Dylan Wise
Carmon Zachary Almon

Project Report

Project Description:

For this project we decided to choose the Animal Shelter dataset. One of the large reasons we chose this project is because of our love for our own pets. So having a dataset that would give us more insight about what happens in animal shelters and understanding more about what traits help an animal have a higher chance of being adopted. Once we know which traits have a higher effect on whether an animal has a favorable outcome we can either focus on animals that do not contain those favorable traits to help them get adopted or focus our resources on helping the ones with favorable traits because we know that they already have a good chance and why not raise it more. What we ended up finding overall is that unsurprisingly that age has the highest effect on whether a pet, dog or cat, gets adopted. However it is worth noting that our handling of the data may not have been optimal and there could very well be a trait that also has a significant effect on whether an animal is adopted. If it lies anywhere it lies within breed or color since we handled that very basically.

Approaches

For this dataset we decided that the best approaches that we could do would be Naïve Bayes, which is a classifier, and Apriori, which uses association rules. Our main reason for using a Naïve Bayes classifier since the entire reason we are doing this is to predict what the outcome is for a particular animal solely based on the traits provided which is exactly what a Naïve Bayes classifier does. For this we decided to use the Naïve Bayes classifier that is within scikit learn. The reason we chose this versus our own is because for our own Naïve Bayes classifiers we build it with the outcome being the last attribute in mind. We were already crunched for time and we were sure that scikit learn's classifier was better than ours anyway. The reason we chose Apriori as our second method is because it would give a great idea as to which traits apply to whatever outcome the most. This would help us understand whether a trait is worth keeping or not, the higher the rule confidence the higher the effect.

Results:

For Naïve Bayes the quality of the model is entirely based on its ability to predict the test data. Initially when we ran the classifier we were able to achieve an accuracy of 67%. However this was before we realized that we had accidently left the outcome subtype attribute in the testing and training data. This unfairly skewed the results in our favor because one of the subtypes, partner, was always paired with transfer and this particular pair appeared in 30% of the total dataset. So once we found the error and dropped the subtype attribute our accuracy lowered to 57%. After messing with which attributes to drop we discovered that dropping color, breed and name was able to raise the accuracy to 62%. Deleting further attributes either has no effect or lowers the accuracy. Rather surprisingly deleting the sex attribute had a larger affect that deleting the age attribute. This tells me that we probably should have focused on refining the sex column some more. Possibly deleting those that were unknown and separating the animals into fixed and not fixed. Basically what learned from how we processed the data is that sex and age are the most important aspects in changing whether an animal is adopted with name actually lowering the results when we first concluded that it actually helped due to a rule generated

from Apriori. This rule however mainly applied to cats so maybe if we only had cats in this dataset we would have a greater effect.

The Apriori application we used was the one that we used during our first HW. We decided that the best settings to use were a relatively low overall support of 10%, to see at least one of the subtypes and a rule confidence of at least 75%, we decided that was good to see significant effects. For this we saw that age, particularly young age, was great in determining if an animal was adopted or not. Unfortunately due to our extreme focus on the adopted outcome in looking at the rules we missed some important rules regarding sex that would have increased our accuracy. Transfers we heavily influenced by the sex of the animal and if had focused on that a little more we would have known that we should have adapted our preprocessing techniques to improve the sex attribute some.

Both of our techniques gave us favorable results in the end. Naïve Bayes performed better than it would have is somebody had randomly guessed what the result was for the outcome. It certainly could have been improved more with better handling of breed, color and sex in particular. We truly believe dropping unknowns and splitting attributes based on fixed or not would have yielded a more accurate model. The results of the apriori test could have been improved with better human interpretation of the results. We focused so heavily on adoption that we didn't focus on improving traits that effected returned to owner or transfer. So overall human error is what resulted in our results not being as good as they could have been.

Related Work:

None of the publically submitted scripts on Kaggle used Naïve Bayes or Association Rules so we am unable to directly compare to any of their submissions. However the very top post which used random forest also had a section where they tested the importance of each attribute. They predicted that age had the greatest importance in determining an outcome and whether they were intact or not was number two. This is slightly different from the conclusion we came to where we saw that sex had a higher weight. But, they separated sex and whether or not the animal was fixed or not into two distinct attributes and we found that those two together had a greater impact, at least on naïve bayes, on the results. They also came to a similar conclusion about whether the animal was a mix or not into the results. We both found that the results were not affected all that much. These results were generated by Megan L. Risdal in the post Quick and Dirty Random Forest if you wish to see the weights for each attribute they came to themselves. Our group is very critical of our results as you can see above. Megan also has the highest rated post on the data so it's a pretty good assumption that her conclusions are probably quite accurate.

Discussion:

Unfortunately the results from our data mining methods didn't tell us too much that we already couldn't assume. However it is nice to see what our assumptions about age and whether an animal was fixed or not work well with real math. We think with better preprocessing we could have had more results that would have told us not only the information we already knew but some more information

that one wouldn't expect like with the colors or breed. We truly believe those were the weakest parts of our data. For dogs doing anything with the breeds would have been pretty difficult. However for cats there were relatively few breeds and we may have been able to see which breed is more favorable. For color we could have separated the data into how light or dark a color was. This would still be very difficult because of how much variation there was. We were disappointed that subtype wasn't able to be effectively used and we have no idea how it could have been with the data we were given. We believe that our biggest misstep wasn't doing anything with the attribute sex upon outcome since apparently its effect was so great. Not only would it have been a simple change we could have also separated it into two separate columns for sex and if the animal was fixed or not. Given more time we would have done more tests with techniques like random forest which seemed really popular, however time just simply wouldn't allow it.

All in all we think this project was our favorite part of the entire class. We liked how open ended it was and we certainly leaned a lot about data preprocessing data even if it we didn't do it as well as we could have. If given the chance to do it again we have an idea of what we would do and how we would go about it. We think our greatest negative here was simply how crunched for time we were. This was caused by a combination of conflicting schedules, work, other projects and poor time management on both of our parts. The experience we would say was overall positive.

Reference:

[1]"Shelter Animal Outcomes | Kaggle", *Kaggle.com*, 2016. [Online]. Available: https://www.kaggle.com/mrisdal/shelter-animal-outcomes/quick-dirty-randomforest. [Accessed: 30- Apr- 2016].