Predicting pet adopting speed using machine learning models

Abstract

Millions of stray animals suffer on the streets or are euthanized in shelters every day around the world. If homes can be found for them, many precious lives can be saved — and more happy families created [6]. Petfinder.com have thousands of pet adoption postings. The paper proposed to use machine learning models such as random forest, gradient boosting and supported vector machine to predict adoption speed of a pet that is listed. Multiple trials of model training were performed on different subset of the data. The random forest is the best performing model. Feature engineering and feature selection were applied to analyze features. Age and color are the only two features of pet that are important to the adoption speed. Type, gender, health status and vaccination etc. are weakly considered when people decide on adopting a pet. The results also show that changes in the number of photos, the quality of the photos, and the number of words in the description can facilitate the adoption speed. In the future, the project can expand the analysis in images and text to establish robust metrics that better evaluate photo and description quality of a listing.

Introduction

Data source

The dataset was acquired from Kaggle. It contains information about pets listed in PetFinder.com Malaysia site. Each observation is an online profile of a pet. The features include an animal's name, type, age, breed, color, gender, fur length, size, status of denormalization, sterilization, health and vaccination, the number of photos and videos attached, adoption fee, location and a description written by the owner. The target variable is the adoption speed. Additionally, the data source also provided image metadata and sentiment metadata that was generated by Google Vision API and saved in Json files. The image metadata introduces more details on image properties such as color distribution and picture size. The sentiment metadata provides details on the magnitude of the description and the sentiment score for the description.

Dataset

The project was performed on the dataset that added the engineered features. The full set has 24 features containing both categorical, numerical and ordinal data types. The target variable--adoption speed has five classes. There are 11, 805 observations. The full description of the features is listed in the appendix list 1.

Methodology

There are five steps in the process: exploratory analysis, preprocessing, feature engineering, feature selection, model training and results analysis.

Preprocessing

Since the dataset contains data types that are not meaningful for statistical analysis these variables are RescuerID, description, color 1, color 2, color 3, breed 1, breed 2, name. To incorporate their presence in the models, preprocessing was done on these variables. 1) The distribution of pets' age is right skewed. Applying logarithmic transformation, a normal distribution was preserved. 2) breed 1 and breed 2 account for the breed types of pets. For each pet type there are more than 300 breed types. A binary variable "Mixed" was created to combine the values in breed 1 and breed 2. If breed 1 and breed 2 share different values, then the pet is mixed. If breed 2 is zero, then the value for mixed is one. 3) Has_name is the other binary variable checks whether a pet has a name in the profile. 4) a count function was applied on colors, Rescuer ID and description. If a pet has three values in all color 1, color 2 and color 3 column, then three is recorded for colors variable. There are rescuers that have multiple listings, which indicates the pet is from an adoption agency. Count of listings gives information about the source of the pet and the size of the rescuer. Description length counts the number of words in the description write up. After the preprocessing, Rescuer ID and description etc. were removed from the dataset.

Feature engineering

Feature engineering involves the selection of a subset of informative features and/or the combination of distinct features into new features in order to obtain a representation that enables classification [1]. Based on the zipcodes of the states, information about states' GDP, population, area, birth rate and the unemployment rate were attached to the original set of features aiming to discover patterns in regard to state and listed features. Image metadata and sentiment metadata which provide additional information about photos and description were also incorporated into the dataset.

Metadata analysis

Figure 1 in appendix is the correlation plot of added features. Vertex_x, vertex_y, dominant_pixel_frac, bounding_confidence are higher correlated. The distributions of these variables are skewed as well. Combining with five-number summary results, it is decided to only keep dominant score as a measure for photo quality and sentiment score, sentiment magnitude as a measure for description. Therefore, the final set of features has a total of 24 variables.

Feature selection

Low variance filter method, wrapper selection, univariate selection and feature importance selection methods were applied to perform feature selection for the dataset. The results of the methods share agreement in seven features: image score, the number of photos associated with the listing, sentiment score, length of the description, colors and age of the pet, and count of listings the rescuer has. The analysis of selected features is presented in the discussion section.

Machine learning models

The dataset was split into 70% of training set and 30% of test set. Five-fold cross validation was also applied to test model consistency. Parameter tuning was applied in training of random forest and gradient boosting to select optimal number of estimations, grading criteria, the depth of the trees, minimum sample split and learning rate etc. Grid search was applied to select the best set of parameters of supported vector machine. The metrics that evaluate the model performance are: accuracy on the test set, accuracy on the cross validation sets, weighted F1 score and run time of the model.
Random forest, gradient boosting, supported vector were built on full set of features, selected bases features and adding engineered features to the selected set.

Results

Accuracy and selected features

Table 1 Training on base feature

Final models of base features				
With Feature selection (6/17)				All features
	Random forest	Gradient boosting	SVM	Random forest
Test accuracy	**0.392**	0.388	0.35	0.431

Weighted avg. f1	**0.38**	0.37	0.33	0.42
Cross validation	**0.38 (+/- 0.02)**	0.40 (+/- 0.01)	0.35 (+/- 0.01)	0.43 (+/- 0.01)
Run time	**6.012**	8.893	32.582	7.7
Selected features	PhotoAmt, Count_of_listings, Description_length, Colors, Log_Age			

Table 2 Adding engineered features

Final models of adding engineered features				
	With Feature selection (7/24)			All features
	Random forest	Gradient boosting	SVM	Random forest
Test accuracy	**0.397**	0.384	0.325	0.428
Weighted avg. f1	**0.38**	0.36	0.24	0.41
Cross validation	**0.39 (+/- 0.01)**	0.39 (+/- 0.01)	0.33 (+/- 0.01)	0.44 (+/- 0.01)
Run time	**8.493**	11.412	29.491	8.54
Selected features	PhotoAmt, Count_of_listings, Colors, Log_Age, SentMagnitude, SentScore, dominant_score			

Table 1 shows the results of models that were built on all base features and selected base features. Table 2 shows the results of models that were built on adding engineered features to the base features and the selected features. Although, for both trials the accuracy from feature selection is reduced compared to all feature models. It is preferred to have feature selection because more than half of the features were removed which increased the interpretability of the model. Meanwhile, the model run time of selected features is shorter than models that were trained on all features. To compare across three types of models, random forest and gradient boosting generated comparable results. Random forest outperformed in run time. SVM with rbf kernel not only has the lowest accuracy, but the run time is at least three times longer than random forest. As the support vector classifier works by putting data points, above and below the classifying hyper plane there is no probabilistic explanation for the classification. SVM algorithm is not suitable for large data sets. The unbalanced distribution of target classes might cause underperformance [3]. Overall, the best preforming model is random forest for this dataset.

Table 3 Reducing target classes

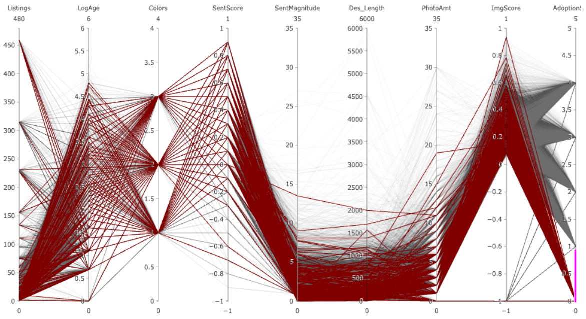
Final models of removing no adoption observations with feature selection (7/24)			
	Random forest	Gradient boosting	SVM
Test accuracy	0.404	0.387	0.368
Weighted avg. f1	0.42	0.37	0.24
Cross validation	0.37 (+/- 0.02)	0.41 (+/- 0.02)	0.41 (+/- 0.02)
Run time	6.284	6.702	6.507
Selected features	PhotoAmt, Count_of_listings, Log_Age, SentMagnitude, SentScore, dominant_score		

The accuracy results from models classifying five target classes are not satisfying. An experiment was done on removing no adoption observations and allow models to train on adopted cases only. An improvement of accuracy was expected. However, models achieved similarly with the above results. Since the project was also designed to understand what features are important to making classifications and the neural networks have black box issues, more advanced models were not experimented. In the future, neural networks can be trained to increase performance. The table 1 in the appendix has results of predicting binary target. The models achieved 77% of accuracy in predicting whether a pet will be adopted or not.

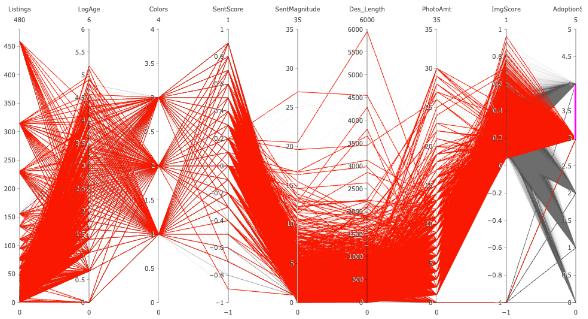
All models fitted on adding engineered features are consistent in selecting the same set of features, which means all the selected features are important in predictive models.

Discussion

1. *Feature importance*: List 2 in appendix shows feature importance ranking. Because the nature of the task is classification. Feature importance score is served as a major criterion for evaluating features. With this method, it is possible to select the feature subset in which the features are most beneficial to the subsequent tasks while the redundancy among them is minimal [2]. The features that relate to owners of the listings have higher rankings and are all selected. Features that describe a pet ranked lower. Information about the state is not significant. It reveals that things that pet owner can do to make the profiles more appealing are more important than the features of pets.
2. *Feature selection*: Two engineered features were added to the original selected set of features. Those two barely make contributions to the performance, yet they are important in making predictions. It opened up opportunities to further analyze on images and sentiment to create more robust evaluation metrics.
3. *Feature analysis*: Parallel plots were created with only the selected features to visualize some distinctions. They compare the two extremes. Adopted on the same day or within the first week versus Adopted after the 3rd month or no adoption. The biggest differences are the length of the description the number of photos, and the variation in the count of listings. Pets that were adopted faster are younger. Their profiles have number of photos and description length less than average.



Adopted on the same day or within the first week



Adopted after the 3rd month or no adoption

From the analysis, it is advised to send the pet to an agency or post less than 10 higher quality photos and write about 300 words description in order to speed up the adoption. The goal is to create a profile that is appealing to the audience.

Future work

1. *Feature extraction:* Exploratory Factor Analysis is a method for identifying the factor structure of a set of multiple indicators or variables without imposing an a priori structure on the factors. EFA is performed to consolidate variables and generate new hypotheses about underlying theoretical processes [4]. Using Factor analysis to create factors that describe pet owners, pets, and photo quality and description quality could improve the prediction accuracy.
2. *Image analysis.* The image score is proven important in making predictions. But the Google Vision API generated ones are not proficient. The results of image analysis would be evaluation metrics that measures the quality of photos in order to improve prediction results.
3. *Text mining.* Similar words occurring quite often in the text said to have little information to distinguish different documents and also words occurring very rarely are also possibly of no

significant relevance and can be removed from the documents [5]. The results of text mining are expected to provide a score to measure the description and weights of certain words. So that rescuers will have additional aid in constructing description for their pets.

Conclusions

The project constructed random forest, gradient boosting and supported vector machines on different facets of the data. Results among models from all features, subset of the features, and subset of the target variables were compared. The final model for the problem is random forest with 200 estimators using entropy scoring. It delivered an accuracy of 40% with a run time of 6 seconds. And cross validation results prove the model performed consistently. The chosen models failed to generate a higher accuracy rate. In the future, more advanced models such as neural networks can be tested to improve the performance metrics. The findings suggest that a pet owner can consider adoption agency because pets from the agencies were adopted faster. The younger the pet, the easier it gets adopted. Posting 10 high quality photos of the pet and write a 300 words description of the pet would also increase the adoption speed. It is not suggested to post more than 10 photos or write really long post. The improvements in evaluation metrics of photo and sentiment are expected for the future work. The results of these work can improve predictive performance of the models.

References

1. Garla, V. N., & Brandt, C. (2012). Ontology-guided feature engineering for clinical text classification. *Journal of biomedical informatics*, *45*(5), 992-998.
2. Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, *282*, 111-135.
3. Noble, W. S. (2006). What is a support vector machine?. *Nature biotechnology*, *24*(12), 1565
4. Reio Jr, T. G., & Shuck, B. (2015). Exploratory factor analysis: implications for theory, research, and practice. *Advances in Developing Human Resources*, *17*(1), 12-25.
5. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.
6. https://www.kaggle.com/c/petfinder-adoption-prediction/overview

Appendix

List 1: list of features

Type	Name	Values
Binary	Gender	[0,1]
	Sterilized	
	Dewormed	
	Vaccinated	
	Mixed	
	Type	
	Has_name	
Ordinal	FurLength	1=short;2=medium; 3=long;0=not specified
	MaturitySize	1=short; 2=medium; 3=large; 4=extra large; 5=not specified
	Health	1=health; 2=minor injury,3=serious injury;0=not specified
	Colors	[1,3]
Continuous	dominant_score	[-1,1]
	SentMagnitude	[0, 32]
	Count_of_listings	[1, 459]
	Log_Age	[0, 5.55]
	SentScore	[-0.9, 0.9]
	PhotoAmt	[0, 30]]
	Fee	[0,2000]
	state_population	[0.87, 54.62]
	state_unemployment	[0.9, 7.8]
	state_gdp	[5.98, 280]
	state_area	[91, 124450]
	state_birth_rate	[0.127, 0.233]
	VideoAmt	[0, 6]
Target	0: The same day as it was listed 1: Between the first week 2: Between the first month 3: Between the 2nd and 3rd month 4: no adoption	

Figure 1: Correlation plot of metadata

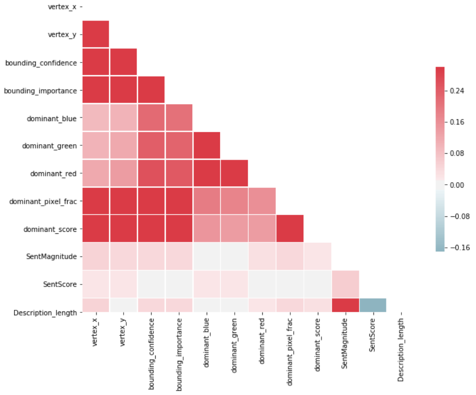


Table 1: Binarizing target

Final models of binary target with feature selection (7/24)			
	Random forest	Gradient boosting	SVM
Test accuracy	0.743	0.736	0.734
Weighted avg. f1	0.72	0.68	0.62
Cross validation	0.74 (+/- 0.02)	0.74 (+/- 0.02)	0.74 (+/- 0.02)
Run time	6.532	6.747	6.475
Selected features	PhotoAmt, Count_of_listings, Log_Age, SentMagnitude, SentScore, dominant_score		

Target distribution (binary target)	
No adoption	0.266
Adoption	0.733

List 2: Ranking of features by importance scores

Feature Importance Ranking		
Rank		Score
0	dominant_score	0.12347211
1	SentMagnitude	0.11038654
2	Count_of_listings	0.10772727
3	Log_Age	0.09537489
4	SentScore	0.08261576
5	PhotoAmt	0.08129957
6	Colors	0.04465272
7	FurLength	0.03516391
8	MaturitySize	0.03255325
9	Fee	0.02865666
10	Gender	0.0284679
11	Sterilized	0.02801049
12	Dewormed	0.02665842
13	Vaccinated	0.02649836
14	Mixed	0.02292765
15	Type	0.02079622
16	state_population	0.01764697
17	state_unemployment	0.01727905
18	state_gdp	0.0166564
19	state_area	0.01664556
20	state_birth_rate	0.01472949
21	Has_name	0.008321
22	VideoAmt	0.00700268
23	Health	0.00645714