UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Master in Innovation and Research in Informatics (MIRI)

Algorithmics for data mining course

First Delivery

Animal Shelter prediction

Worked by: Nejada Karriqi

March, 2019

Description:

Animals are something I love so I decided to develop this competition in order to find insights that might help those animals find a new home. Based on the statistics in US, approximately 7.6 million animals end up in shelters. There reasons are different, starting from loss, sickness, aggressivity and more others. Our dataset is taken from Austin Animal Center records from October, 2013 to March, 2016 and Shelter animal statistics from ASPCA. The dataset includes information such as breed, color, sex, and age. The analysis of this dataset might help shelters understand the trends in animal outcome, helping them on developing a specific structure to help animals with specific needs.

Objectives:

My objective is to predict what happens with those animals after they leave the shelter. In my predictions will be included: adoption, transfer to another place, euthanasia or return to owner. The prediction will be based on the most important variables such as age, color, gender, name, ect. The result might be useful for all the animal centers in order to help them build up a structure for helping animals with specific needs get adopted or find their owners.

Hypothesis:

- Young animals older than a month have a high chance to be adopted.
- Old dogs have a high rate of finding their owner.
- Old cats are mostly euthanized.
- Color of animals has high impact on lost animals.
- Compared to cats, dogs in general have more chances to be returned to their old family.
- Having a name, improves the adoption rate for the animals.

Data description and analysis:

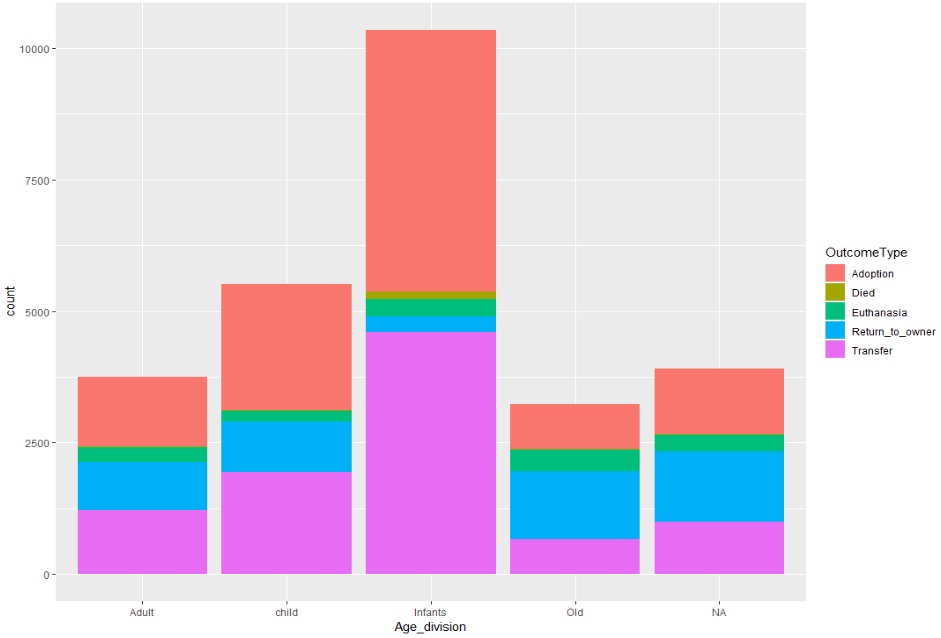
This data set contains 38185 observations and 10 variables for each observation, plus more added by me for analysis purpose. The dataset was divided by Kaggle into train and test dataset. The first thing I did was to check for the missing values and try to impute them in order to not let them affect the final result. Here are all the variables of the dataset.

- ID - During their intake, all the animals get a unique ID. This variable is unchangeable and will be considered as factor (doesn’t impact the outcome).
- Name – is the animals name. The column name has lots of missing values. The first impact is to ignore this variable but instead of ignoring it, I will try to use it as a factor to help the increase of adoption rate.
- DateTime – contains the date and the hour when the animals arrived at shelter.
- OutcomeType – will contains info about what has happened after shelter leave.
- AnimalType – we have two types of animals in dataset: dogs and cats.
- SexuponOutcome - here are considered four groups: intact male/female, spayed, neutered and unknown.

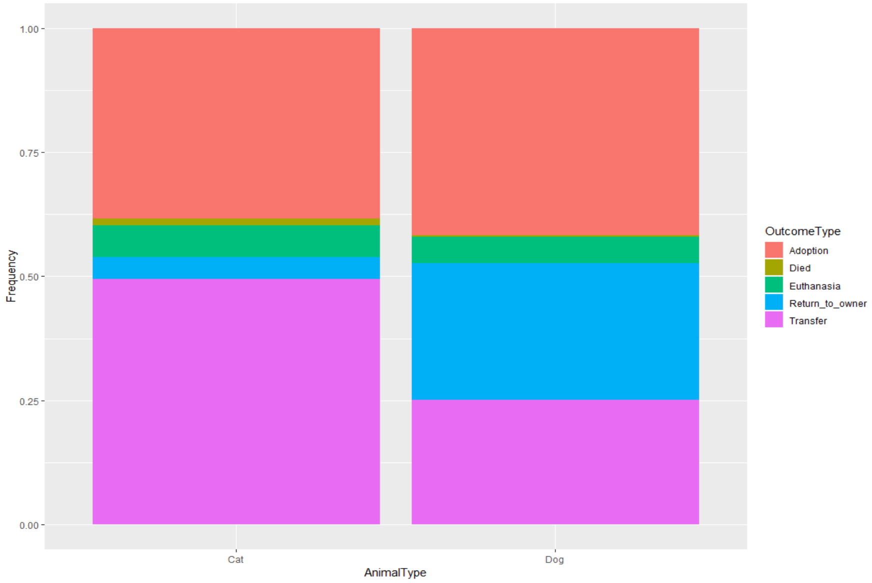
- AgeuponOutcome – is the animal's age. To properly predict the outcome and to see how different factors impact the result I found it reasonable to change the scale of animals age. Converting it to days old will be easier to compare between all the animals of different ages.
- Breed – contains information about the animal's origin (mix or pure).
- Color – the animal's color

Procedure:

After pointing the missing values and substituting them I decided to check the correlation between the OutcomeType (the variable I want to predict) and other variables I find to be important on this prediction. The column with more missing values was name, so I replaced them with Unnamed. Then I checked all the variables in order to identify which one of them was continuous and which one was to be considered as factor. The first one was the correlation of OutcomeType with age so in order to have a fair result I changed the age into AgeinDays (same scale for all the animals) and then I separated the animals into 4 categories: Infants, Child, Adult and Old. These categories are stored in a new column and was added to the original dataset.



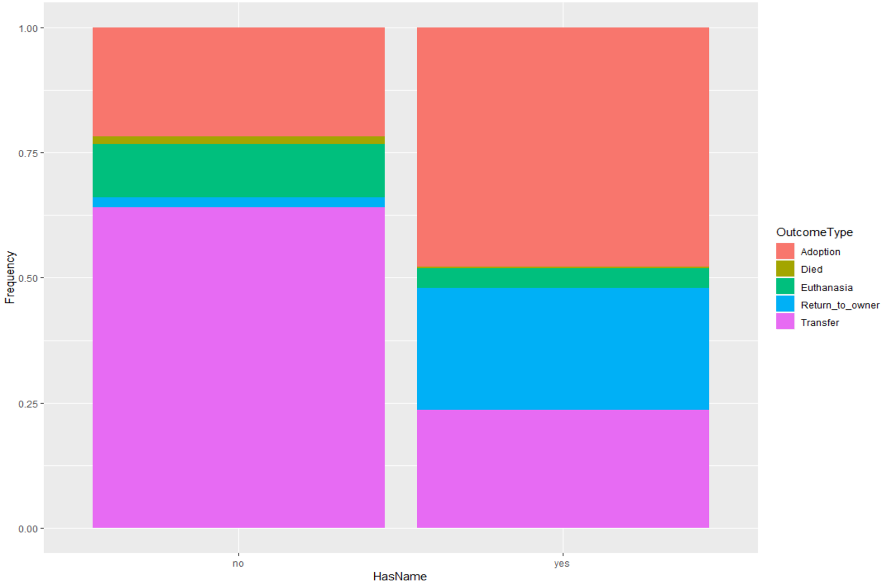
Here is the graph of the outcome type based on the animal type:



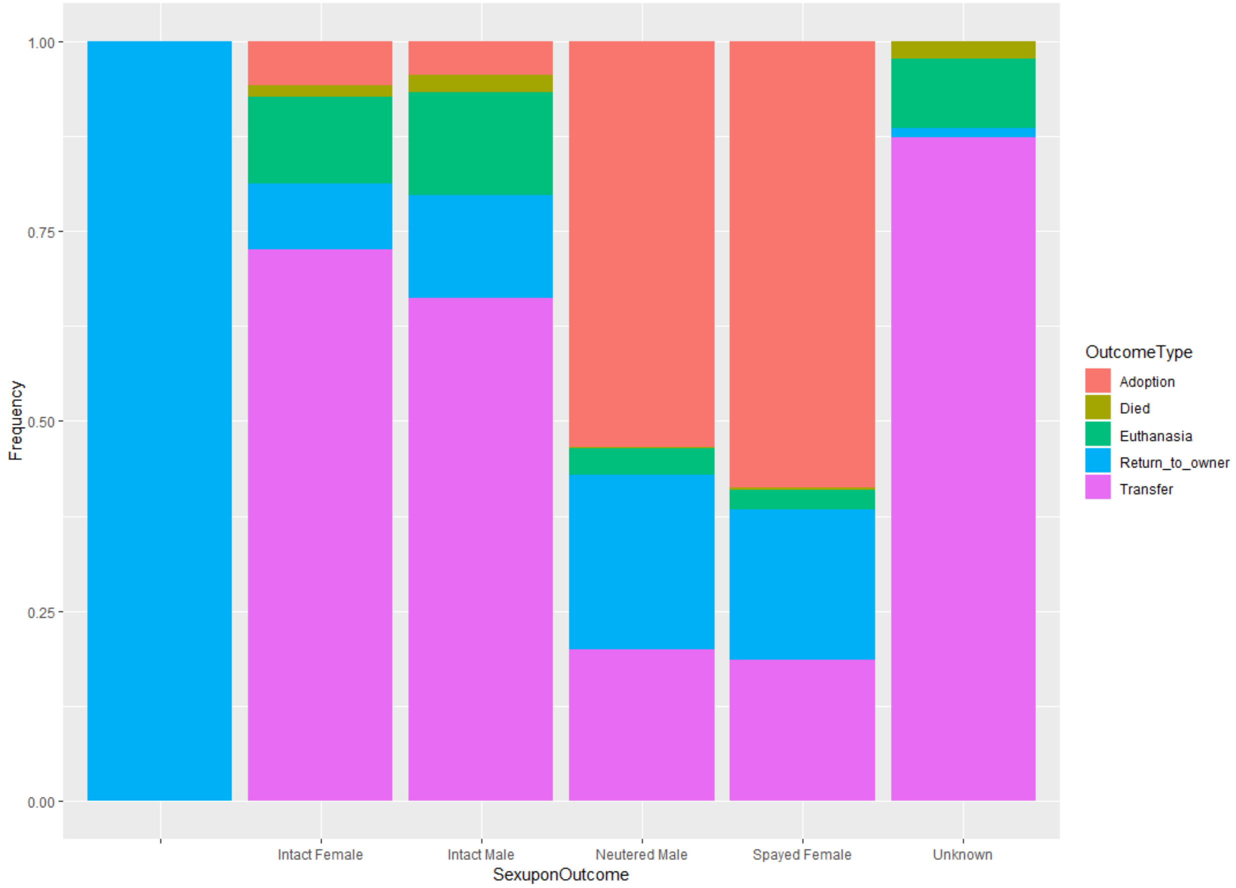
I did the same for HasName (a new column created to identify if the animal has name or not and added to the original dataset), AnimalType, SexuponOutcome and Color.

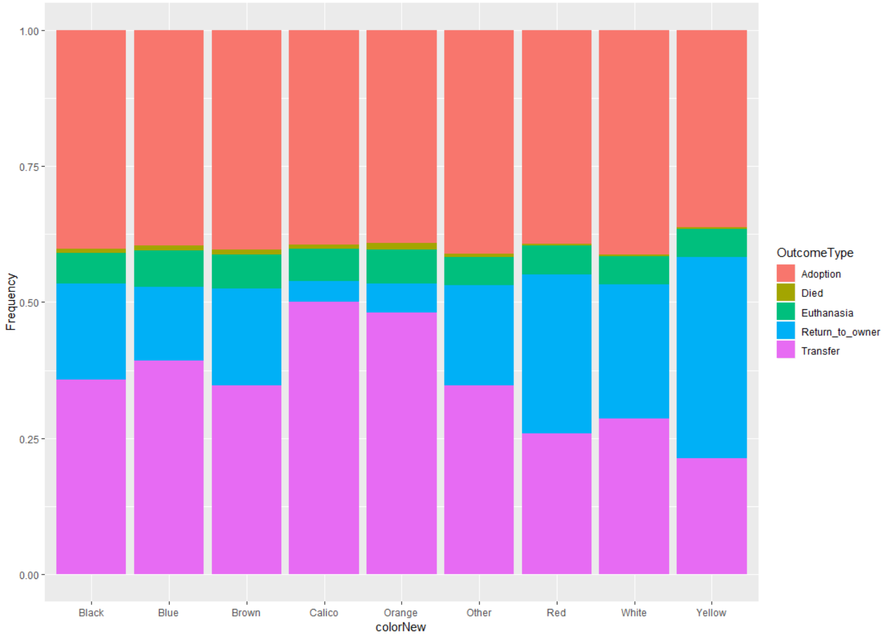
	Adoption	Died	Euthanasia	Returned to owner	Transfer
yes	47.7518647	0.4044542	3.8869629	24.3355394	23.6211787
no	21.8177090	1.5602652	10.5968015	1.9893382	64.0358861

This table shows the percentage of the animals for each Outcome Type considering if they have or not name and the graph bellow allows us to visibly see the correlation of name and Outcome type.



The graph of SexuponOutcome:





I decided not to consider all the variables when predicting my model so I created a new training set containing only the variables I thought were significantly important, mostly for OutcomeType adoption and return to owner, and I divided the observations in two parts . This is a decision I took based on the objectives I had (helping the shelter build a strategy to help animals get adopted and find their owner). The model I chose to create is the GLM model (generalized linear model). GLM is more expanded than the linear model and by using this model we are able to fit model binomial data with logistic regression, run logistic regression and to get the regression success on the numeracy and anxiety scores. The standard formula to create a GML is:

glm(*formula***, family=***familytype***(link=***linkfunction***), data=)**

My GLM model was based on variable age division, variable has name and variable color. The data used is just the training and the family is binomial. From the results of the predictions I got these results:

The color was not significant for the adoption, so it had very small values; the age is highly correlated (infant and child negatively correlated) to the output type and has name is negatively but highly correlated. The final score of the initial dataset is 60%. The result is acceptable.

Conclusions:

From the dataset I understood that people prefer dogs more than cats, so in the most of the cases the cats were transferred. The age is very important for all the types of output I predicted. Usually the animals older than one month have higher chances to get adopted, compared to younger animals and for the dogs at a mature age (old) have a very high chance of returning to their owner. Even though the color seems not to be so correlated to the output type, from the graph its very clear the impact that color has on return to owner output type. So, as we understand the descriptive variables may not affect adoption but for sure they affect a lot on finding the lost animals. The name seems not important variable but the affect it has on the output type is amazing. When having, the number of all the output types is doubled or more. So, except of the descriptive part (affecting to return to owner output), having a name doubles the adoption rate. Another important characteristics is the sex upon outcome: intact male, neutered, sprayed. It helps people decide on adoption especially if the owner doesn’t want to have animal babies in his home.

Finally, I can state that all my hypothesis were true. I got the expected result and I hope it will be helpful.

References:

Kaggle competitions https://www.kaggle.com/competitions

Austin animal center http://www.austintexas.gov