified a few interesting outliers, with an analysis of the faulty prediction. At the end of this report, we concluded the project constraints and challenges, business insights from the prediction results, and proposals on how to improve and expand the project in the future.

PROPOSED APPROACH

Dataset Collection and Summary

The Amazon dataset consists of 2,643,619 transactions of pet products (informed retrieved from Huggingface co.) The dataset include 15 features including product title, star rating, helpful votes, total votes, review headline, review body, review date,etc. We primarily focused on product title, product id, and reviews for our model analysis.

Baseline Model Selection

The classification model we intended to develop contains various layers of taxonomy (refer to the appendix for the classification model prototype). The first level of the model can classify different animals. Following to the first layer, the second level will be the subcategory classification of different products under each animal from the first layer.

To meet the need of the intended model output, we decided to use a multi-classification model with the assumption that each model belongs to one class. The clustering model is not applied in our project as the clustering model inherently contains bias on its own. To develop a supervised model, we selected a various number of manually labeled sample items to train the model, test the model, and evaluate model performance. If the model performs well on the training model, we can run the model on the whole amazon dataset and conduct performance evaluation (accuracy testing and statistical testing). The detailed execution process will be explained further in the model development section.

Dataset Processing and Evaluation

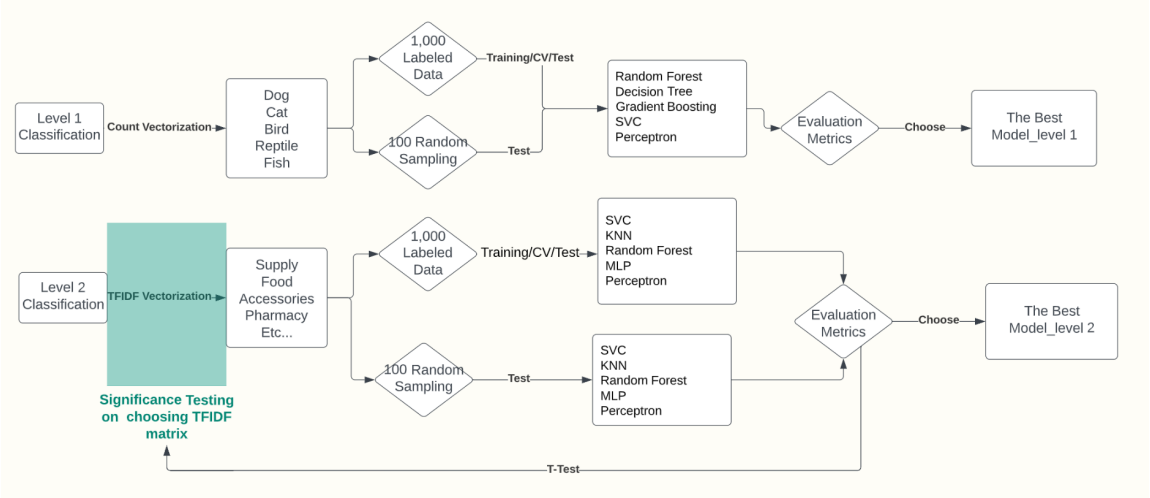
To develop the baseline model, we have manually selected 1,000 amazon reviews, and labeled the level 1 and level 2 classifications. For example, for a product for “cleaning cat privacy tent for litter cleaning,” we labeled “cat” as the first level and “cleaning” as the second level.

Due to the imbalanced dataset (64% of dog reviews, 30% of cat reviews), we stratified the labeled dataset to reduce the imbalance issue. With 1000 labeled samples, we retrieved roughly 50,000 data as our training dataset to fit models, using the product ID.

Some other data preprocessing procedures we have performed include Lemmatization, Stopword (non-English brand names, punctuation, numerical values), TFIDF, and String Contains (grouping same context words together such as doggie, dog, puppy **).**

Model Development

Step 1: Model Selection



Figure[1]: Model construction flow chart

Level 1 Classification

In the first level of model selection, we used word frequency to build the model. We searched the frequency of pet species. (for example dog, cat, fish, birds, and reptile) in the product title, review title, and review body. To categorize the pet species, we randomly selected and manually labeled 1,000 training data. We trained the pet species' word frequency and labeled data as the first step.

Further, we randomly sampled another 100 data. To ensure the randomness of the 100 test data, we first randomly sampled 10 data from the entire data and repeated the process 10 times. As a result, we are able to get a very random 100 test data and this 100 test data does not have any selection bias. In other words, the distribution of these 100 test data will be similar to the distribution of the whole data.

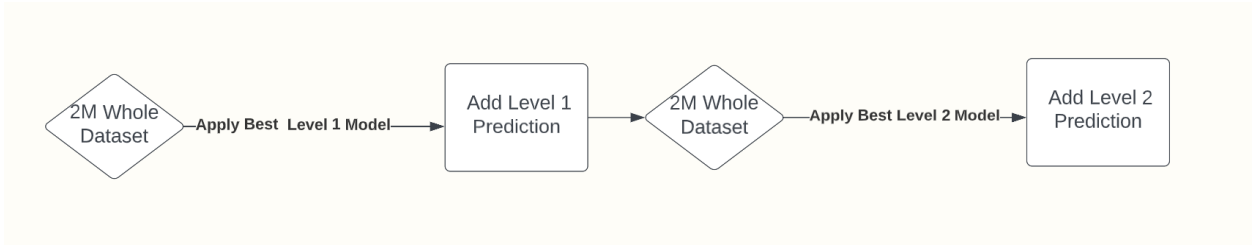
We used a cross-validation method to compare the models' performance by comparing their accuracy, precision, recall, and F-1 score. We also used a secondary method to test the performance of each model. We used 100 test data comparing their accuracy, precision, recall, and F-1 score to find the best model as well. With two selected evaluating methods, the best model we picked is the random forest model.

Level 2 Classification

At the second level, we used TF-IDF vectorizer to vectorize the data. For corpus selection, we had the option to use either the product title or a combination of the product title and review body. We were not sure which one would give a better model. There is no free lunch in machine learning so we tried both metrics. We used both metrics to build up SVC, KNN, Random Forest, Multilayer-perceptron, and Perceptron. The next question is to select the model with the best performance. We used the same method as we used at the first level. Therefore, for both metrics, we have their accuracy, precision, recall, and F-1 score.

Based on significance testing performed, we found out that there is no significant difference between using only product titles or using a combination of product titles and review bodies. We chose to use the combination of product title and review body as a corpus to build the TF-IDF matrix as we believe it will contain more information than only using product titles.

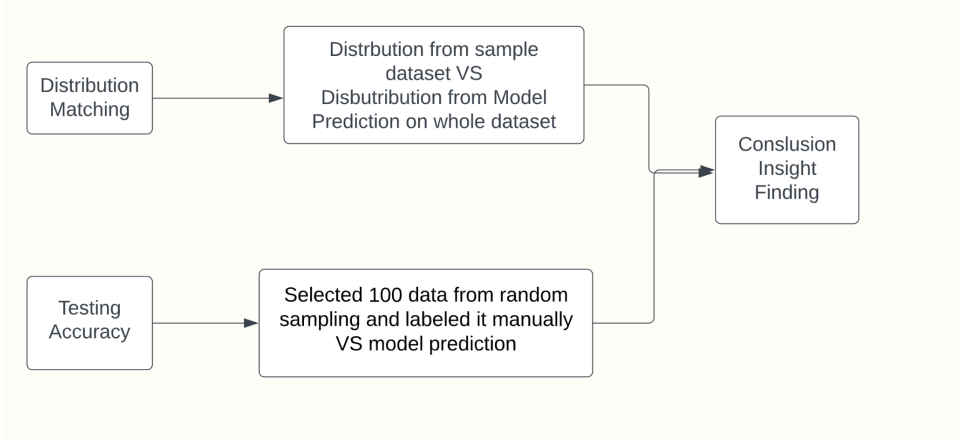
Step 2: Apply Prediction



Figure[2]: Application flow chart

For the second step, we used the best model from the first level and the best model from the second level to predict the entire data so that we can have a prediction for the entire data.

Step 3: Verify Prediction



Figure[3]: Prediction flow chart

We used two methods here to verify the prediction.

The first method is to get the distribution of our 1,000 manually labeled training data. Because we have the prediction of the entire data and we can get the distribution of the prediction of the entire dataset. We made the comparison between the two distributions and discovered that there is no big discrepancy. This indicates that our training set does not have selection bias.

The second method is using 100 randomly selected manually labeled data to verify our prediction again. This 100 data does not have a selection bias problem. After fitting these 100 test data and we still get a good performance, we can conclude that our training data do not have a selection bias issue. The result of the accuracy, precision, recall and F-1 score of this 100 test data is good, implying that we do not have a selection bias issue.

EXPERIMENTAL RESULT

Model Evaluation

To evaluate the actual results of our models, we looked at 4 measurements, average Cross-Validation Score for training and accuracy for testing, macro-weighted average Precision, macro-weighted average Recall, and macro-weighted average F1-score.

Among them, the most significant one is the F-1 score which is the harmonic mean of precision and recall, and it can balance the 'tradeoff' between precision and recall. Below is an example of our code output from one of our models.

	Bird	Cat	Dog	Fish	Reptile	accuracy	macro avg	weighted avg
precision	0.969605	0.984039	0.979516	0.996564	0.905405	0.980417	0.967026	0.980524
recall	0.990683	0.955204	0.991374	0.973154	1.000000	0.980417	0.982083	0.980417
f1-score	0.980031	0.969407	0.985409	0.984720	0.950355	0.980417	0.973984	0.980363
support	322.000000	2969.000000	6608.000000	298.000000	67.000000	0.980417	10264.000000	10264.000000

Figure[4]: Model evaluation for level 1

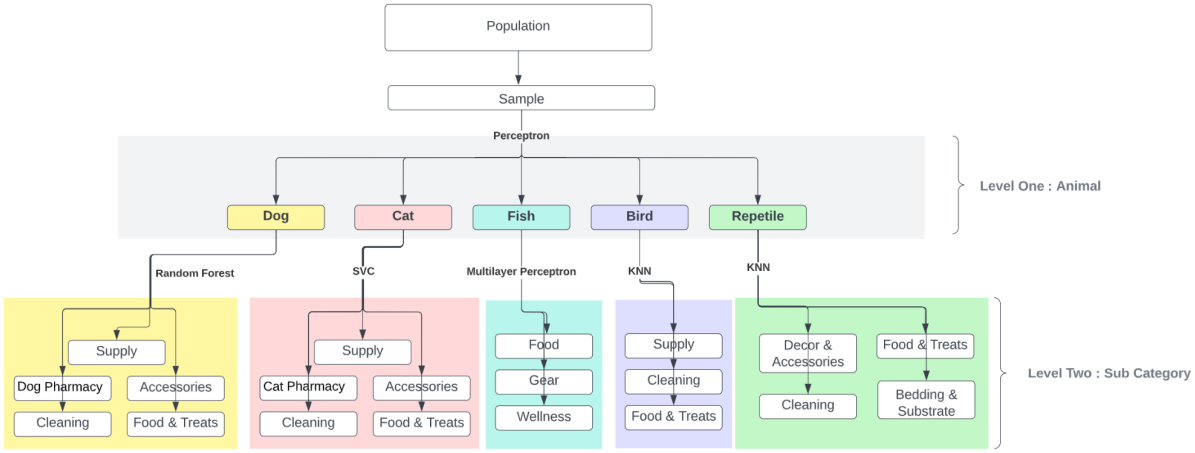
For each model, we calculate those metrics and compare them to get the best model for training and testing.

For our level 1 classification, we tried approximately 10 different models, we listed the top five. We were able to achieve a great CV and F-1 score in general, with the random forest being the best model while others are close.

As those models were applied to the testing dataset, the accuracy and Macro Weighted Avg. F-1 Score still looks promising, indicating that the models are scalable and the possible selection bias does not impact the models significantly.

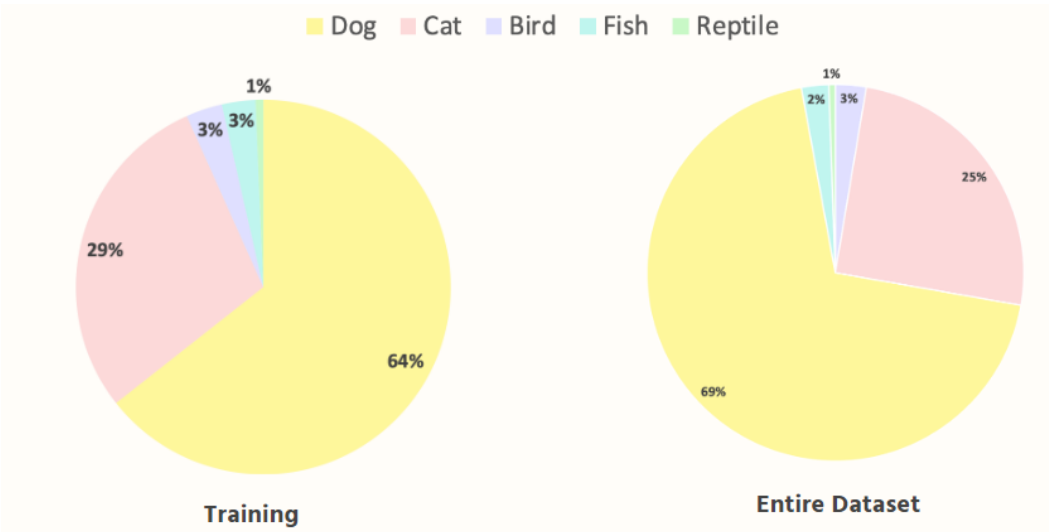
These are the level 2 dog models and it looks like the random forest is the best model which continues to be the case for testing

Below is a summary of our structure and best models based on the measurements mentioned above.



Figure[5]: Best models summary

The other evaluation method we used is to check the distribution for our manually labeled training sample and prediction after labeling the entire dataset. As you can see the two distributions are very similar. We checked for both level 1 and level 2 for all five animals, and they all indicated similarity.

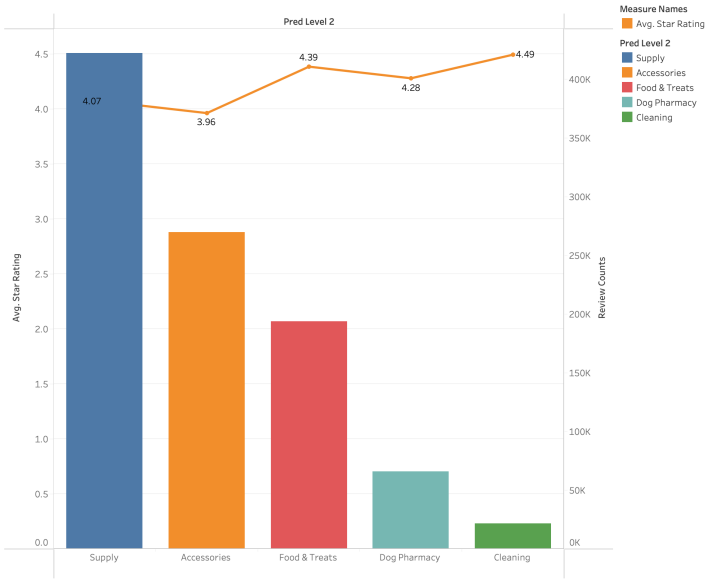


Figure[6]: Distribution checking for training and prediction

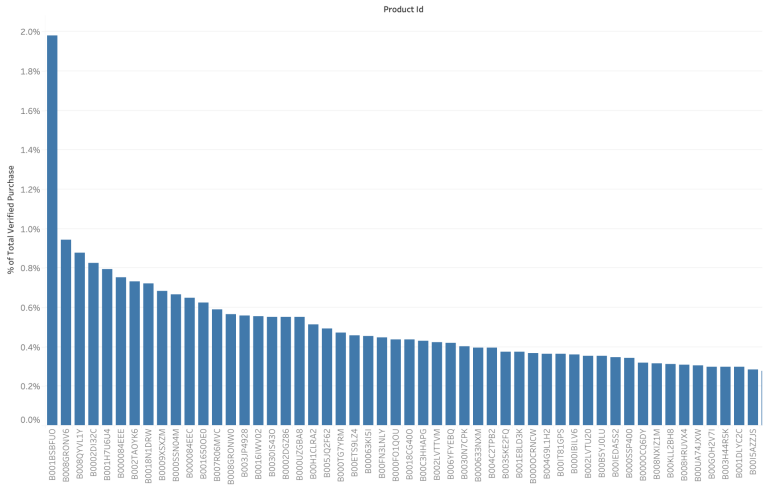
DISCUSSION

Dog sub-categories analysis:

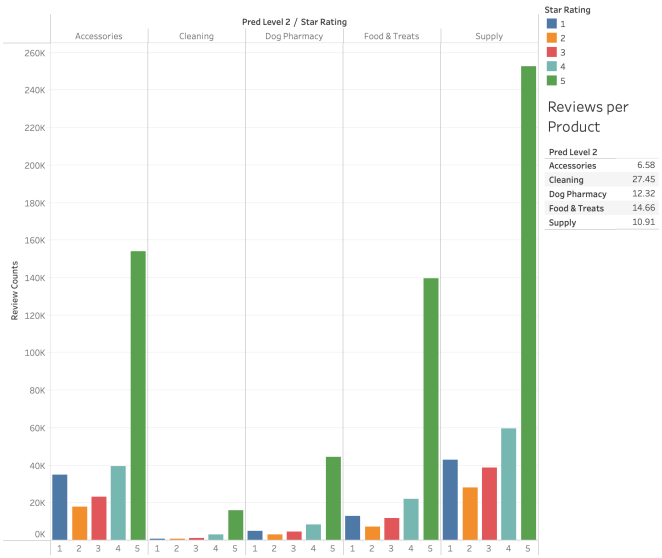
1. The supply sector has the most reviews and 4 out of 5-star ratings on average, which indicates that there is a lot of competition in this sector.
2. The accessories sector has a lot of reviews but the lowest star ratings, which means there is still some space for new entries in this sector, and further analysis is needed for business growth opportunities.
3. For each sector, the purchase follows a 20/80 distribution among products, which means roughly 80% of the purchases happen in the top 20% of products. This trend is very obvious on the Dog Pharmacy section: one product (Allergy Formula) takes 2% of the total purchase
4. From the reviews per product, we found that customers are pickier and more satisfied in the Cleaning Sector: on average, there are 30 reviews for one cleaning product and most of the reviews are five-star ratings.



Figure[7]: Dog sub-categories distribution & average rating



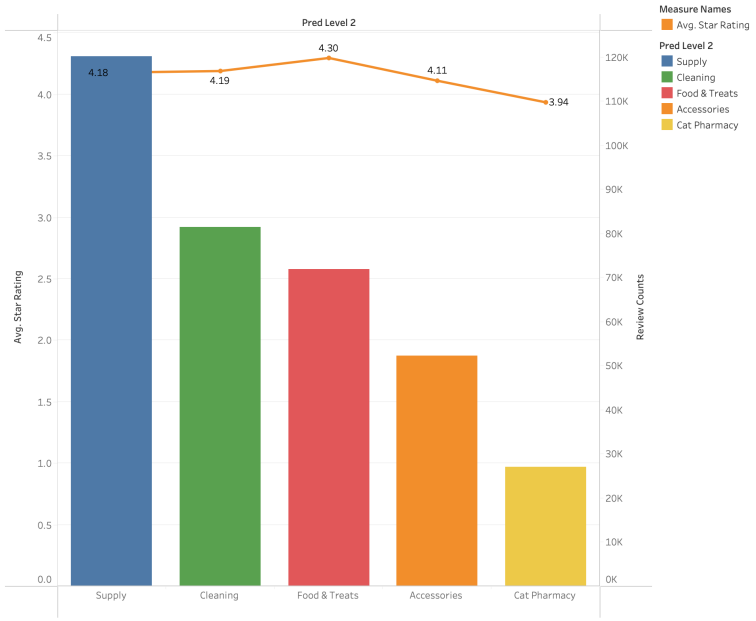
Figure[8]: Dog sub-categories product distribution



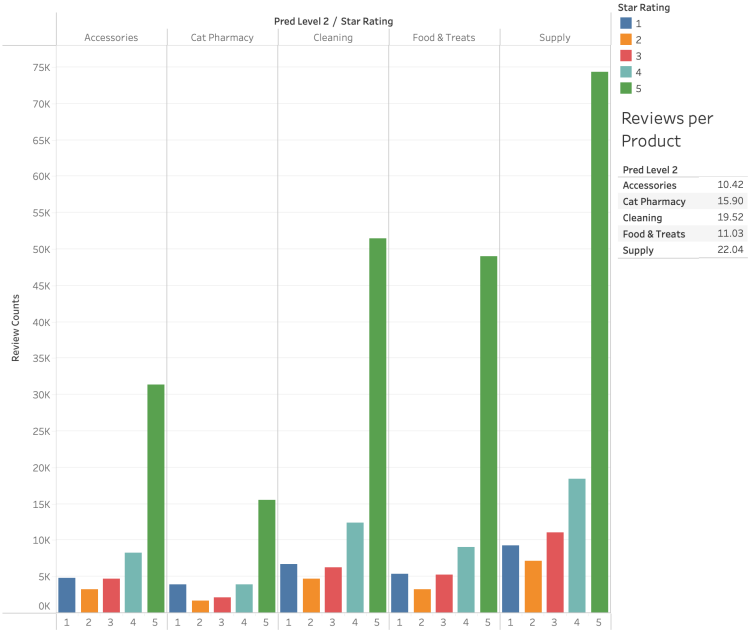
Figure[9]: Dog sub-categories review per product distribution

Cat sub-categories analysis:

1. The Food & Treats sector has the highest rating 4.3 out of 5.0, which means customers are satisfied with the Food & Treats products and it will be relatively harder for new entries to share the market.
2. The average rating on Cat Pharmacy sector products is the lowest, which indicates that customers are not happy with the current products and there might be business growth opportunities for new products.
3. The purchase follows a 20/80 distribution among products.
4. The Accessories and Food & Treats sector has fewer comments per product. It might be this sector has a more standardized requirement that customers are commonly shared so that they don't need to post reviews.



Figure[10]: Cat sub-categories distribution & average rating

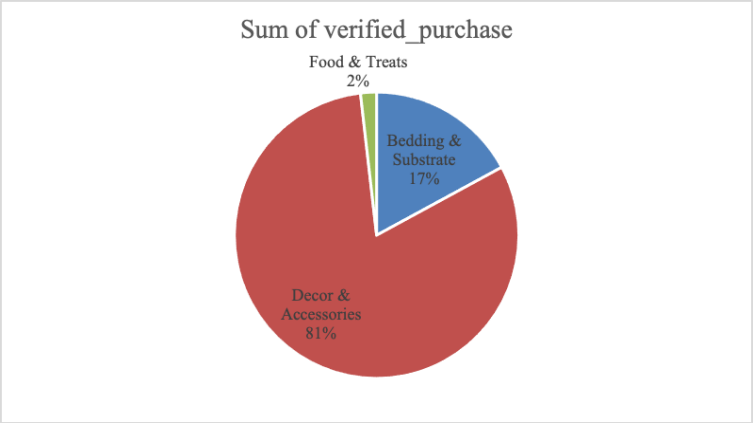


Figure[11]: Cat sub-categories review per product distribution

Reptile sub-categories analysis:

1. The Food & Treats sector only takes 2% of total purchases and 81% of purchases happen in the Decor & Accessories sector.
2. The average rating for all of the three sub-categories is below 4 stars, which means customers are not satisfied with Reptitle's product and improvement is needed for all

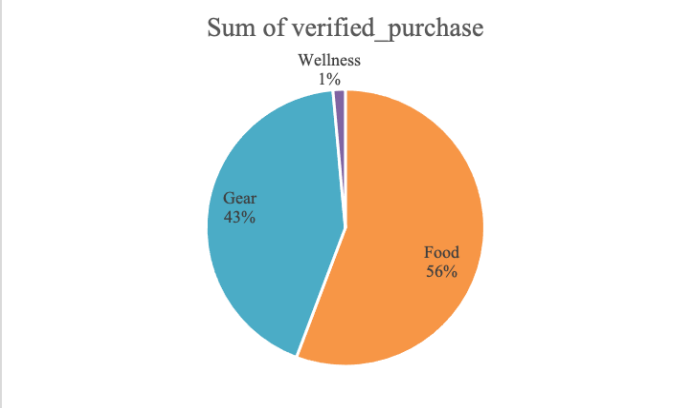
sub-categories.



Figure[12]: Reptile sub-categories distribution

Fish sub-categories analysis:

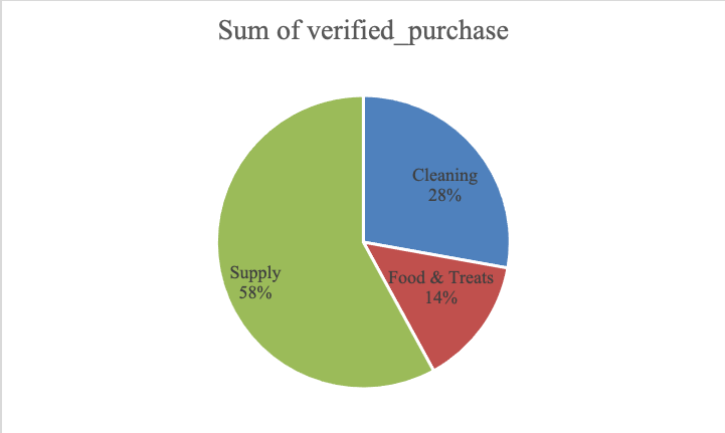
1. The gear sector has 43% of purchases and the average rating is 3.86 out of 4, which means there might be space for new business opportunities.



Figure[13]: Fish sub-categories distribution

Bird sub-categories analysis:

1. Supply sectors has the highest percentage of purchases.



Figure[14]: Bird sub-categories distribution

CONCLUSION, FUTURE WORK, AND LESSON LEARNED

Constraints & Challenges

We were concerned about the training imbalance due to the uneven distribution of pet product purchasing records. We, therefore, used the smote methodology to oversample the training dataset to elevate the problem. We also suspected the possibility of selection bias due to stratification. For that, we used 100 randomly selected and labeled data as test data to evaluate the model performance, and as stated earlier, the prediction result for the test set was positive.

The nature of the dataset determined that we won't have a set of the ground truth of the whole dataset to test the model performance. We also, due to the constraint of time, couldn't develop a multi-label model that can put different labels on the same product. These are the constraints that we have to work with during this project.

What We Learned and Would've Done Differently

We realized that we would have developed multi-label classification models to enable a more precise model after we are done with data labeling and model building for level 1. At that point, we were too far into the project process, and can't start over due to the time constraints. Moving forward, we would do a thorough inspection of the project flow design before starting the concrete process.

How To Improve and Expand

Aside from switching to a multi-label classification model, we can improve our project through:

Add more depth of hierarchy to provide more precise classification results: we have the level three label available. Due to the time constraint, we couldn't finish the model building and selection as we need to train 100 models for level 3

Optimize the coding efficiency: data pre-processing and model training takes a long time to finish. If we can optimize the coding efficiency, we will be able to bring this project to a larger scale with more layers.

Find insights in ambiguous reviews: we found that review rather introduces factors of misclassification, but carries business insights. For example, these factors could indicate that this product can be used for different animals, or provide reasons for why this product is bad. It would be worthwhile spending more time exploring ambiguous reviews to generate business insights.

Category performance analysis: we generated quite some business insights from the result of the classification, and hope to expand on that process

REFERENCES AND APPENDICES

Data source:

The HF Datasets community. (n.d.). *Amazon_us_reviews · datasets at hugging face*. amazon_us_reviews · Datasets at Hugging Face. Retrieved October 10, 2022, from https://huggingface.co/datasets/amazon_us_reviews#annotations

Classification model prototype