A Home for All

AUTHOR:	Caitlin Jansson, CaitlinJEJansson@gmail.com
DATASET:	https://www.kaggle.com/c/petfinder-adoption-prediction
CODE:	https://github.com/CJEJansson/A-Home-for-All
SLIDES:	Final Presentation Slides

ABSTRACT

A local animal shelter has requested that a methodology be developed to help them more strategically process animals and ultimately improve adoption outcomes for their current residents. By leveraging the Kaggle Petfinder data and analyzing the trends between dog's features and adoption outcome, a set of models were developed. These models can be used by the client to develop a specific strategy for each type of dog and ensure the funds needed to support an animal finding a home are best leveraged. These models revealed that it's easy to predict which animals are adopted in under 90 days. Those animals that take longer to adopt can be sent to intensive training, or resituated with a better positioned rescue to ensure their placement in the correct "Fur-ever" Home.

Models used to analyze the data and accuracy scores are summarized in the table below:

Model	Accuracy Score			
	Same Day		No Same Day	
	Training	Validation	Training	Validation
Logistic Regression- lbfgs, restricted features *	0.39208	0.37146	0.40754	0.38645
Logstic Regression - lbfgs, all features	0.38579	0.39237	0.41541	0.37516
Logistic Regression - liblinear, all features	0.38579	0.39483	0.40369	0.37641
Naïve Bayes	0.23880	0.26568	0.33626	0.34003
Linear SVC	0.24918	0.36162	0.27247	0.24467
SVC	0.39071	0.40590	0.42099	0.44668
kNN	0.36885	0.39606	0.39698	0.41656
Random Forest	0.46940	0.47109	0.47571	0.48557
Light GBM	0.45027	0.46740	0.48130	0.47679

* Features Included: Rescuer ID, pet age, first breed classification, amount of photos

** Kaggle PetFinder Leadboard Winning Accuracy 0.45338

PROBLEM STATEMENT

There are more potential pets than pet owners, and shelters across the country are struggling to keep up. But what if there were a way to increase the likelihood that an animal gets adopted? Does the name matter? The color? Is one breed more likely to be adopted than another? Do any of these things influence the speed at which a pet is adopted? Will understanding historic data improve a shelter's ability to find an animal placement in a "forever home?"

Local area shelters cannot help but feel discouraged. The problem is that intake capacity is limited and the number of animals that need help just keeps growing! A local shelter specializing in large breed dogs has asked that some research be done to try and improve their adoption rate. Are better pictures needed? Do they need to "rebrand" the animals by listing a different primary breed? Or is a simple name change going to help? Which animals need to go to all the adoption events and which will be adopted simply because they're listed? While they can utilize strategic "marketing" techniques for each animal, ultimately the animal has to stand on its own characteristics to succeed. Telling an interested individual that an animal is high energy, only to have that animal be returned 3 months later for being destructive is both heartbreaking and frustrating.

A U.S. based client has stated that they want to start focusing their energies in a more targeted manner. Which pets are going to be easily adopted out, and which will be the problem puppies? If they better understand the systemic trends based on the dog's identifying features, they can make more strategic decisions in directing funds and time to get animals face time with adopters. They will know who to take to adoption events, and they can better help those animals in their care. Those animals that will be long term with the rescue can be sent to more in depth training courses, and possibly moved to another rescue where the market is better for that specific breed or temperament of pet. The intent of this project is to develop a model to analyze an incoming animal and define an individualized approach to success.

GENERAL OVERVIEW OF THE DATASET

The data is provided from the Kaggle.com PetFinder competition, targeted at increasing adoption speeds for homeless pets. It's divided into a test and train dataset and contains information on over 10,000 animals, including cats and dogs. The data is pulled from the Malaysian region, and is categorized by Malaysian state. Feature data is a combination of ordinal features and descriptive fields (name, description of animal). Those fields that have been converted to an ordinal format have separated descriptive data in an independent table for each feature.

The data was relatively clean, with the exception of the "Name" and "Description" columns, which will be addressed later. The first step was to remove all data on cats from the dataset, as the customer is only concerned with the dogs at this time. It was verified that there are no animals listed as being a cat and having a corresponding breed classification indicating that it is a dog.

Any values that are null in breed occur only in Breed 2 or Breed 3 column, indicating that a dog is listed as being a purebred dog. Null values were filled with zeros to balance out the data set for later use, and a row for 0 was entered into the breed table with the value "No Breed". An additional column was created to account for mixed vs. purebred listings called "BreedCount", and counts the number of breeds listed for each animal. Mixed breeds have a "BreedCount" value greater than one, as expected. The same process was applied to the animal's colors, and the creation of a column called "ColorCount."

Two rows in the Description column were found to have null values, these were originally included. However, further investigation showed that there is a lot of unclean data in the description column. As the final intent of this project will not include Natural Language Processing, the null description values were filled with “No Description”. The intent going forward will be to leave the description column largely unaddressed. This is specifically due to the request of the client to focus only on animal feature data.

After investigation of the "Name" column, in the attempt to address Null values, it was determined that just because an animal had a value in the name column, did not mean that the animal actually had a name. Due to the size and extent of the problem, name's were manually cleaned using excel. Some of the following are examples of name entries that were found and were replaced with a null include, but are not limited to: “Lost Dog”, “Cute Puppies”, “Boy”, “Girl”, Miscellaneous breed names (labrador, terrier, etc.), “No Name Yet”, “Please Name Me”, “Save me or I'll Die”, “Urgent home needed”, “Puppy”, and various descriptions of puppies (happy puppy, big eyes, sad puppy, bouncy puppy, etc.) An additional feature was added to convert the presence/absence of a name into an ordinal value, with 1 indicating presence of a name.

Animals that are actually unnamed were left as null values. This allowed for the creation of an additional column “Name_bool”, used to indicate whether an animal had a name or not. The value 0 was filled for no name, and 1 indicated that the animal had a name.

For final analysis, the RescuerID field was also converted into an ordinal, with each unique Rescuer facility being assigned a random integer. For machine learning purposes the index, description, text RescuerID, and individually unique pet ID assigned by the system were dropped. The description was omitted because it will not be used for analysis, and the others were omitted because they contain no pet-specific information to help distinguish one listing from another.

INITIAL FINDINGS

The data was then explored via Exploratory Data Analysis (EDA) and analyzed using inferential statistics. Knowing that the majority of the dataset is categorical (discrete) data, Chi-squared testing was heavily used to explore the relationships between features. Additionally, Cramer's V technique was used to generate a heat map to show correlation between variables[1]. A complete summary of statistical results can be found in Appendix A.

It is known from animal rescue work in the United States that dogs with the least likely chance of adoption are senior dogs, dogs with health problems, and dogs that have black fur. Also, the breeds categorized as “aggressive,” pitbull breeds, german shepherds, dobermans, etc. have a lower chance of adoption. Overall, approximately 50% of American animals (approx. 1.2 million) go unadopted, annually. It is suspected that this is also the case in dog rescue globally.

When looking at the Kaggle dataset it quickly becomes apparent that a large number of dogs are also going unadopted, but not to the severity in the U.S. Approximately 30% of dogs go either unadopted or unreported as being adopted (Fig. 1, Page 4). As expected, a very low number of dogs are adopted on the same day they are listed, at 2.08% Of these animals, the younger dogs are adopted more quickly than the senior, which is consistent with expectations. Statistical testing confirms that there is a relationship between age and adoption speed at a significance level of 95%.

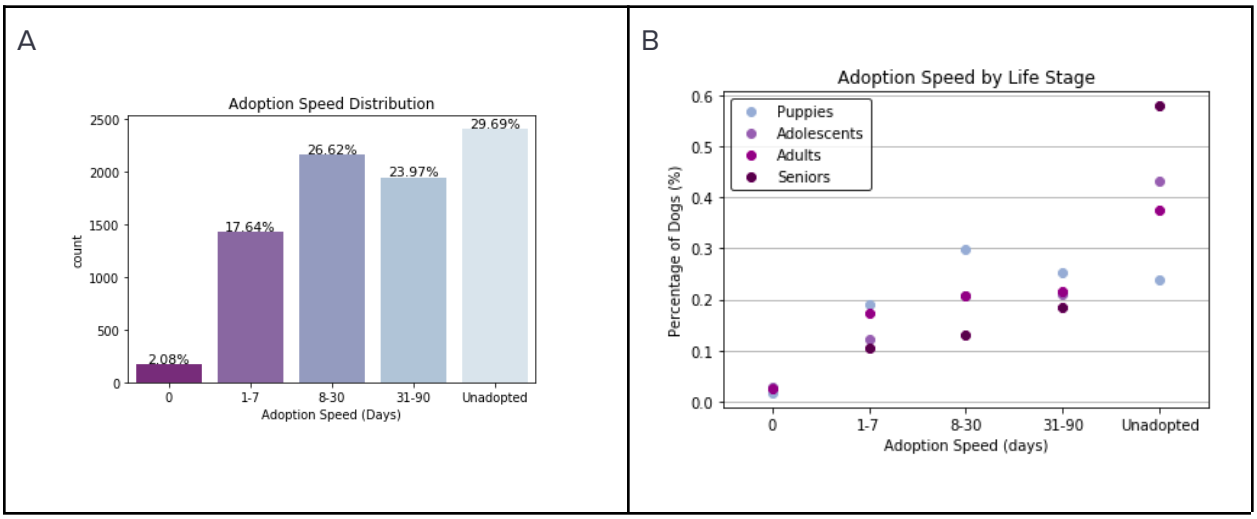
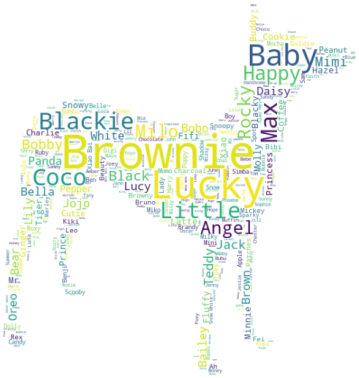


FIGURE 1: a) Distribution of adoption speeds for dogs across data set. b) Scatterplot showing adoption speed distributions by life stage

When it comes to names, dogs that have already been given a name tend to be adopted more quickly in the U.S. It's possible that the presence of a name indicates the dog's history is known, and it gives a sense of personality. First, the names listed in the dataset were explored to see what were the most popular names, and if they had any bearing on adoption speed. Most often, the color of the animal was its name, with the most popular names in this dataset being "Blackie", "Brownie", "Baby" and "Lucky". Statistical testing showed that the specific name of an animal is independent of adoption speed. However, the presence of a name did matter, and tested as being statistically related to adoption speed (Fig. 3)



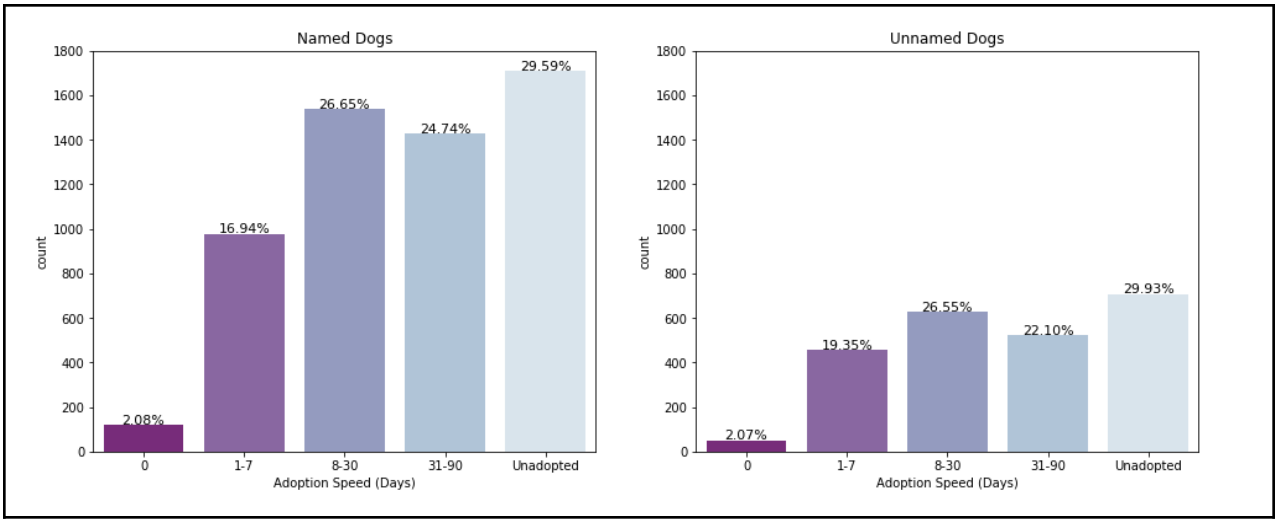
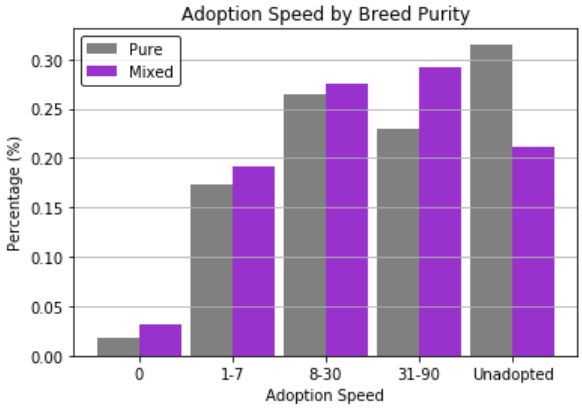


FIGURE 3: Adoption distribution of named (left) vs unnamed (right) dogs

When it comes to breed choices in the U.S., purebred dogs tend to be more popular. Particularly when it comes to labradors and other hunting dogs. The data set seems to demonstrate that in Malaysia this is not the case, with mixed breed dogs being more popular (image to right). This may be due to the availability of mixed breed dogs being higher. However, labradors are just as popular as they are here in the U.S. Shih Tzu's and Terriers are the second and third most popular dogs. (Fig. 4, pg.4) Statistical tests confirm that there is a relationship between breed and adoption speed, and between mixed/purebred and adoption speed. However, the breed has more of an impact than whether a dog is mixed.



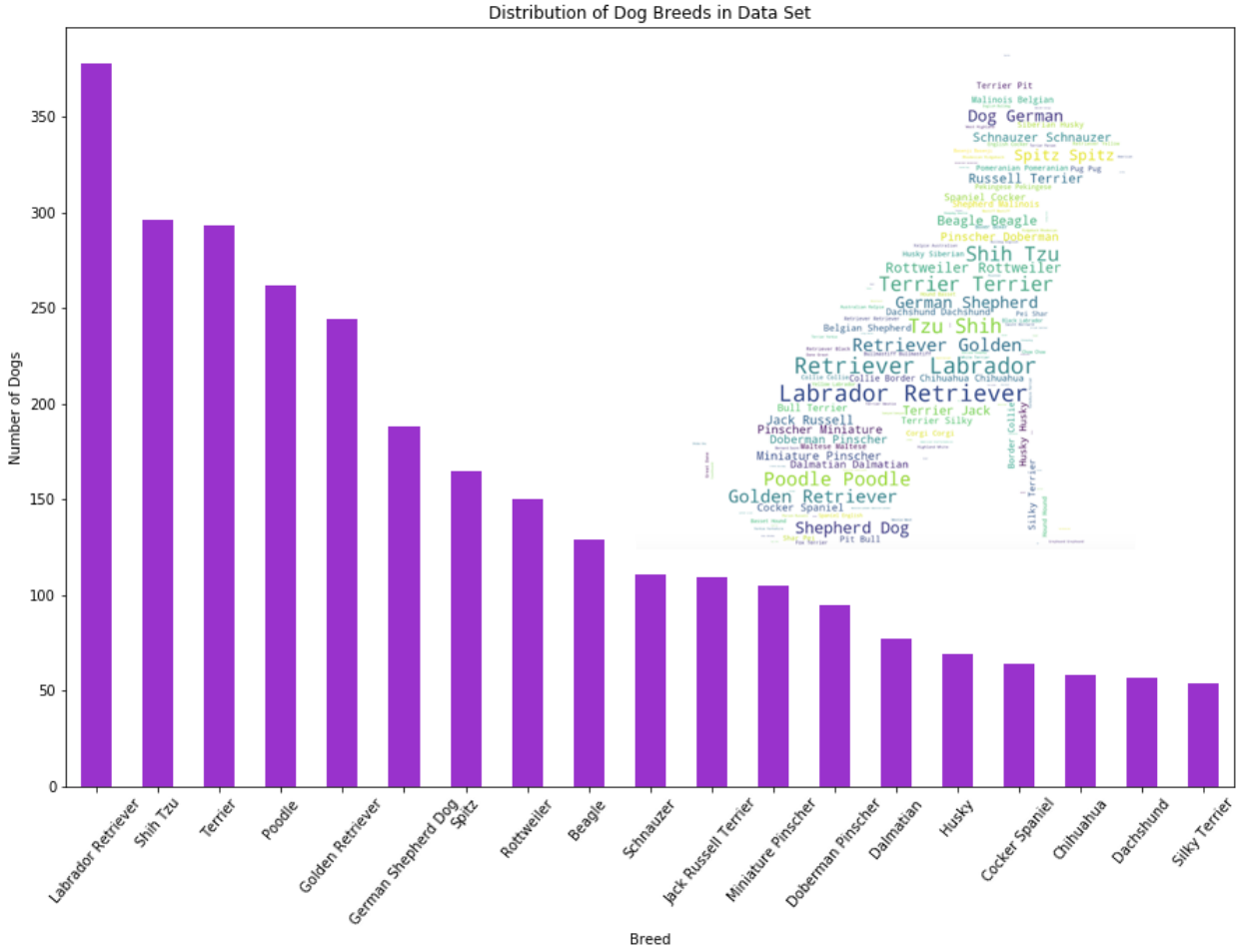


FIGURE 4: Distribution of dog breeds in data set and corresponding word map

The size of the dog also plays a role here in the U.S., with medium size breeds being more popular. In Malaysia, giant breeds are very quickly adopted, likely due to their scarcity, as they make up 0.22% of the data set. Otherwise, size has very little effect on adoption speed in Malaysia. Statistical tests confirm this relationship (Fig. 5).

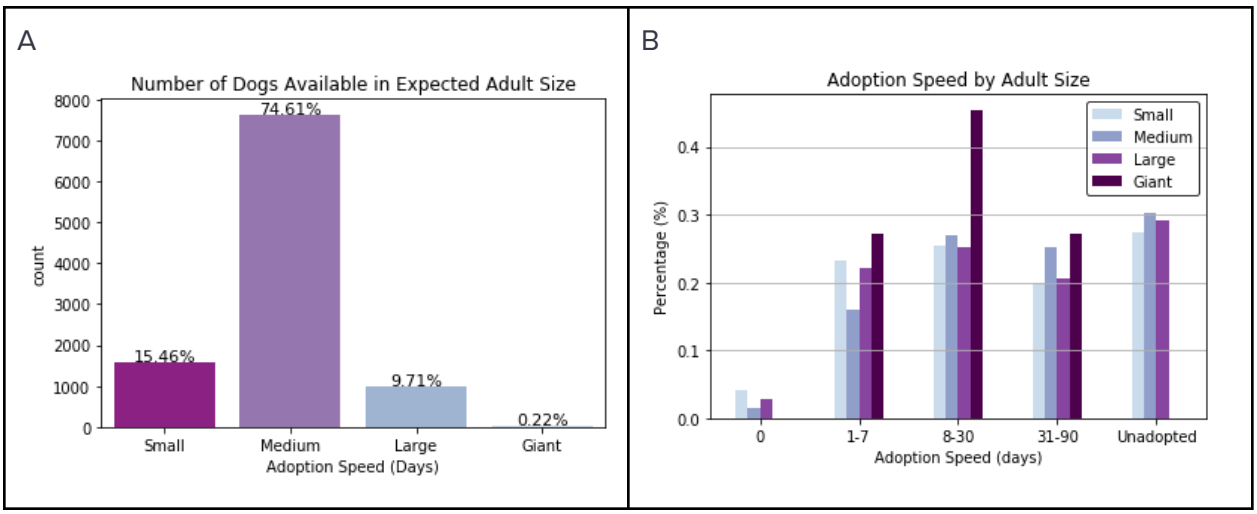


FIGURE 5: a) Distribution of dogs by adult size, b) distributions of adoption speed by adult size

When it comes to appearance, in the U.S., there's definitely an impact. Fur length is more of a factor in certain regions, but overall doesn't make much of a difference in adoption speeds. In the dataset, it appears that dogs trend towards shorter coats. While there is some difference between the distributions (Fig. 6), statistical testing shows that between fur length and adoption speed are not highly correlated but there is a relationship between the features.

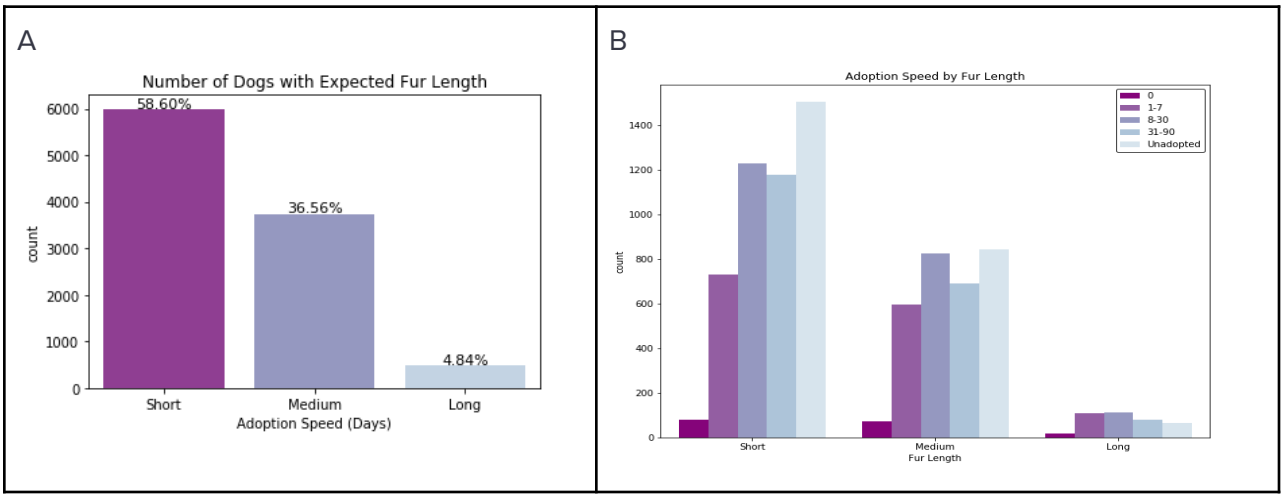


FIGURE 6: a) Distribution of dogs by fur length, b) distribution of adoption speeds by fur length

While fur length doesn't make much of a difference in the dog's adoption speed in the US, the color does. There's a well known phenomenon in the U.S. adoption groups that indicates black, or darker colored dogs go largely unadopted. This is thought to be because they appear intimidating, and they also do not photograph well. In Malaysia fur length also has very little effect on adoption speed. However, as expected, darker color dogs go unadopted more often. This may in part be due to the large number of dark colored dogs that are available. Additional investigation would be needed for conclusive evidence, but it appears the trend holds (Fig. 7).

One interesting observation of note is that golden colored dogs and yellow colored dogs are separated into two groups. At first glance it appears that yellow dogs are also very unlikely to be adopted, but when combined with golden dogs this trend disappears.

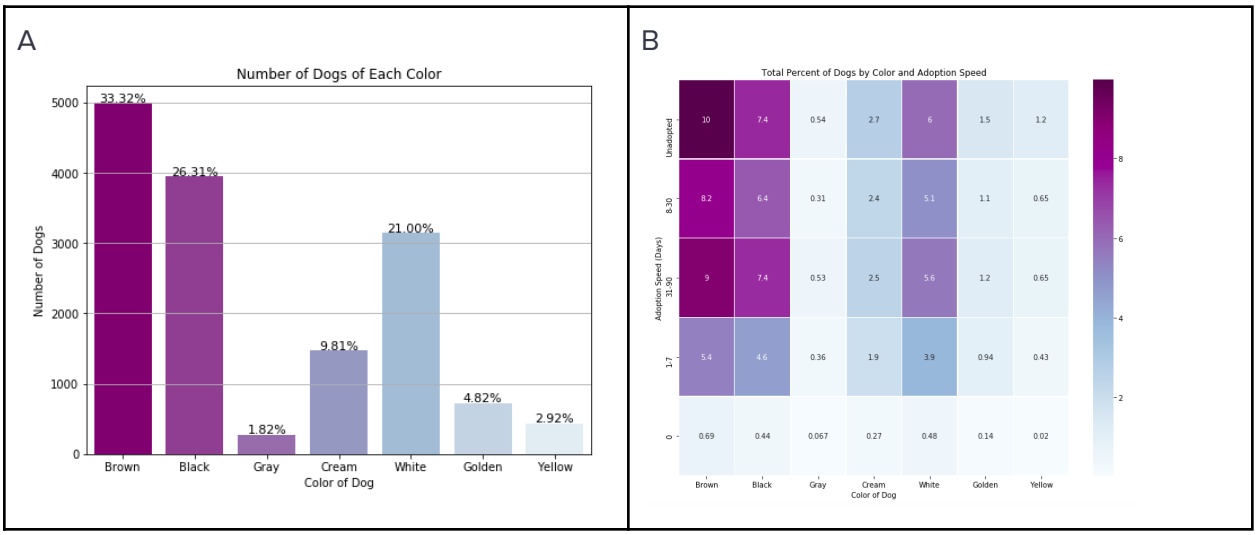


FIGURE 7: a) Distribution of dogs by color, b) heat map of percentage of dogs in each adoption speed category

Also investigated was the number of colors in a dogs coat. There were no visible trends during EDA, but there is a slight correlation between this and adoption speed.

As with dogs in the U.S., the way the listing profile is set up has an effect on the speed with which an animal will be adopted. The presence of photos and as much information as possible does make a noticeable difference here. This trend holds true in Malaysia. Also of note, the presence of a video makes little difference in adoption speed (Fig. 8).

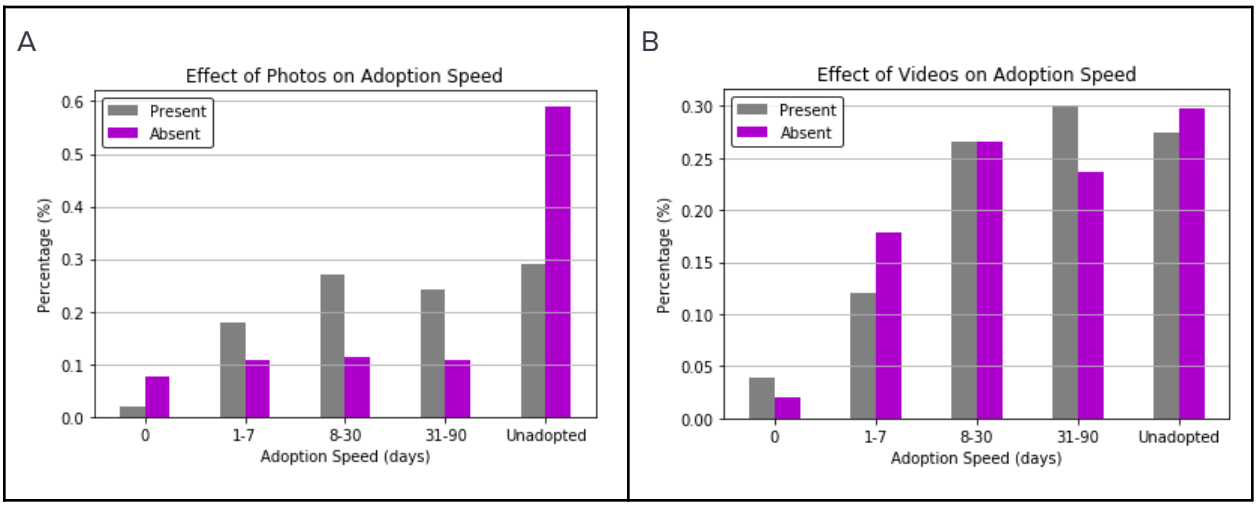
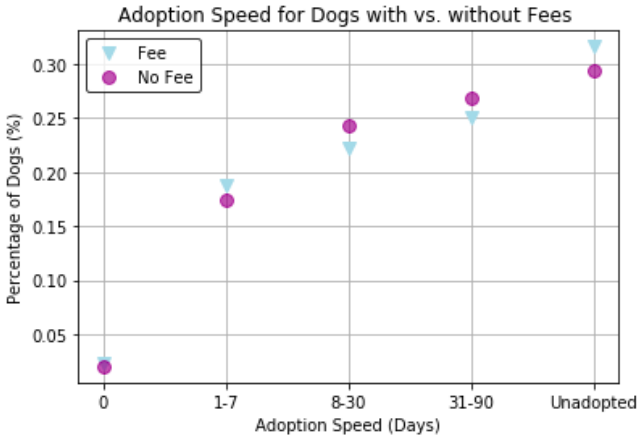


FIGURE 8: a) Distribution of adoption speed with and without photos, b) distribution of adoption speeds with and without videos

The cost of a rescue is often a contentious subject when adopting a dog. If an animal is from a particular breeder and is sold as a puppy, it's often no issue. The perception in the US is that a rescue is not as "high class", and should thus be cheaper. It was expected that the more expensive a dog is listed as, the less likely they are to be adopted. However, the opposite was found to be true when examining the data (see figure on page 8). This is potentially due to the



fact that a specific type of dog or a purebred dog can be listed at a higher price and many people consider them to be more desirable. While cost is not highly correlated with adoption speed, statistical testing does show it had some impact.

Investigation showed that listings with more than one dog are less likely to be adopted. It's possible that listings with larger numbers of dogs that do not require co-adoption (for litters vs adopting siblings together) are obscured because there is no way to record adoption speed for each animal, and the longest adoption time is all that's captured. Statistical testing was not performed on this feature due to the inability to perform correlation on a combination of continuous and discrete data.

In the U.S. the health of an animal plays a large role in adoption speed. To the point that almost all shelters and rescues guarantee the dog will be sterilized (spayed/neutered), up to date on vaccinations, and dewormed. Healthy dogs also tend to be adopted faster as opposed to those classified as "special needs", or those dogs that have a known injury or illness. In the dataset, it became apparent that dogs that were unsterilized, not dewormed, and unvaccinated were fastest adopted (Appendix A, "General Health Plots"). More in-depth investigation showed that this is because these dogs were typically puppies, and puppies are adopted fastest and have the least amount of pre-adoption veterinary care. After removing puppies from the data set the trends more closely mirrored those in the U.S. As expected, dogs that are unhealthy or have a known injury/illness are less likely to be adopted. The one trend that was not particularly elucidating was that of the worming status, with a large percentage of dogs not being dewormed. This could be due to the fact that worm treatment is only being performed when necessary, but the data is not clear as to whether the treatment is preventative or in response to infection. These relationships were confirmed further during statistical testing, which used the entire dataset, rather than removing the puppies.

IN DEPTH ANALYSIS VIA MACHINE LEARNING

It was anticipated prior to beginning analysis that the model would have difficulty predicting same day adoptions as this is generally difficult to predict in a real life setting. The majority of the data chosen was categorical or non-continuous data, so that limited the model choices somewhat.

The first step involved converting some of the features to ordinal classes, to allow for application of specific models and eliminate difficulty of analyzing datasets with mixed data. The two variables specifically targeted were Name and Rescuer ID. Name was converted into a binary presence/absence, and Resucer ID was converted into randomly unique integers by ID. Then all data not being used was dropped, which included that data that was either converted or would be non-useful in analysis. This included the features Name (converted), Rescuer ID (converted), the index column, randomly assigned Pet ID, and the description, which was not used in this analysis.

Then the data was split using training, testing, and validation sets. The first split was done to segregate 10% of the data for later model validation. The remaining data was split into 75% training, 25% test sets, and stratified using the Adoption Speed feature. Stratification was chosen to handle the amount of imbalance between categories.

Logistic Regression

While the model showed some linear behavior it wasn't definitively linear. Since the statistical analysis did show some linearity between certain variables, it was decided that trying a linear model on the data set would be worth trying.

Analysis during development showed that a model using only those pets who were not adopted on the same day as listing and access to all features performed best. However, when testing the training and validation data, it was found that the best model was to use all the features with a liblinear solver, and to keep the "same day adoptions". There is likely a large component of overfitting here, based on the behavior seen. However, the increased accuracy could also be due to simply having access to more data. Final accuracy score for all data was 0.39483.

Naive Bayes

The next model utilized was Naive Bayes. Multinomial Naive Bayes was used because the bulk of the data is categorical, and it tends to not be normally distributed. Gaussian was considered but because of the mix of variables, it was decided that the best choice would be multinomial. The benefit of using Naive Bayes is that it assumes the independence of each feature in the data set, and will hopefully find any instances where a particular feature has more weight than the others. The accuracy of the Naive Bayes model was considerably less than that of the logistic regression model, with a score of 0.23333. However, when removing the "same day" adopted pets, the score increased dramatically to 0.35120. Pursuing hyperparameters again decreased the score, so it was decided that Naive Bayes was probably not going to be the best model for this dataset. A test on the entire data set, including validation data, yielded a score of 0.31242. Since we had good success in predicting across all classes for adoption speed, it would be good to further pursue classification algorithms.

Support Vector Machines (SVM)

Since Naive Bayes was successful, it would probably be beneficial to investigate further something that utilizes a categorization approach. Enter support vector machines (SVM) Two types of support vector machines were investigated, both Linear and Nonlinear SVC . SVC was used, rather than SVM, because this is a multiclass problem.

Starting with Linear SVC, the default argument for multiclass was used. This model did not yield good results. The classification was restricted entirely to 1-7 day adoptions, and the accuracy was the lowest so far, at 0.17705. When dropping pets adopted the same day, accuracy increased dramatically, but the confusion matrix still showed that the model was not classifying correctly. The best score for Linear SVC was 0.27527. It is hoped that the different forms of classification in nonlinear SVC methods will perform better.

SVC uses a one-vs-one approach, which would perhaps perform better than the one-vs-rest approach of LinearSVC. Since the default kernel here is rbf, it will also likely work better with the multiclass categorization of our dataset. Utilizing a gamma of 'auto' was found to be more efficient, which was expected because it both allows the model to run faster, and because it will work to correctly scale the amount of influence each single training example has. These settings yielded an accuracy score of 0.39344, so hyper parameter tuning was pursued to see if the

model could be further improved. Again this model did not do a good job of predicting “same day” adoptions, so it was repeated without those data points and the model score of 0.42099, which is the best model performance so far.

When testing on all the data, including the validation set the scores were 0.40590 when including all data and 0.44668 when removing the same day adoption values. This is interesting, and is probably due to the way that gamma is being utilized to fit the shape of the Gaussian, which may be filtering out noise.

K-Nearest Neighbors (KNN)
kNN was chosen for exploration because it is a method that allows for investigation into both classification and regression problems, so perhaps it will be a good blend of all the previous models. The only concern prior to application is that the data set may be too large to analyze using this method, even though the decisions made through the analysis process allowed for some of the dataset to be eliminated. Analysis found that the brute force method and the kd-tree method with a leaf size of 1 performed the same. Due to the nature of the kd-tree, it was chosen to stick with brute force. Kd-tree is an axis aligned algorithm, which gives it less flexibility when taking shape and is highly feature dependent. This dataset is small enough that brute force can be run without taking too much time. Final score for this model was highest if eliminating the same-day adoption dogs, and was found to be 0.41656.

Random Forest
It is now time to investigate ensemble methods to see how these perform. The first to explore is Random Forest. A classifier was chosen because we’re trying to predict a categorical response variable. Random forest performed best so far of the models, and hyperparameter tuning was pursued. When tuning the number of n-estimators, the score for all animals was 0.47109, and for animals without same day adoption speed 0.48557, which is better than the leader on the Kaggle competition!

LightGBM
It is now time to investigate one final model, LightGBM. This is also an ensemble method and tends to perform well on Kaggle. First, the features were plotted to see if they were consistent with those features that showed the highest weight in our statistical analysis. While breed and age were expected to be heavily weighted, the amount of photos was not after statistical analysis. However when using LightGBM to see the importance of each feature, the photo amount did rank higher than breed. Even with tuning hyperparameters, the model performed best using the default arguments. The best score was again found when dropping same day adoptions, and was 0.47679.

FINAL MODEL RESULTS:

A comprehensive model summary as seen in the abstract can also be found in Appendix A .

The best model choice ended up being Random Forest, which was not unexpected. This method along with LightGBM and XGBoost tend to be most successful in Kaggle competitions. Some of these findings were very unexpected, having looked at many of the other Kaggle notebooks submitted during this competition. The difference in findings can easily be attributed to dropping cats out of the dataset. However, it was interesting to see the same or better accuracy on models when certain parts of the dataset were intentionally dropped. For example, without analyzing the

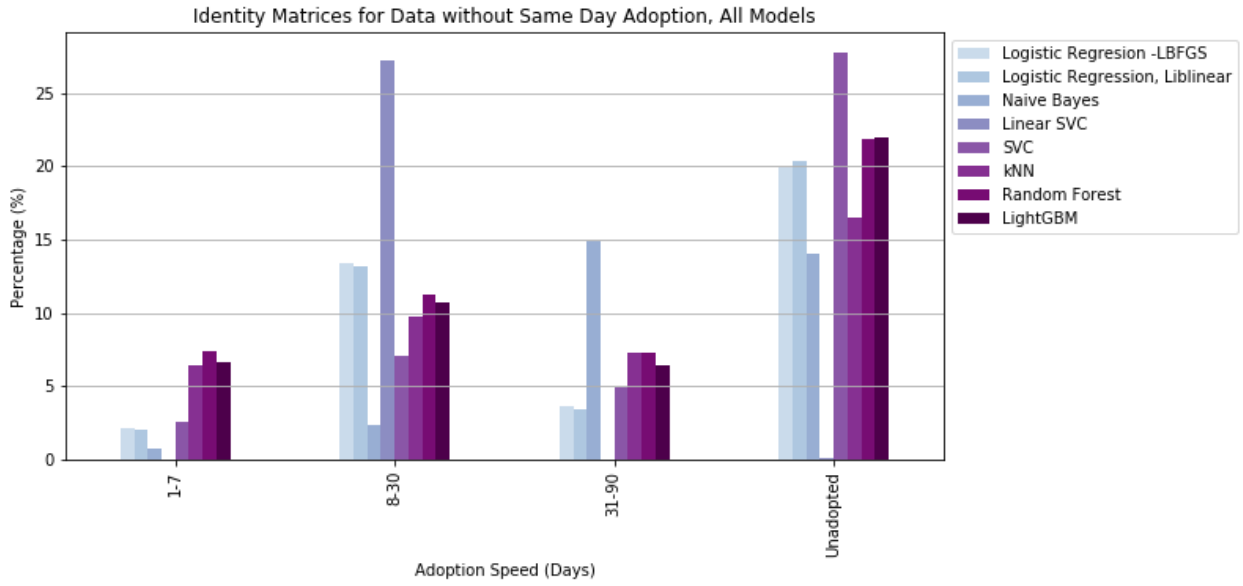
descriptions or photos of the pets, the final score was comparable to that found by the leader in the kaggle competition. It's possible that by eliminating the presence of cats, noise was eliminated, improving prediction accuracy. It's also possible that the algorithms developed to handle the "name" feature of each pet listing was not as effective as manually cleaning the data.

Overall, the findings were consistent with what is understood in real-world dog adoptions. It's very difficult to predict which animals will be adopted on the same day, and is more or less random. The likelihood a dog will be adopted probably has less to do with the animal and more to do with pet trends, seasons, and what people are looking for in a dog. Or, perhaps it's like every other relationship we have as humans, and it's all based on that instant connection an adopter has when meeting a pet, which a machine simply cannot predict.

RECOMMENDATIONS TO THE CLIENT

The client's request was to better understand how to address incoming animals so a strategic plan can be developed. This led to the development of multiple models to determine which would best predict outcomes for animals within certain time ranges, to allow for more effective care, training, and adoption event planning on a case-by-case basis.

A summary of each model's accuracy was created by pulling only the True-Positive classifications out of the identity matrix and plotting them side by side. This summary was found to be most useful when eliminating data from the models for those pets that were adopted on the same day, due to the randomness of this event. However, a summary plot of accuracy for each prediction by model when studying all animals can be seen in Appendix A. The summary chart for only animals who were not adopted on the same day as being listed can be seen in the figure below.



The chart was used to reveal the following recommendation system for the client to implement:

- The best chance of predicting a same day adoption is to use the logistic regression model using a liblinear solver.
- The second step in analyzing a new pet during intake should be to run LinearSVC, and keep those pets who will be adopted in the first month if capacity is available.
- To determine those pets who will require a more medium duration stays (<90 days), utilize the Naive Bayes model
- To predict those pets who will require long term care, and possibly rehoming to a more targeted rescue, specialized training, etc. utilize the following models: SVC, kNN, Random Forest, or Light GBM.
- An averaged prediction for all models can also be utilized.

If the client feels this will appropriately suit their needs based upon the original consult, additional work can be completed to develop a targeted recommender-style prediction program.

DATASET LIMITATIONS

While the dataset does a good job of providing examples for PetFinder listings, and a lot of work has been done to investigate how to improve prediction of adoption speed based on a listing, there are some points about the data that are lacking in clarity. They include, but are not limited to the following:

- Understanding interactions between listings opening and closure- how is “adoption speed” defined. Are animals that go unadopted simply not updated listings?
- Understanding how unique identifiers are assigned - is it possible for a pet to change rescues and have multiple listings for the same animal with different outcomes?
- Understanding how breed is determined for listings - is it appearance based or DNA based?

OPPORTUNITIES FOR FUTURE WORK

This project is as comprehensive as possible, given the constraints. However, there are additional opportunities to expand on this work, including, but not limited to:

- Exploring the effect of specific names on adoption speeds
- Exploring the effect of description on adoption speeds
- Exploring the appearance of pictures and their effect on adoption speed, particularly since some relationship was indicated between photos/adoption speed
- Understanding animal trends in other countries and being able to compare relationships between the data set analyzed here and local trends
- Exploring the relationship between specific breeds and adoption speeds to uncover bias

- Exploring the relationship between animal size/color and any unintended bias in adopter response
- Examining breed/size restrictions in different states and correlating to uncover additional trends
- Develop an effective algorithm to process the “name” feature of the data and clean it, rather than cleaning manually, and repeat the analysis
- Perform a data cleaning on the entire dataset, and repeat analysis to determine accuracy of the models without using description or photo analysis.
- Exploring housing and other socioeconomic trends in the area surrounding each rescue to determine those effects on adoption speed.
- Development of a targeted recommender style adoption speed predictor system to predict adoption speed and provide a recommended course of action. This model could be further developed by analyzing the rescue specific data, and data of those rescues in the surrounding area. The data could then be used to include recommendations for more target-appropriate rescues to send difficult-to-home pets that may succeed better if placed with a different rescue.

CITATIONS

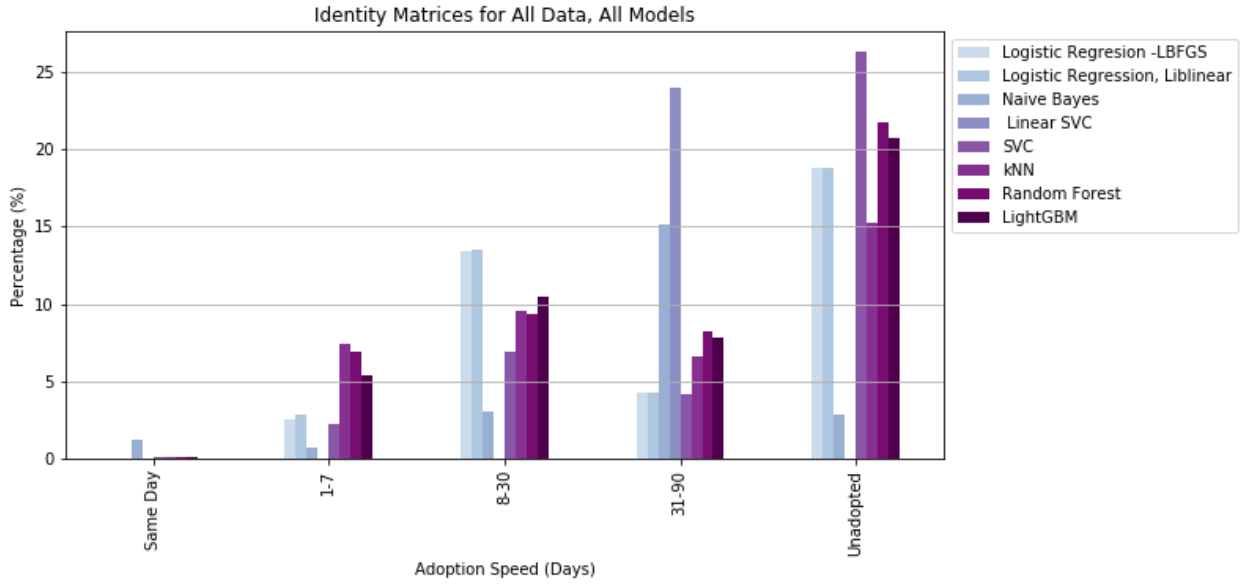
[1] https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9

[2] https://www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics

[3] https://www.kaggle.com/c/petfinder-adoption-prediction/discussion/81597

Appendix A

MODEL ACCURACY SUMMARY FOR ALL PETS, ALL MODELS:



MODEL PERFORMANCE SUMMARY:

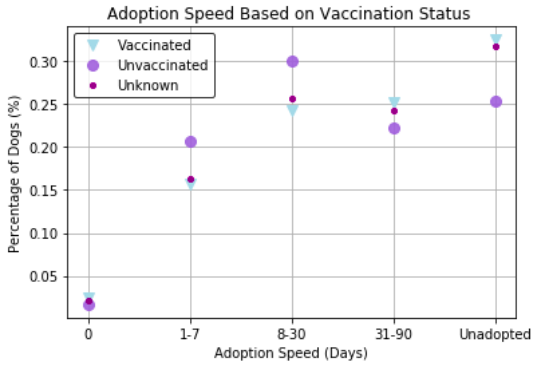
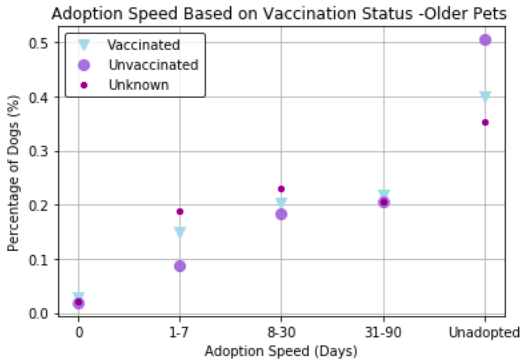
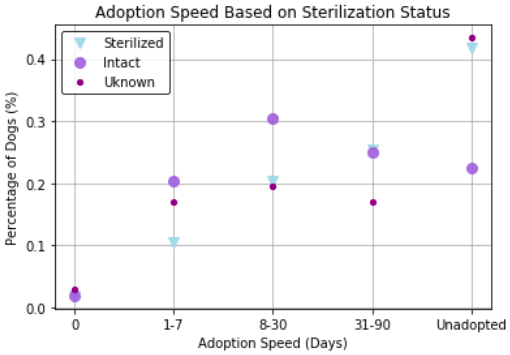
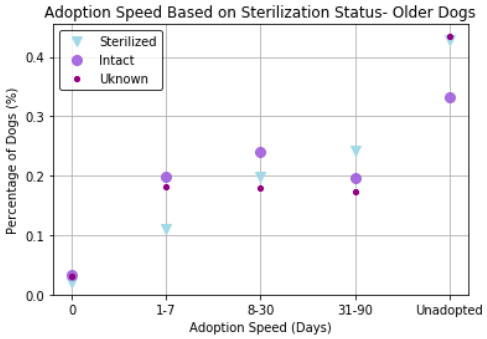
Model	Accuracy Score			
	Same Day		No Same Day	
	Training	Validation	Training	Validation
Logistic Regression- lbfgs, restricted features *	0.39208	0.37146	0.40754	0.38645
Logstic Regression - lbfgs, all features	0.38579	0.39237	0.41541	0.37516
Logistic Regression - liblinear, all features	0.38579	0.39483	0.40369	0.37641
Naïve Bayes	0.23880	0.26568	0.33626	0.34003
Linear SVC	0.24918	0.36162	0.27247	0.24467
SVC	0.39071	0.40590	0.42099	0.44668
kNN	0.36885	0.39606	0.39698	0.41656
Random Forest	0.46940	0.47109	0.47571	0.48557
Light GBM	0.45027	0.46740	0.48130	0.47679

* Features Included: Rescuer ID, pet age, first breed classification, amount of photos

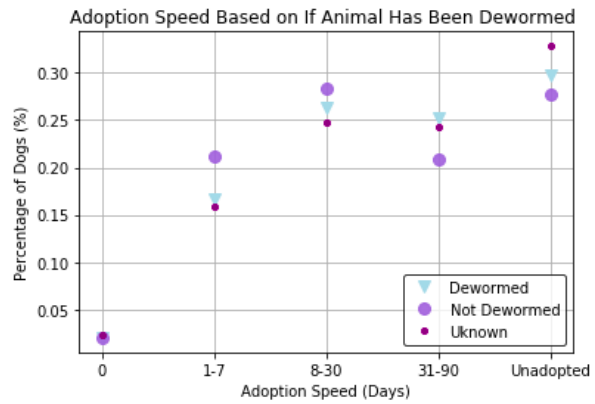
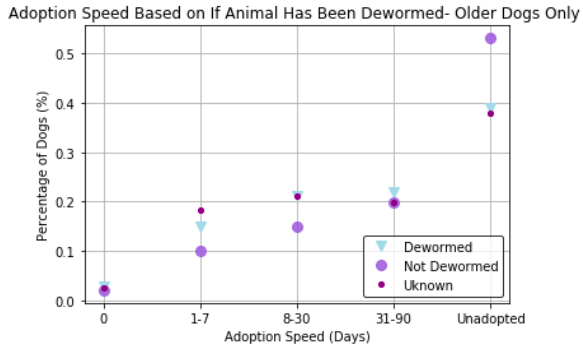
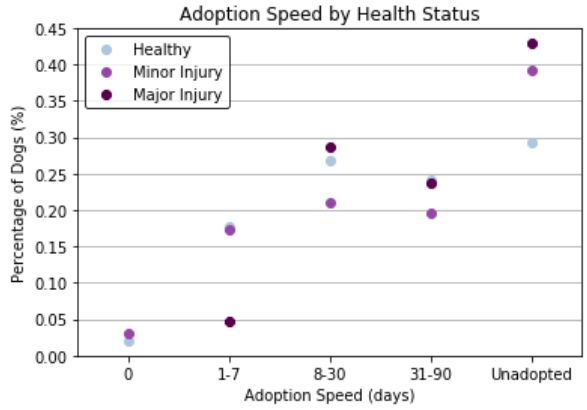
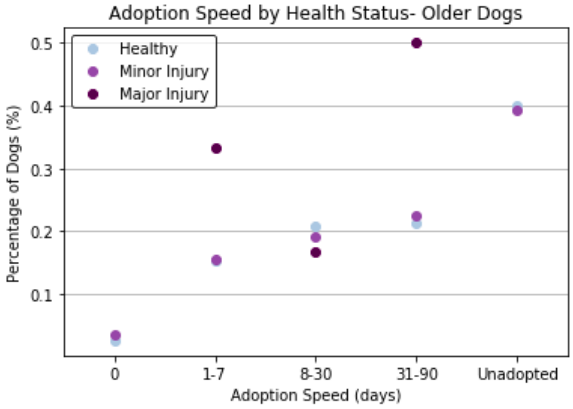
** Kaggle PetFinder Leadboard Winning Accuracy 0.45338

GENERAL HEALTH PLOTS

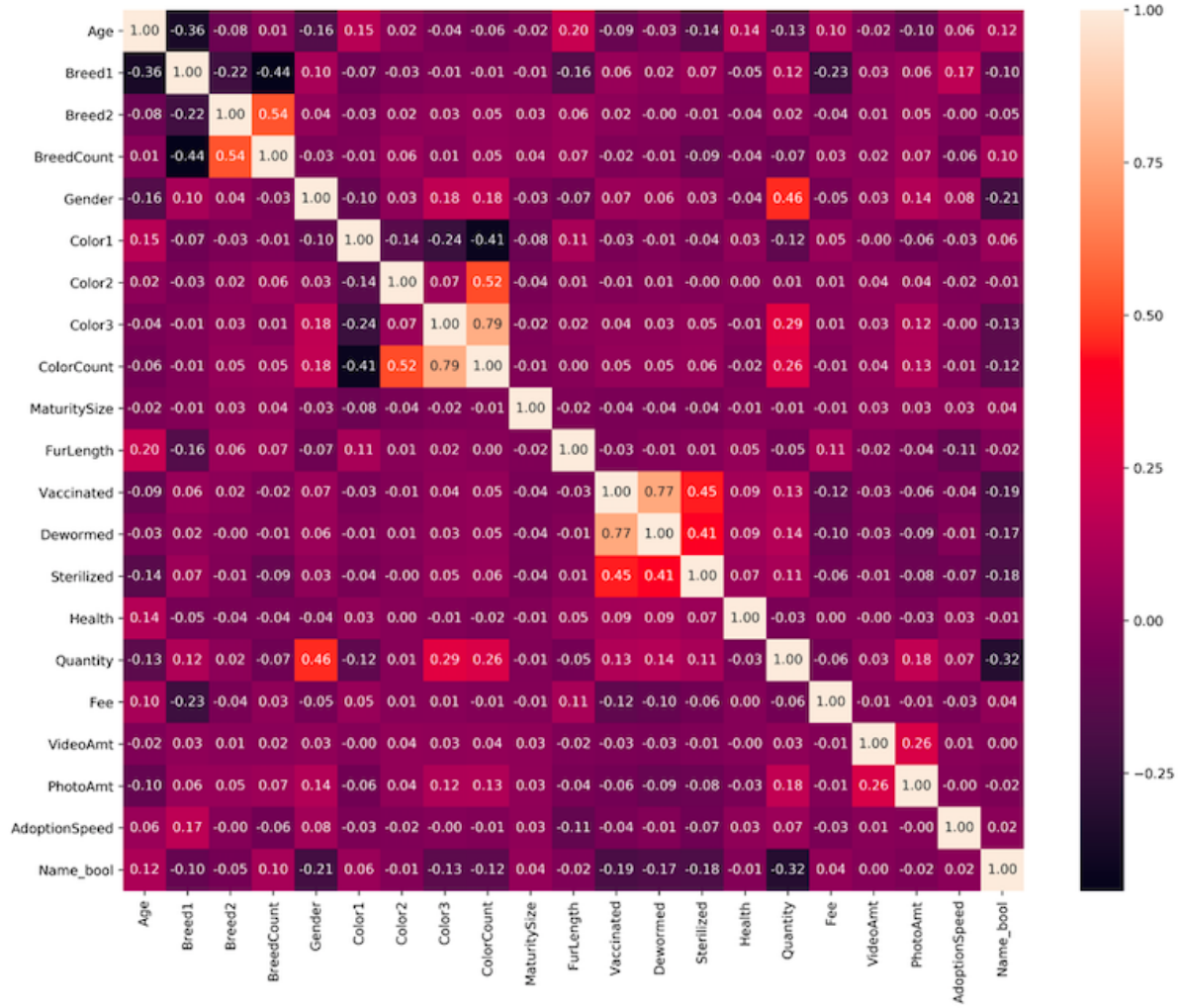
Adoption speed as compared to a) vaccination status for all animals, b) vaccination status for only older pets, c)) sterilization status for all animals, d) sterilizations status for only older pets,

ALL PETS DATA	OLDER PETS ONLY DATA
A 	B 
C 	D 
Charts E-H on next page	

CONTINUED FROM PREVIOUS PAGE: Adoption speed as compared to e) digestive status (dewormed, not dewormed) status for all pets, f) digestive status for only older pets, g)) overall health (pre-existing conditions) for all pets, and h) overall health for only older pets.

ALL PETS DATA	OLDER PETS ONLY DATA
E 	F 
G 	H 

CRAMER'S V HEATMAP SHOWING CORRELATION BETWEEN FEATURES:



STATISTICAL RESULTS OF FEATURE SPECIFIC INVESTIGATION:

*Cramer's V Value reported is vs adoption speed. Green indicates

FEATURE	CONCLUSION	χ^2 VALUE	P-VALUE	CRAMER'S V VALUE*
Age/Life Stage	Age and adoption speed are related	939.16	1.31 E -44	0.06
Name	The dog's name and adoption speed are independent.	13,096.11	0.98	-
Presence of Name	Whether the dog has a name affects adoption speed.	10.46	0.03	0.02
Breed	There is a relationship between breed and adoption speed.	2738.21	2.39 E -41	-0.06
Mixed vs Purebred	There is a relationship between mixed vs purebred dogs but it's not as significant as the relationship between specific breed and adoption speed.	75.32	1.70 E -15	0.17
Size	There is a relationship between an animal's adult size and adoption speed.	102.48	1.82 E -16	0.03
Fur Length	There is a relationship between fur length and adoption speed.	106.14	2.36 E -19	-0.11
Color	There is a relationship between color and adoption speed.	74.19	4.96 E -7	-0.03
# of Colors	There is a relationship between whether a dog is a single color and adoption speed, but it has less impact than color alone.	30.85	1.4 E -4	-0.01
# of Dogs in Listing	Not tested due to the nature of data	-	-	0.07
Photos	There is a relationship between photo amount but it is not a large correlation.	401.19	5.34 E -32	-0.00
Videos	Videos and adoption speed are independent.	33.80	0.38	-0.01
Cost	There is a relationship between adoption fee and adoption speed	327.04	1.5 E -4	-0.03
Vaccination	There is a relationship between vaccination status and adoption speed.	90.03	4.59 E -16	-0.04
Sterilization	There is a relationship between sterilization status and adoption speed.	431.30	3.75 E -88	-0.07
Wormed vs Dewormed	There is a relationship between whether a dog has been treated for worms or not and adoption status.	42.40	1.14 E -6	-0.01
Pre-existing Conditions	There is a relationship between an animal's health status (healthy vs pre-existing conditions) and adoption speed.	18.51	0.02	0.03