

Pet Expense Management

A primary pain point within the pet industry is poor data quality and information overload: pet owners are not equipped to record pet data in a structured manner and also do not understand which information is critical. Paw á Peau is attempting to address this problem by building an app to deliver actionable insights from pet data and improve the overall quality of pet care. The challenge of this data science project is thus to translate the benefits of tracking information into financial terms, in order to motivate pet owners.

Objectives

In Germany, pet ownership has increased by over 30% since the pandemic and currently lies at 15.7 million cats and 10.7 million dogs (2020, Zenz-Spitzenweg). Dogs are the second most popular pets, but dog owners spend far more than any other category of pet owner and thus will be the focus of this study.

The project aims to:

- Determine the key criteria most correlated with dog ownership expenses.
- Identify risk factors associated with higher costs, such as breed category.
- Explore KPIs that could be used as early warning indicators for avoidable expenses.

Machine Learning

In this phase of our project, we were to choose a machine learning algorithm to train and tune a model to predict the categories of yearly expenses. We created five categories, labelled 'lowest', 'low', 'medium', 'high' and 'highest', representing yearly expenses binned by calculated quartiles.

Feature Engineering

The feature engineering process took a lot of time due to missing, dirty, or incorrectly formatted data. We combined many different datasets (as can be seen in the references section) and therefore we had to make decisions about which format to prefer over the others and subsequently transform the complementary datasets accordingly. Additionally, we enriched the costs using figures from scientific literature and our own primary research.

We standardized the data by ensuring certain thresholds for breed inclusion in the dataset, this included the minimum number of dogs which belonged to a given breed population, as well as the acceptable amount of missing key data. For example, dog breeds missing our key behaviour traits were dropped from the dataset. We also chose to represent mixed breeds by using the mode for missing data, i.e. behaviour traits, as well as figures presented in literature such as longevity, reduced insurance cost, and the decision to not associate the population with specific genetic ailments.

We analyzed the dataset with one hot encoded transformed 'breed' data, as well as without 'breed' data, and with breed labels encoded simply with numerics. Finally, we realized that 'breed' as a variable was not essential and that representation via 'breed_group' was sufficient. The foundation dataset had provided categorization based on the American Kennel Club, but we later changed this to the categorization provided by the Fédération Cynologique Internationale (an umbrella organization governing pedigree dogs in most of regions of the world) as these were more numerous, purpose-specific, and more relevant to Germany.

We dropped the following variables, due to irrelevance, better representation through other variables, or insufficient data:

'neutered', 'grooming_required', 'food_cost_year', 'lifetime_cost',
'adapts_well_to_apartment_living', 'bite_statistics'

We also expanded 'genetic_ailments' through one hot encoding, in order to have a more detailed breakdown of the costs. Following the calculations of individual genetic ailments and common training costs (local cost data in Berlin was used) to address certain behavioural problems, we removed all cost data apart from 'yearly_final_costs', which we used as our target.

Model Selection

After realizing that many individual cost parameters were missing from our dataset, we decided to change our approach away from regression and instead to classification. We tested various classification machine learning algorithms:

- Random Forest
- Decision Tree
- Boosting
- K-Nearest Neighbour
- Support Vector Machine

Our first step was preparing two common datasets, *combined_data.csv* (with breeds and ailments encoded) and *nobreed_data.csv* (without 'breed', only 'breed_cat'). We then separately tested two models each, along with various techniques such as Pruning and LIME. Additionally, various types of feature importance analyses were performed to determine the most important variables for the model's calculation. We selected Random Forest as our final model for the completion of the project due to its performance in comparison to the other models.

Model Interpretation

Random Forest

One strategy we employed to standardize the data was balancing the samples of different cost categories, so that there was no model bias due to an overabundance of any given demographic. The expense classes were thus reduced to 4, as opposed to the original 5 ('lowest', 'low', 'medium', 'high'). Using the Random Forest algorithm, we tried removing the variables that had little relevance - defined as ≤ 0.01 importance and re-testing the model's performance. The features of low importance which were removed included 'breed', some genetic ailments and 'grooming_required'.

Breed was already sufficiently represented by breed groups, which broadly categorized the dogs by purpose, for example 'category_working' includes breeds such as Giant Schnauzer, Dobermann, German Shepherd, Malinois and others which are typically used by police or security. Breed groups tend to share similar characteristics such as size and temperament traits which could in some cases be translated to similarity in costs. When it comes to physical traits such as 'grooming_required' and 'tolerates_cold_weather', there were implications of dogs with a high score translating to having a thick coat or long fur. The use of both variables would in such cases be unnecessary.

After reducing the feature set, the model still showed perfect scores of accuracy for the train and test sets:

```
Cross-validation scores: [0.76377953 0.84839895 0.60488189 0.46094488
                          0.61503937]
Mean cross-validation score: 0.6586089238845144
```

The cross-validation scores indicate that the model's performance is variable across different data samples, with an average performance of approximately 0.66. This result is indicative that the model may have still been overfitting, or that the features are not entirely representative of the target variable.

Performance after Regularization

```
New Model Performance with Regularization:
Training Accuracy: 0.9876771653543307
Testing Accuracy: 0.986745406824147
Precision: 0.9868037873845078
Recall: 0.986745406824147
F1-score: 0.9867435260246461
```

Precision, Recall and ROC AUC after RF Regularization

Training Classification Report:

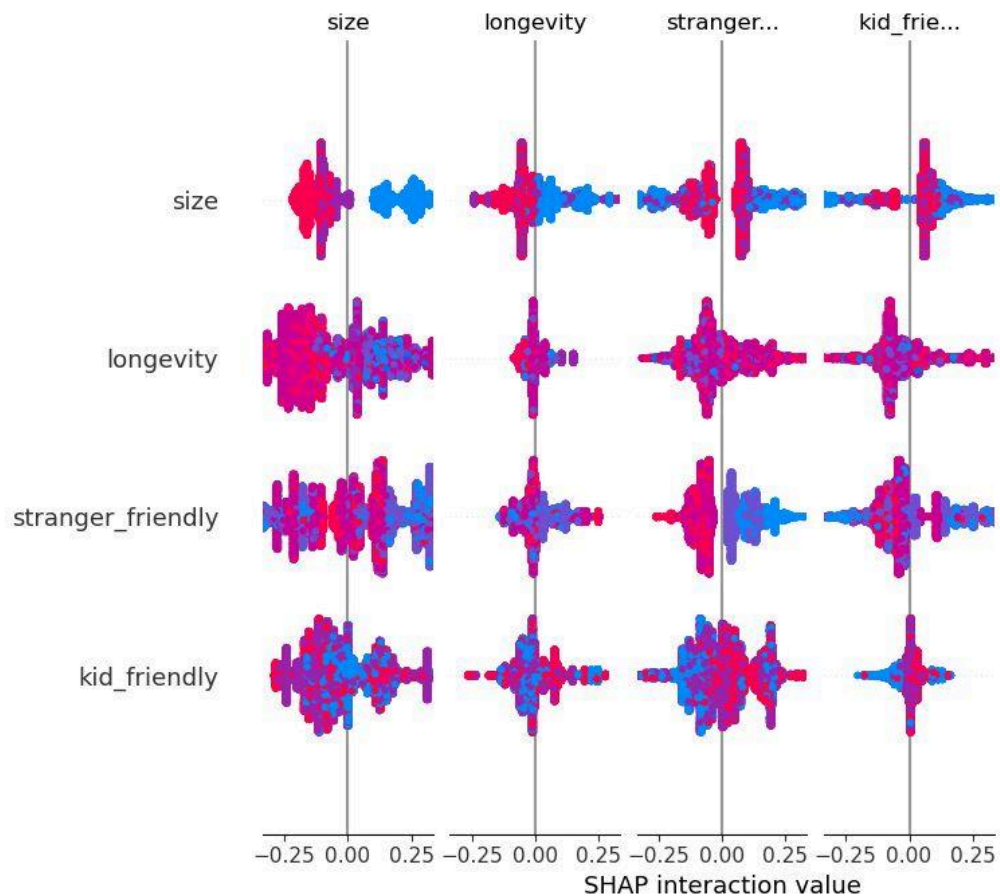
	precision	recall	f1-score	support
High	1.00	0.98	0.99	37833
Low	0.98	0.99	0.99	38189
Lowest	0.99	1.00	1.00	38340
Medium	0.98	0.98	0.98	38038
accuracy			0.99	152400
macro avg	0.99	0.99	0.99	152400
weighted avg	0.99	0.99	0.99	152400

Testing Classification Report:

	precision	recall	f1-score	support
High	1.00	0.98	0.99	9339
Low	0.98	0.99	0.98	9726
Lowest	0.99	1.00	1.00	9416
Medium	0.98	0.98	0.98	9619
accuracy			0.99	38100
macro avg	0.99	0.99	0.99	38100
weighted avg	0.99	0.99	0.99	38100

```
Training ROC AUC: 0.9998189004379342
Testing ROC AUC: 0.999768424071771
```

SHAP Chart (Random Forest)



From the SHAP chart, we can see that large size and poor longevity is positively correlated with a higher cost class, and that smaller size leads to better longevity and lower cost. For other variables, the interpretation was not so clear cut.

LIME (on Random Forest)

LIME stands for “local interpretable model-agnostic explanations”, and is a technique and generic framework to uncover black boxes and provides an explanation for predictions or recommendations with a local, interpretable model to explain each value. LIME values represent the extent to which each feature contributes to the prediction in a specific instance. Positive values indicate features with a higher probability of predicting the class (low, lowest, medium, high), while negative values indicate features with a low probability of predicting the target classes. Features with larger absolute LIME values (whether positive or negative) have a stronger impact on the model's predictive performance.

Below is the result of LIME run on the RF model.

Feature	Value
size	-1.19
category_terrier	2.18
category_sporting	-0.48
stranger_friendly	0.48
sensitivity_level	-0.64
dog_friendly	0.33
longevity	1.38
kid_friendly	0.01
thyroid	-0.58
exercise_needs	0.38
wanderlust_potential	0.89
prey_drive	0.66
tolerates_cold_weather	-0.05
severity_score	-0.38
potential_for_mouthiness	1.14
tendency_to_bark_or_howl	1.53
tolerates_hot_weather	0.26
tolerates_being_alone	1.38
intelligence_category	0.56
energy_level	0.45
category_working	-0.46
category_companion	-0.50

Feature Perturbation

A feature perturbation analysis was performed on the first 4 instances.

- **Common Sensitivities:** across all instances, features such as 'size', 'intelligence_category', 'sensitivity_level', and certain behavioral traits (e.g. 'tendency_to_bark_or_howl') consistently influenced prediction probabilities.
- **Impact Variation:** Perturbations often lead to shifts towards predictions that are adjacent (e.g., Lowest to Low), indicating robustness of the model's predictions in most cases.
- **Complexity:** Some features showed minimal impact on predictions, suggesting that they either had a lower importance in the model's decision-making process or that their interactions with other features were not fully captured in isolation.

We examined a few of the features more in detail:

- 'stranger_friendly' represents a dog's predisposition to friendliness (or alternatively, aggression) when encountering a stranger. Perturbation in the case of this feature had a noticeable impact across all instances, altering the prediction probabilities significantly. This suggests that this feature is quite influential in the model's decision-making process.
- The features 'size', 'severity_score', 'intelligence_category', and 'longevity' affect the prediction in most instances, which indicates their high relevance for the model.
- Features such as 'category_companion', 'category_sporting', and 'category_working' showed little to no change in prediction probabilities when perturbed across all instances. These features might therefore have minimal influence on the model's predictions or already be balanced within the model's considerations.

Decision tree

In order to understand the high performance of our Random Forest model, and with the idea of comparing it with a similar branching model, we decided to apply a Decision tree. Below is an analysis of the Decision Tree performance, along with a classification report.

```
Decision Tree Model Performance:
Training Accuracy: 0.9662073490813649
Testing Accuracy: 0.9642782152230971
Precision: 0.9654977356639471
Recall: 0.9642782152230971
F1-score: 0.9644424253729907
Cross-validation accuracy scores: [0.72440945 0.90981627 0.83685039
0.52007874 0.6244357 ]
Mean cross-validation accuracy: 0.7231181102362205
Training Classification Report (Decision Tree):
```

	precision	recall	f1-score	support
High	0.99	0.95	0.97	37833
Low	0.97	0.94	0.96	38189
Lowest	1.00	0.99	0.99	38340
Medium	0.91	0.98	0.94	38038
accuracy			0.97	152400
macro avg	0.97	0.97	0.97	152400
weighted avg	0.97	0.97	0.97	152400

Testing Classification Report (Decision Tree):

	precision	recall	f1-score	support
High	0.99	0.94	0.96	9339
Low	0.97	0.94	0.96	9726
Lowest	0.99	0.99	0.99	9416
Medium	0.91	0.98	0.94	9619
accuracy			0.96	38100
macro avg	0.97	0.96	0.96	38100
weighted avg	0.97	0.96	0.96	38100

Training ROC AUC (Decision Tree): 0.998118718690002

Testing ROC AUC (Decision Tree): 0.9978297076792021

Random Forest Performance

- The Random Forest model showed high performance in terms of accuracy, precision, recall, F1-score, and ROC AUC, both on the training and testing datasets.
- The cross-validation scores show a significant drop, indicating potential overfitting despite regularization.

Decision Tree Performance

- The Decision Tree model has slightly lower overall performance metrics compared to the Random Forest.
- However, it has better cross-validation accuracy scores, suggesting it generalizes better than the Random Forest.
- The performance gap between training and testing metrics is smaller, indicating less overfitting.

Hyperparameter Tuning with RandomizedSearchCV:

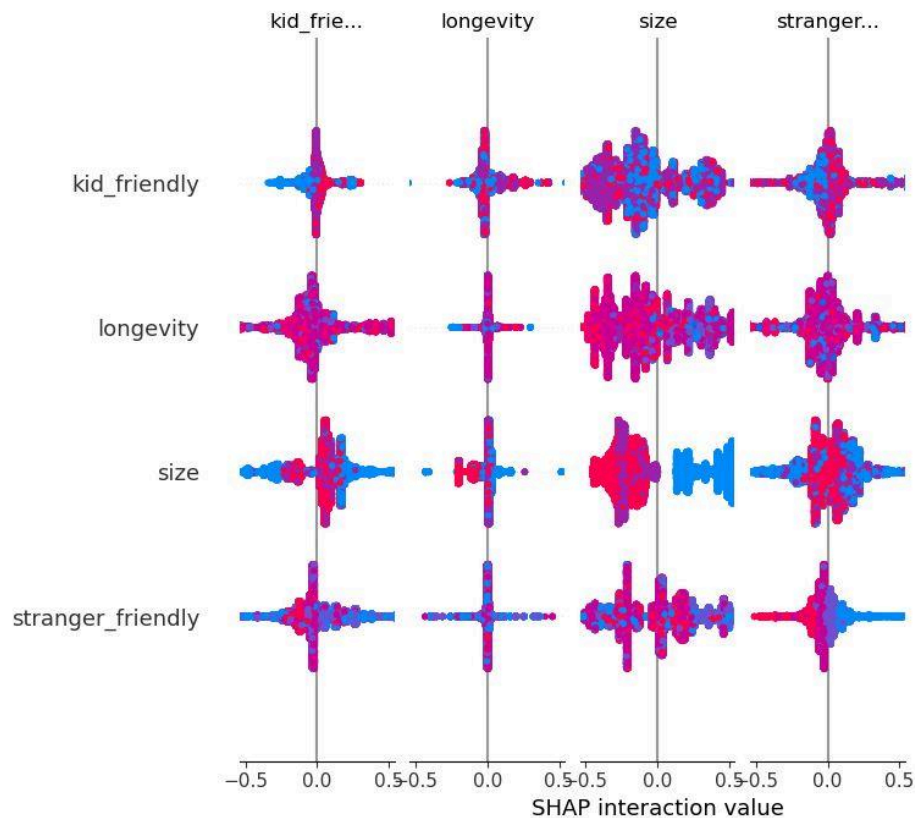
Random Forest Best Parameters: {'n_estimators': 50, 'min_samples_split': 20, 'min_samples_leaf': 10, 'max_depth': 20}

Random Forest Best Cross-validation Score: 0.9955774278215224

Decision Tree Best Parameters: {'min_samples_split': 20, 'min_samples_leaf': 2, 'max_depth': 20}

Decision Tree Best Cross-validation Score: 0.9992979002624672

SHAP Chart (Decision Tree)



Binarizing the Data

As a second strategy, we decided to use **dogs_expenses.csv**, a version of the dataset in which we modified the target variable to a binary in order to analyze the effect on the model. The variable 'health_severity_score' was calculated based on the health insurance cost and the severity score. Both variables influence the severity of 'genetic_ailments' for the different breeds. A threshold, based on the median, was used to binarize the score (0 for low severity, 1 for high severity).

```
### Feature Importances ###
      feature  importance
0      longevity  0.098617
2    size_category  0.097173
6  tolerates_cold_weather  0.076247
3  intelligence_category  0.067648
14    exercise_needs  0.060961
13  wanderlust_potential  0.056421
10  potential_for_mouthiness  0.053552
11      prey_drive  0.053138
7    tolerates_hot_weather  0.048574
15    energy_level  0.044288
```

12	tendency_to_bark_or_howl	0.043209
5	tolerates_being_alone	0.041560
4	sensitivity_level	0.036072
9	friendly_towardstrangers	0.030588
1	grooming_required	0.030143
23	category_Working Dogs	0.028650
17	age	0.027278
22	category_Terrier Dogs	0.022014
8	incredibly_kifriendly_dogs	0.021601
19	category_Herding Dogs	0.017662
21	category_Sporting Dogs	0.014496
20	category_Hound Dogs	0.011657
18	category_Companion Dogs	0.011102
16	gender	0.007348

Classification Reports

Listed below are a collection of classification reports from various techniques:

```
### Classification Report-Random Forest###
              precision    recall  f1-score   support

         0           1.00        1.00        1.00        20419
         1           1.00        1.00        1.00        17681

 accuracy              1.00              1.00              1.00        38100
 macro avg              1.00              1.00              1.00        38100
weighted avg              1.00              1.00              1.00        38100
```

ROC AUC Score: 1.0000

```
### Cross-Validation Scores (AUC)- Random forest###
[0.81284402 0.79371676 0.74240547 0.67848494 0.69157041]
Mean AUC: 0.7438
Standard Deviation AUC: 0.0534
Model and scaler saved successfully.
```

```
### Classification Report ###
              precision    recall  f1-score   support

         0           1.00        1.00        1.00        20419
         1           1.00        1.00        1.00        17681

 accuracy              1.00              1.00              1.00        38100
 macro avg              1.00              1.00              1.00        38100
weighted avg              1.00              1.00              1.00        38100
```

ROC AUC Score: 1.0000

```
### Cross-Validation Scores (AUC)- Logistic regression ###
```

[0.64048978 0.72754026 0.63176795 0.54849899 0.52785082]

Mean AUC: 0.6152

Standard Deviation AUC: 0.0716

Model and scaler saved successfully.

Classification Report

	precision	recall	f1-score	support
0	0.72	0.75	0.74	20419
1	0.70	0.66	0.68	17681
accuracy			0.71	38100
macro avg	0.71	0.71	0.71	38100
weighted avg	0.71	0.71	0.71	38100

ROC AUC Score: 0.8069

Cross-Validation Scores (AUC) with Balanced Class Weights

[0.63948701 0.73019646 0.63120315 0.55111711 0.52770767]

Mean AUC: 0.6159

Standard Deviation AUC: 0.0719

Balanced model and scaler saved successfully.

Classification Report for Best Logistic Regression

	precision	recall	f1-score	support
0	0.73	0.69	0.71	20419
1	0.66	0.71	0.68	17681
accuracy			0.70	38100
macro avg	0.70	0.70	0.70	38100
weighted avg	0.70	0.70	0.70	38100

ROC AUC Score: 0.8045

Best Parameters: {'C': 0.001, 'penalty': 'l2', 'solver': 'liblinear'}

Best Cross-Validation AUC: 0.6229830458624127

Cross-Validation Scores (AUC)- Gradient Boosting Classifier

[0.70585514 0.75835648 0.68035704 0.66656849 0.70124045]

Mean AUC: 0.7025

Standard Deviation AUC: 0.0314

Best Parameters: {'n_estimators': 200, 'max_depth': 7, 'learning_rate': 0.01}

Best Cross-Validation AUC: 1.0

Best GBM model and scaler saved successfully.

Training Accuracy: 0.9953608923884515

Testing Accuracy: 0.9949606299212599

Classification Report for Best GBM

	precision	recall	f1-score	support
0	1.00	0.99	1.00	20419
1	0.99	1.00	0.99	17681
accuracy			0.99	38100
macro avg	0.99	1.00	0.99	38100
weighted avg	1.00	0.99	0.99	38100

ROC AUC Score: 1.0000

Training Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	81581
1	1.00	1.00	1.00	70819
accuracy			1.00	152400
macro avg	1.00	1.00	1.00	152400
weighted avg	1.00	1.00	1.00	152400

Testing Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	20419
1	1.00	1.00	1.00	17681
accuracy			1.00	38100
macro avg	1.00	1.00	1.00	38100
weighted avg	1.00	1.00	1.00	38100

Training ROC AUC: 1.0

Testing ROC AUC: 1.0

K-Nearest Neighbors

K-Nearest Neighbours is another algorithm that was employed on the dataset. It was used on **clean_split_data.csv** (created by splitting **combined_data.csv**) with encoded breeds and **nobreed_data.csv** - without breeds.

Splitting the Dataset

A third strategy was employed in which we split the dataset vertically, to include only behaviour-related features, and alternatively only health-related features. The aim here was to determine whether certain types of features were having an exaggerated influence on the model's accuracy. We also decided to employ additional calculations for training costs based on certain behavioural traits and had previously used 'severity_score' to represent 'genetic_ailments' following their one hot encoding, and no financial multiplier for behaviour-related variables such as 'sensitivity_level', 'kid_friendly' etc.

Behaviour Only (with Breed)

Behaviour costs added to the 'final_yearly_cost' were calculated using regular training costs positively correlated with a high score for most behaviour variables. A negatively correlated score for 'kid_friendly', 'stranger_friendly', 'dog_friendly' was calculated using an increased multiplication of the regular training costs. As we did not have raw, uncalculated data, it was likely the reason for the model performing too perfectly.

BEHAVIOUR DATASET

Accuracy Score: 0.9997900262467192

Classification Report:

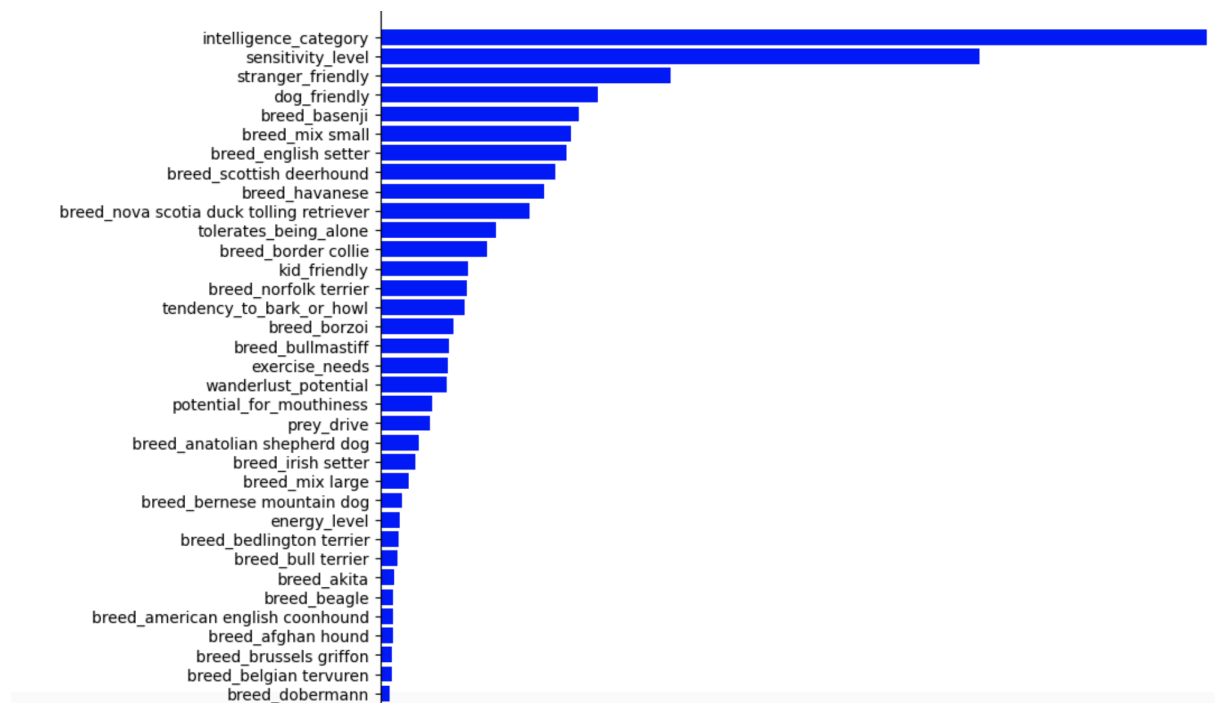
	precision	recall	f1-score	support
2	1.00	1.00	1.00	3366
3	1.00	1.00	1.00	3275
4	1.00	1.00	1.00	31459
accuracy			1.00	38100
macro avg	1.00	1.00	1.00	38100
weighted avg	1.00	1.00	1.00	38100

Confusion Matrix:

```
[[ 3360    6    0]
 [    0 3274    1]
 [    0    1 31458]]
```

Model Accuracy: 99.98%

LightGBM (with Breed) Behaviour Feature Importance Ranking



Behaviour Only (without Breed)

Removing the individual breeds did not improve the model's performance, likely due to the overall low importance of breeds and the cost figure calculation being simplistic as we lacked raw data.

----- BEHAVIOUR DATASET

Accuracy Score: 1.0

Classification Report:

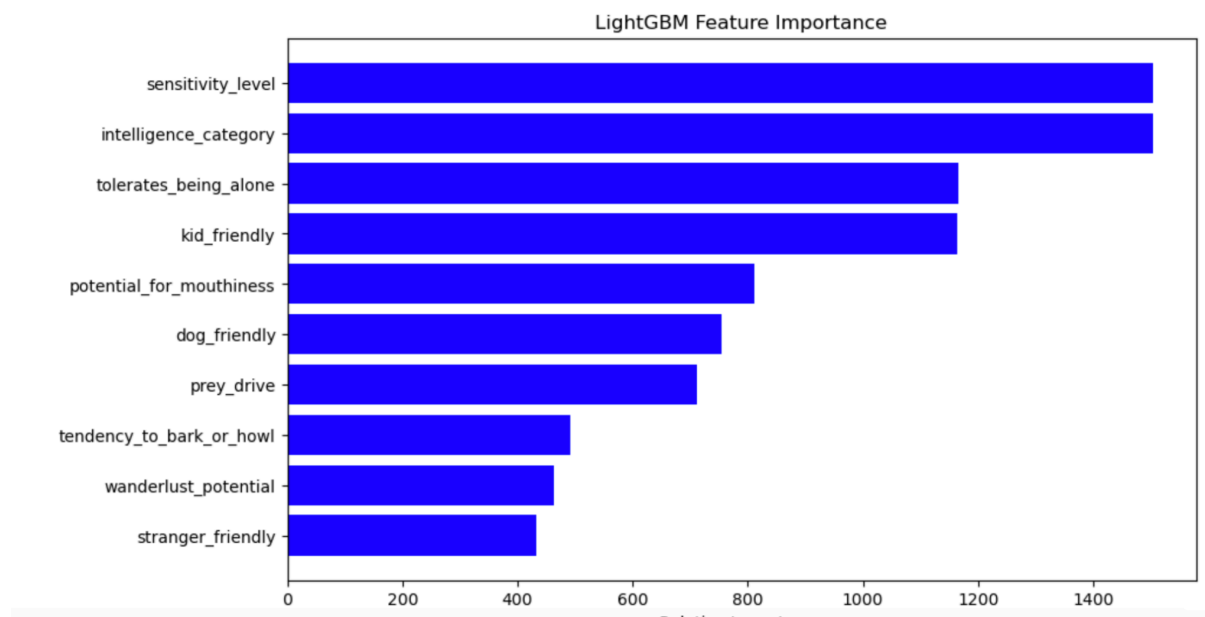
	precision	recall	f1-score	support
2	1.00	1.00	1.00	192
3	1.00	1.00	1.00	1145
4	1.00	1.00	1.00	36763
accuracy			1.00	38100
macro avg	1.00	1.00	1.00	38100
weighted avg	1.00	1.00	1.00	38100

Confusion Matrix:

```
[[ 192    0    0]
 [   0 1145    0]
 [   0    0 36763]]
```

Model Accuracy: 100.00%

LightGBM (without Breed) Behaviour Feature Importance Ranking



Ailments Only (with Breed)

The ailments only half of the dataset performed comparatively well. This could be due to the complexity of the different calculations related to health data.

AILMENTS DATASET

Accuracy Score: 0.9896850393700788

Classification Report:

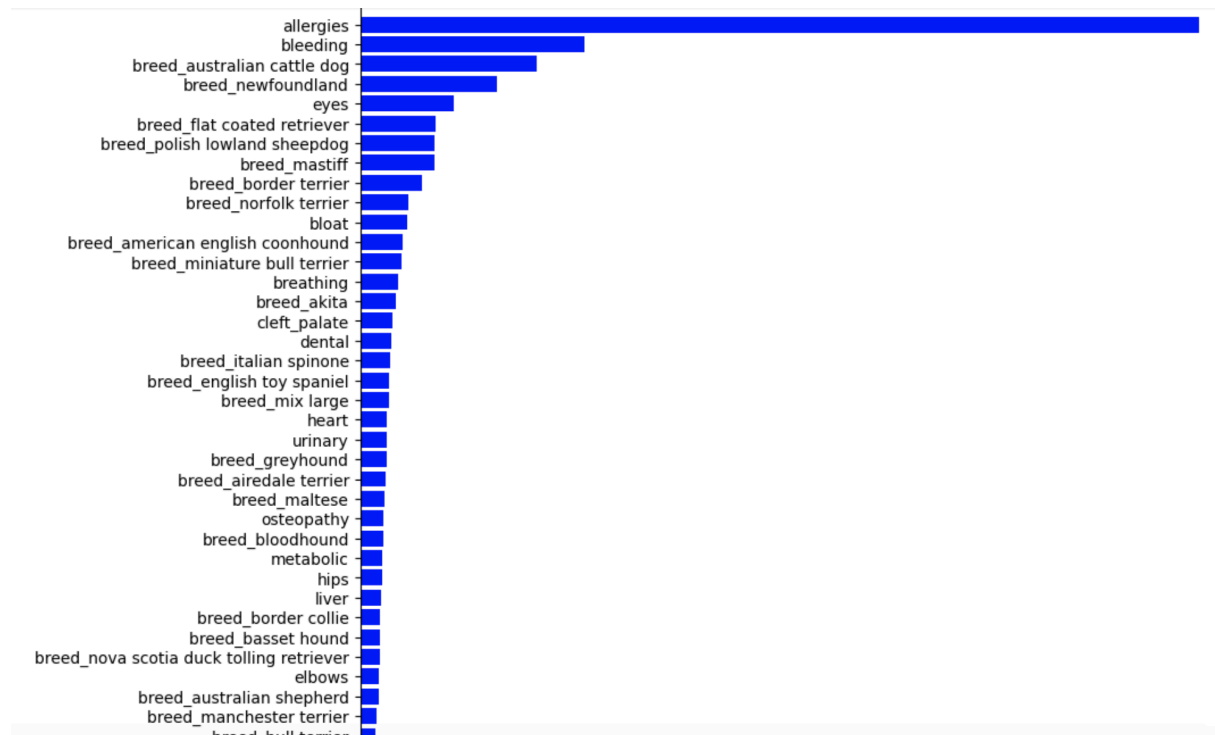
	precision	recall	f1-score	support
2	0.00	0.00	0.00	192
3	1.00	0.82	0.90	1145
4	0.99	1.00	0.99	36763
accuracy			0.99	38100
macro avg	0.66	0.61	0.63	38100
weighted avg	0.98	0.99	0.99	38100

Confusion Matrix:

```
[[ 0  0 192]
 [ 0 944 201]
 [ 0  0 36763]]
```

Model Accuracy: 98.97%

LightGBM (with Breed) Ailments Feature Importance Ranking



Ailments Only (without Breed)

Removing breeds appeared to reduce the overfitting of the model.

AILMENTS DATASET

Accuracy Score: 0.8589501312335958

Classification Report:

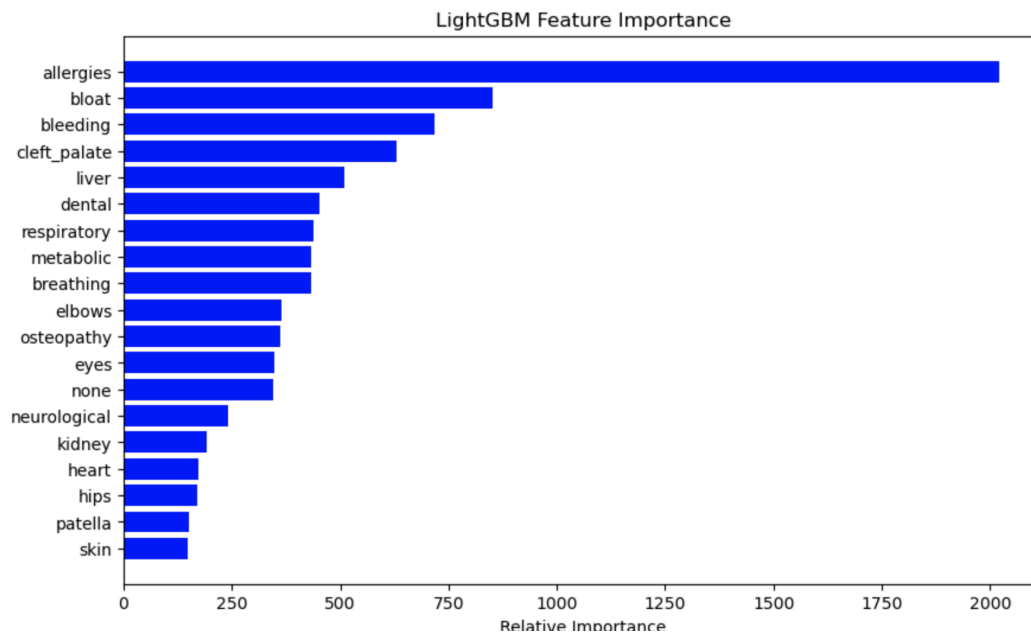
	precision	recall	f1-score	support
2	0.61	0.71	0.65	3366
3	0.49	0.13	0.21	3275
4	0.90	0.95	0.92	31459
accuracy			0.86	38100
macro avg	0.67	0.60	0.60	38100
weighted avg	0.84	0.86	0.84	38100

Confusion Matrix:

```
[[ 2393   74  899]
 [  364  437 2474]
 [ 1187  376 29896]]
```

Model Accuracy: 85.90%

LightGBM (without Breed) Ailments Feature Importance Ranking



Support Vector Machine

Change of Target Variable

In an effort to understand the interrelation of different variables and their influence on 'breed', a final strategy was employed in which the target variable was changed to 'breed' and the SVM algorithm was used. The **clean_split_data.csv** dataset was used, assigning breeds numeric labels instead of one hot encoding. Two approaches were taken: feature importance ranking based on a specific breed and feature importance ranking common to all breeds.

The feature ranking for an individual breed is pictured below:

Feature ranking:

1. Feature 'tendency_to_bark_or_howl' (Weight: 0.06265928754222089)
2. Feature 'grooming_required' (Weight: 0.05698572439392033)
3. Feature 'category_companion' (Weight: 0.055208839980648204)
4. Feature 'intelligence_category' (Weight: 0.04863402542765306)
5. Feature 'hips' (Weight: 0.04390942863089128)
6. Feature 'age' (Weight: 0.00013784036702132324)
7. Feature 'gender' (Weight: 4.6331362080209715e-09)
8. Feature 'potential_for_mouthiness' (Weight: 3.469446951953614e-18)
9. Feature 'osteopathy' (Weight: 0.0)
10. Feature 'longevity' (Weight: 0.0)

The confusion matrix of all breed feature importance follows, along with the classification report. In both cases, the classification reports clearly indicate overfitting, leading us to the conclusion that breed has a low importance in model prediction in relation to the explanatory features.

Confusion Matrix:

```
[[ 300    0    0    0    0    0    0    0    0    0    0]
 [   0 3917    0    0    0    0    0    0    0    0    0]
 [   0    0 5663    0    0    0    0    0    0    0    0]
 [   0    0    0 7655    0    0    0    0    0    0    0]
 [   0    0    0    0 257    0    0    0    0    0    0]
 [   0    0    0    0    0 3579    0    0    0    0    0]
 [   0    0    0    0    0    0 2054    0    0    0    0]
 [   0    0    0    0    0    0    0 3544    0    0    0]
 [   0    0    0    0    0    0    0    0 4180    0    0]
 [   0    0    0    0    0    0    0    0    0 4499    0]
 [   0    0    0    0    0    0    0    0    0    0 2452]]
```

Model Accuracy: 100.00%

Top feature ranking (common to predicting all breeds):

1. Feature 'wanderlust_potential' (Importance: 13.442221909705687)
 2. Feature 'thyroid' (Importance: 12.662010224283588)
 3. Feature 'prey_drive' (Importance: 12.491465372726259)
 4. Feature 'intelligence_category' (Importance: 11.65193807242324)
 5. Feature 'potential_for_mouthiness' (Importance: 11.636687675919866)
 6. Feature 'tendency_to_bark_or_howl' (Importance: 11.468628551244182)
 7. Feature 'dog_friendly' (Importance: 11.12026478100518)
 8. Feature 'tolerates_being_alone' (Importance: 10.862369566127839)
 9. Feature 'size' (Importance: 10.735131079813433)
 10. Feature 'tolerates_cold_weather' (Importance: 10.242371969039608)
 11. Feature 'sensitivity_level' (Importance: 9.41790420423412)
 12. Feature 'longevity' (Importance: 9.393652070283979)
 13. Feature 'energy_level' (Importance: 9.022004070539282)
 14. Feature 'exercise_needs' (Importance: 9.001805284613237)
 15. Feature 'hips' (Importance: 8.433037636016229)
 16. Feature 'none' (Importance: 8.192494553855239)
 17. Feature 'eyes' (Importance: 8.140009618399656)
 18. Feature 'grooming_required' (Importance: 7.974561737652105)
 19. Feature 'heart' (Importance: 7.696262849803286)
 20. Feature 'spine' (Importance: 7.627799903815624)
 21. Feature 'tolerates_hot_weather' (Importance: 7.462840953868363)
-

Conclusion

Since 'breed' was proven to have low importance in model predictions regardless of algorithm and model performance boosting techniques, the switch to 'breed_category' was made. What remains to be seen is whether further variables should be dropped in the continued tuning of our model and the subsequent effect on our business case. Our original problem space was predicting cost categories on the basis of individual breeds, but dog owners are unlikely to be able to choose the correct breed category when using our app. A further point to highlight is mixed breeds, which are further disadvantaged in the case of breed groups as they simply belong to "other" yet make up a significant portion of the dog population.

References

Literature

- [Digital Pet Care Industry: European Market, focus on Germany, France & Bulgaria \(Paw à Peau, 2022\)](#)
- [Increasing adoption rates at animal shelters: a two-phase approach to predict length of stay and optimal shelter allocation \(Bradley & Rajendran, 2021\)](#)
- [Better animal welfare through insurance \(Zenz-Spitzweg, 2021\)](#)
- [The Economic Significance of Pet Ownership in Germany \(Ohr, 2019\)](#)
- [The Economic Significance of Pet Ownership in Germany \(Ohr, 2014\)](#)

Primary Datasets

<https://www.kaggle.com/datasets/paultimothymooney/best-in-show-data-about-dogs>
(foundation)

<https://www.kaggle.com/datasets/jainaru/dog-breeds-ranking-best-to-worst>

<https://www.kaggle.com/datasets/iqamharisfahromi/pet-sales>

<https://www.kaggle.com/datasets/jahangirraina/pet-food-customer-orders-online>

https://data.cityofnewyork.us/Health/DOHMH-Dog-Bite-Data/rsgh-akpg/about_data

<https://www.kaggle.com/datasets/rtatman/animal-bites>

https://www.kaggle.com/datasets/agarwalyashhh/dog-adaptability?select=training_data.csv

<https://www.kaggle.com/datasets/jackdaoud/animal-shelter-analytics>

https://data.bloomington.in.gov/Public-Works/Animal-Shelter-Animals/e245-r9ub/about_data

Supplementary Datasets


(used for enrichment, along with the German pet ownership studies)

<https://rvc-repository.worktribe.com/output/1375552>

<https://data.world/the-pudding/adoptable-dogs-on-petfinder-in-the-us/workspace/file?filename=alIDogDescriptions.csv>

<https://rvc-repository.worktribe.com/output/1375689>

<https://rvc-repository.worktribe.com/output/1558210>

 Pet health pricing model