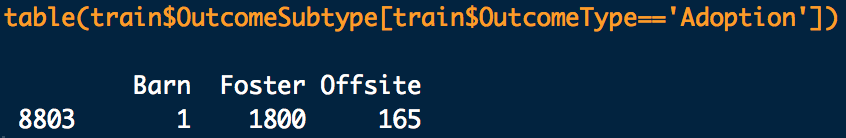
**Overview of Shelter Dataset**

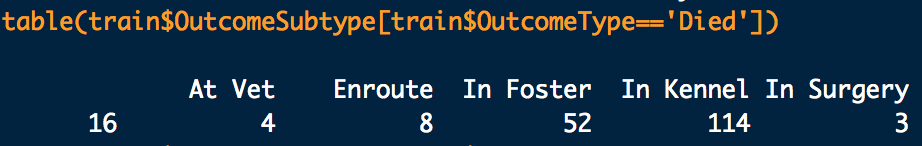
In the train dataset, there are 26729 observations of 10 variables, while there are 11456 observations of 8 variables in the test dataset. The missing two variables are *OutcomeType* and *OutcomeSubType*, which record the destiny of shelter animals. The other 10 variables are *AnimalID, Name, DateTime, AnimalType, SexuponOutcome, AgeuponOutcome,* Breed and *Color.*The cats are counted as 11134 and dogs are counted as 15595 in the train dataset. Nearly 29% of animals don’t have a name.

Basically, animals’ journey in the shelter end in five ways: Adoption, Died, Euthanasia, Return to owner and Transfer. In the train dataset, 40.3% of animals are adopted, 0.7% of animals died, 5.8% of animals are euthanized, 17.9% of animals are returned to owners and 35.3% of animals are transferred. The *OutcomeSubtype* variable is not recorded detailedly and only 13612 observations have values, which are additional remarks to *OutcomeType*. The following screenshots show the various *OutcomeSubtype*s of each *OutcomeType* category respectively.

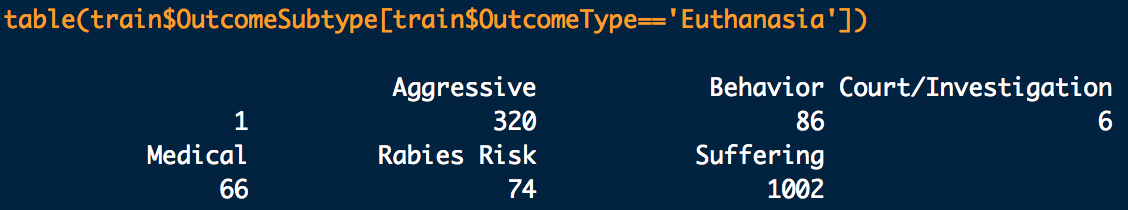
*Figure 1 Subtypes of Adoption*



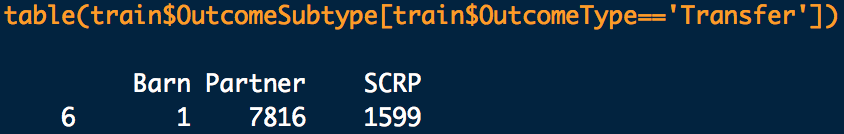
*Figure 2 Subtypes of Died*



*Figure 3 Subtypes of Euthanasia*

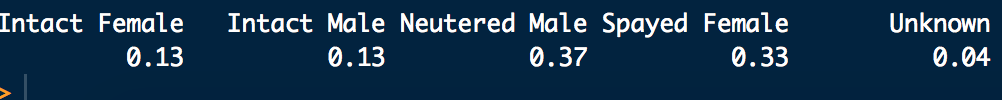


*Figure 4 Subtypes of Transfer*



One of biggest feature of this dataset is that the values in most variables are nested and thus complicated. The *DataTime* variable records the exact date and time when the animal was sent in. Given the nested date and time in this variable, we will split it later to see how outcome will be impacted by the time and the date. The *SexuponOutcome* variable records the combination of animals’ gender and intactness information and there are five groups: Intact Female, SpayedFemale, Intact Male, Neutered Male and Unknown. There is one missing value in this variable and we decided to assign it as “Unknown”. The percentage of these five groups in the *SexuponOutcome* are listed below.

*Figure 5 Groups in SexuponOutcome*

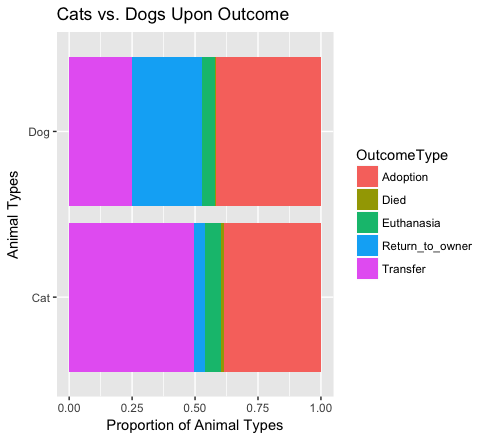


The *AgeuponOutcome* variableis very complicated since the value is composed of numeric value and non-unified units like days, months, weeks and years, resulting in 45 different groups. We decided to unify this age variable later. Similarly, the *Breed* variable even has 1380 different groups and the *Color* variable has 366 different groups.

**Data Exploration**

For the *AnimalType* variable, cats are counted as 11134 and dogs are counted as 15595 in the train dataset. We are interested to explore whether people in Austin have tendency to adopting animals based on types. The Figure 6 clearly demonstrates that dogs and cats are almost equally adopted. Cats are more likely to be transferred than dogs and dogs are more likely to be returned to owners. In general, both death cases and euthanasia cases account for a very small part.

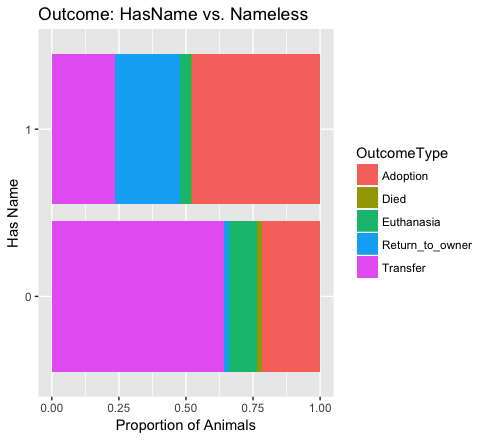
*Figure 6*



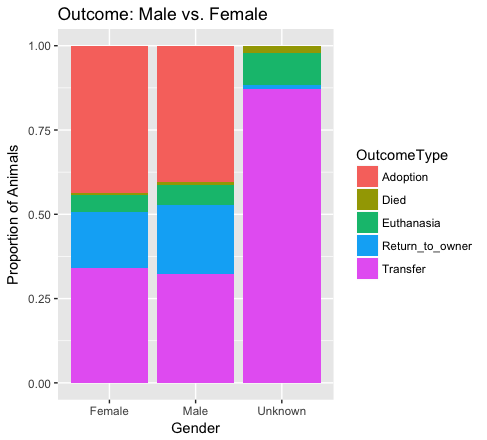
For the *Name* variable, we are interested to explore whether nameless impacts the outcome of animals. The Figure 7 shows that whether an animal has a name or not plays importantly in outcome types. Animals with names are more likely to be adopted or returned to owners than animals with no name. Meanwhile, nameless animals are highly likely to be transferred.

We extracted the gender value out of *SexuoponOutcome* and found that the outcome distribute are almost identical to the male and female groups. But animals with unknown gender are more likely to be transferred and few of them are adopted (Figure 8). The intactness values are extracted from *SexuoponOutcome* as well. Neutered animals are more likely to be adopted than intact animals and unknown group. Additionally, neutered animals are more likely to be returned to owners than the other two groups. It may due to the fact that usually animals who have owners have been neutered. Intact animals are more likely to be transferred. Unknown group are most likely to be transferred. Most euthanasia cases occur in intact and unknown groups.

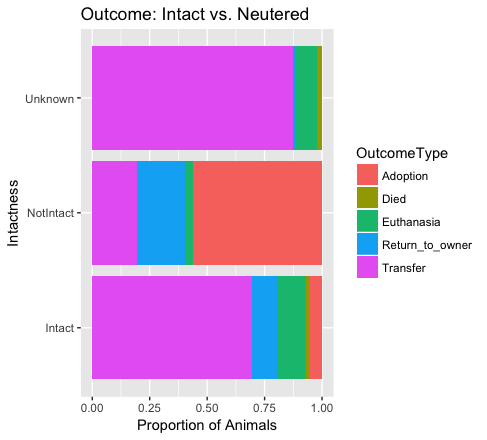
*Figure 7*



*Figure 8*

****

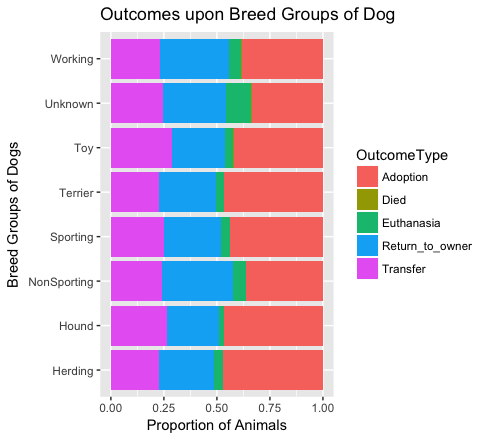
*Figure 9*

****

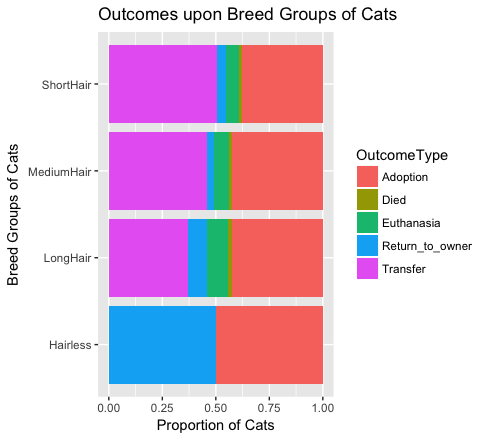
For the Breed variable, we classified dogs’ breed into 8 groups based on the information from American Kennel Group and classified cats into 4 groups based on the length of hair. The result shows that the outcome distribution does not differ greatly between different breed groups of dogs (Figure 10). Generally, Terrier, Hound and Herding dogs are relatively more popular than other breeds in Austin. In contrast, the breed of cats did influence the outcome distribution (Figure 11). Shorthair cats are more likely to be transferred than other breed types. Medium Hair cats and long hair cats have higher chance of getting adopted.

In order to explore the effect of age in outcome, we transformed the values of *AgeuponOutcome* in base unit of day and classify them into two groups: less than 365 days—young, and more than 365 days—mature. Young animals are more likely to be adopted than mature animals. Mature animals are more likely to be returned to owners. It may due to mature animals have more chance of getting lost while outing and young animals does not have too much time going outside (Figure 12).

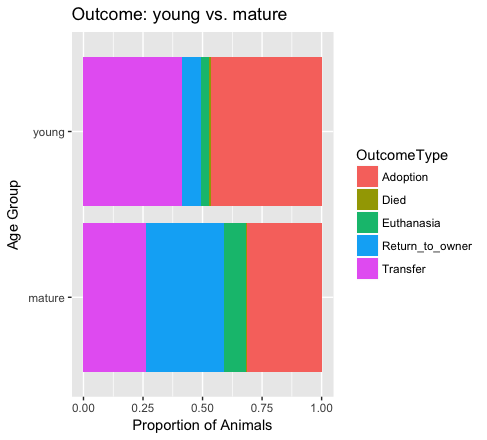
*Figure 10*

****

*Figure 11*

****

*Figure 12*

****

**Model**

1. **Data Lineage**

We summarized the dataset, and found there are 10 variables in the train and 8 variables in the test. Since our final output is to predict the probability of *OutcomeType*s, and the OutcomeSubtype is not detailedly recorded, we decided not include *OutcomeSubtype* in the data analysis.

For the purpose of cleaning data and keeping consistency, we firstly tried to combine train and test dataset as a new combined dataset. We found that the name of ID column is different between train and test, so we renamed the *AnimalID* in train as ID, which is the same as the name in test. Then we created a new *OutcomeType* column for test dataset, the default value of which is NA. Finally, we used rbind function to create new dataset called combine.

As for the variable *Name*, we divided it into two groups based on if animal has a name or not, and the new variable is called *HasName*.

For the second variable *DateTime*, we extracted the value in different time units, such as hour(named as *DateTimeHour*), month(named as *DateTimeMonth*), year(named as *DateTimeYear*).

For the third variable *AnimalType,* we just let it be.

For the fourth variable *SexuponOutcome*, we found there are five groups (Intact Female, Intact Male, Neutered Male, Spayed Female, Unknown). Besides, one observation didn’t belong to any groups, so we assigned Unknown for that object. Then, we also created new variables called *SexIntact* and *SexGender* to indicate Intactness and gender groups.

For the fifth variable *AgeuponOutcome*, we tried to unify the unit. As we can see in the summary, the original data has four different units (day, week, month, year), and we dealt with it by unifying measurement as day. The corresponding new variable is called *AgeDay*. We also used rpart function to predict the missing AgeDay values based on other variables and replaced the missing value with it.

For the sixed variable *Breed*, we firstly extracted the Mix value out based on whether the animal is mixed or not and created a new variable BreedMix. Then, we did the breed classification based on American Kennel Group standards for dogs (http://www.akc.org/), and the length of hair for cats( http://www.catbreedslist.com/). As a result, we divided cats into four groups (Hairless, ShortHair, MediumHair, LongHair) based on the length of hair and divided dogs into eight groups (Herding, Hound, NonSporting, Sporting, Terrier, Toy, Working, Unknown).

For the seventh variable *Color*, we created three new variables *ColorType*, *ColorPattern* and *ColorMulti* to better interpret it. If the value has more than one colors, we assigned *ColorMulti* as 1, otherwise 0. Then we grabbed the first color split from *Color* as its *ColorType*. Additionally, we extracted the pattern value out (e.g. Brindle of “Yellow Brindle”, Tabby of “Cream Tabby”) and assign it to *ColorPattern*.

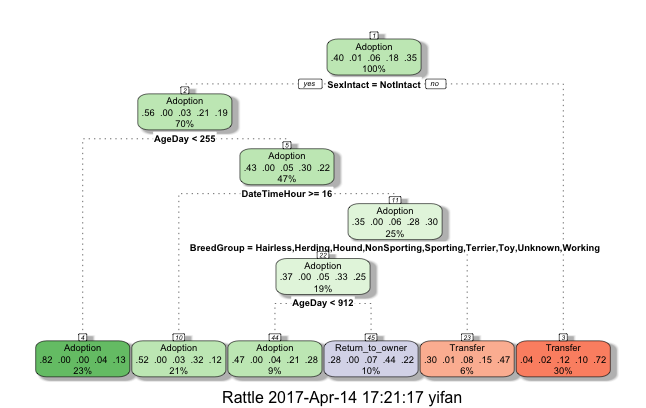
In the stage of building models, we created a new variable named as *TimeofDay* based on hour period of a day in *DateTimeHour*. Similarly, we classified *DatetimeMonth* values into 4 groups in a new variable called *Season*. We also simply divided AgeDay values into 2 groups in *AgeGroup*: adult( equal to or more than 365 days) and baby (less than 365 days).

Overall, apart from OutcomeType and ID, at last we have 16 variables: *AnimalType*, *SexIntact*, *SexGender*, *ColorType*, *ColorPattern*, *ColorMulti*, *BreedMix*, *BreedGroup*, *DateTimeHour*, *DateTimeMonth*, *DateTimeYear*, *AgeDay*, *HasName*, *TimeofDay*, *AgeGroup* and *Season*.

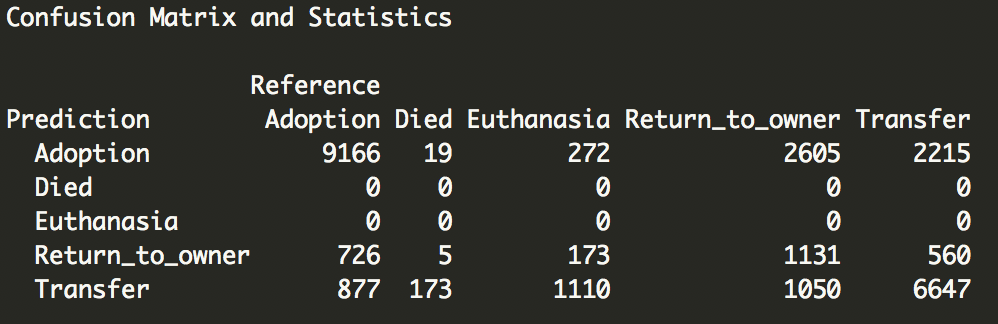
1. **Training results**
2. DecisionTree Model

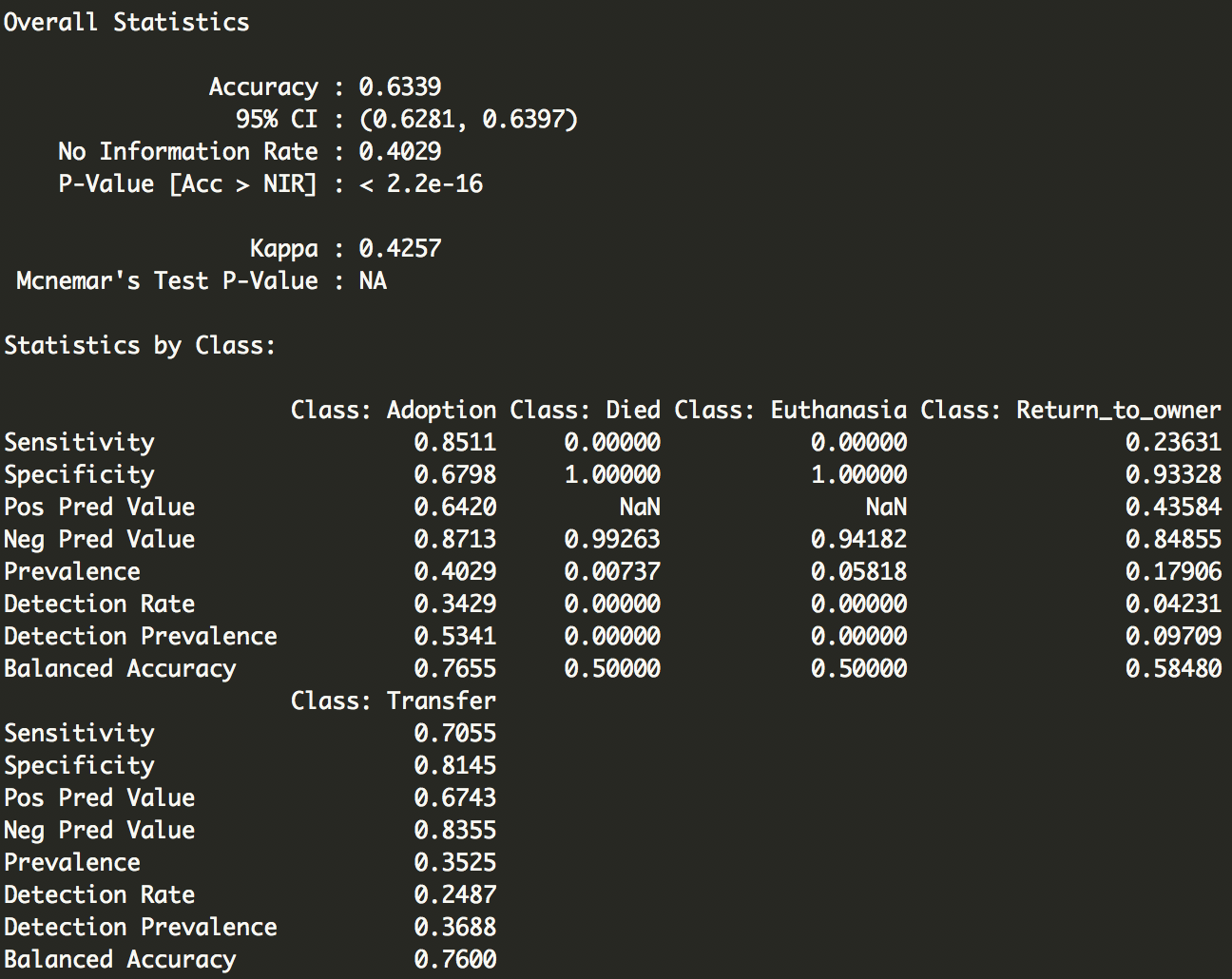
First, we build a decision tree model based on *AnimalType*, *AgeDay*, *SexIntac*t, *SexGender*, *ColorType*, *BreedGroup*, *HasName*, *AgeGroup*, *BreedMix*, *DateTimeHour*, *DateTimeMonth* and DateTimeYear. The overall accuracy is 0.6339. From Figure 13 Tree Plot, we can tell this model is inadequate to predict Died and Euthanasia cases. Besides, from Figure 14 confusion matrix, we can see that this model predicted Adoption while the true results were Return to owners and Transfer.

*Figure 13 Decision Tree Plot*

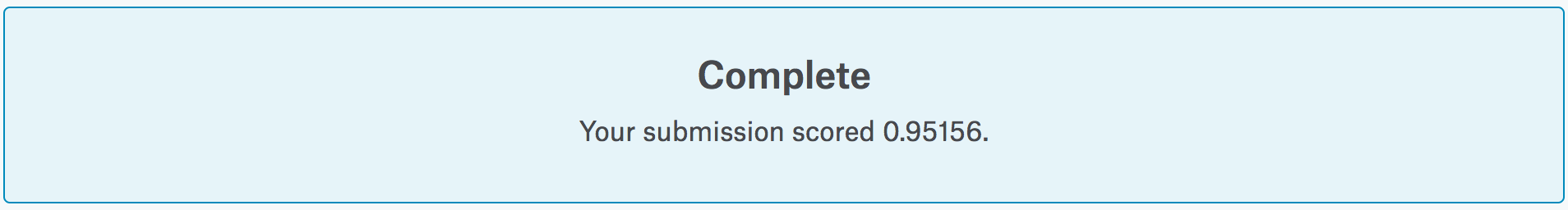
****

*Figure 14 Confusion Matrix of Decision Tree Model*





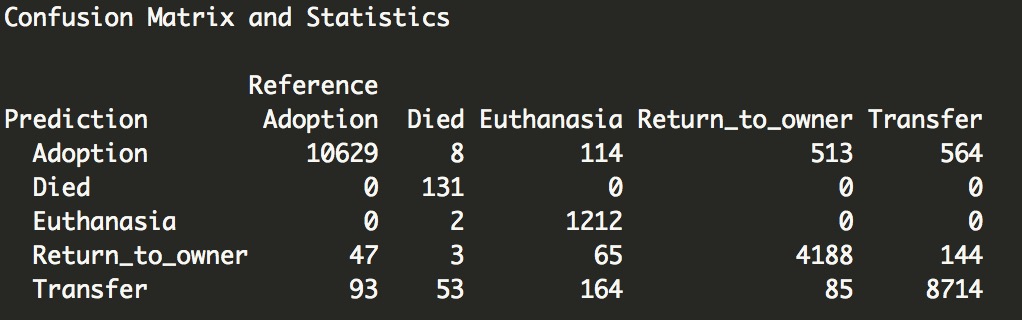
Kaggle Score

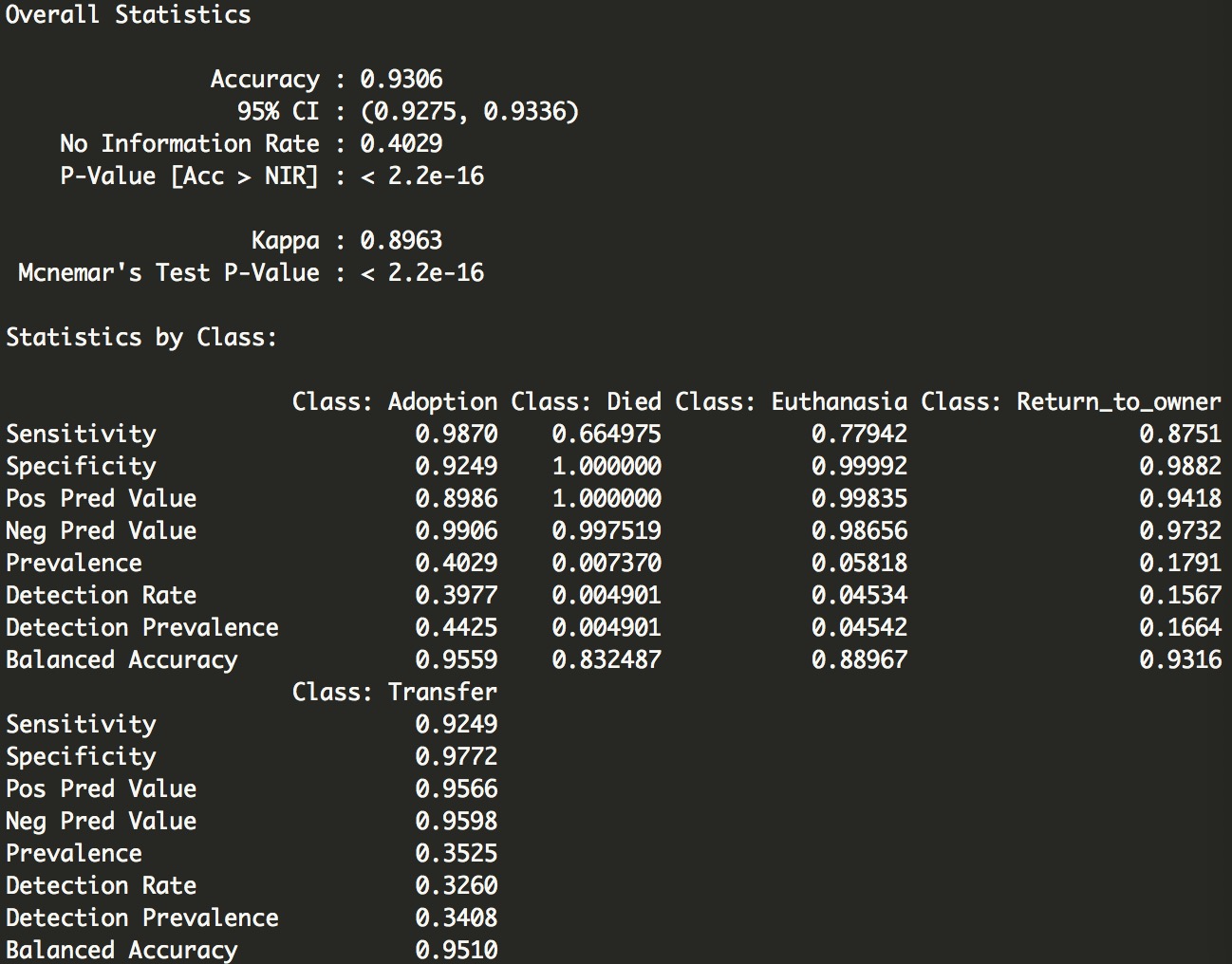


1. Random Forest Model 1

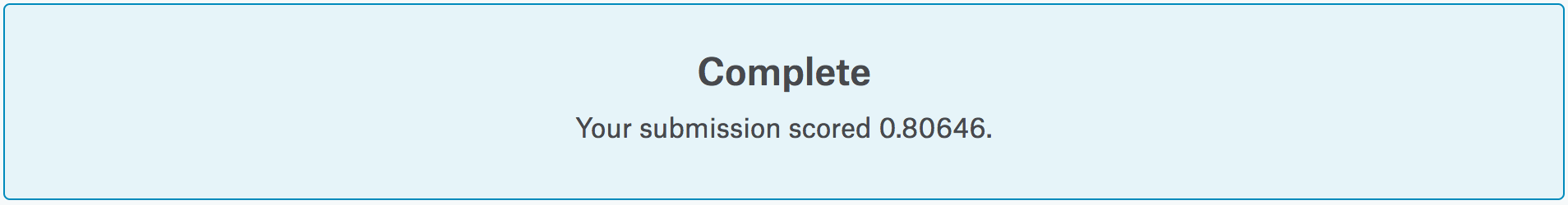
Further, we build a RandomForest model based on *AnimalType, AgeDay, SexIntact, SexGender, ColorType, ColorPattern, ColorMulti, BreedGroup, HasName, BreedMix, DateTimeHour, DateTimeMonth and DateTimeYear*. The accuracy of this model is 0.9306. Sensitivity and Specificity of this model are greatly improved compared to the decision tree model.

*Figure 15 Confusion Matrix of RandomForest Model 1*





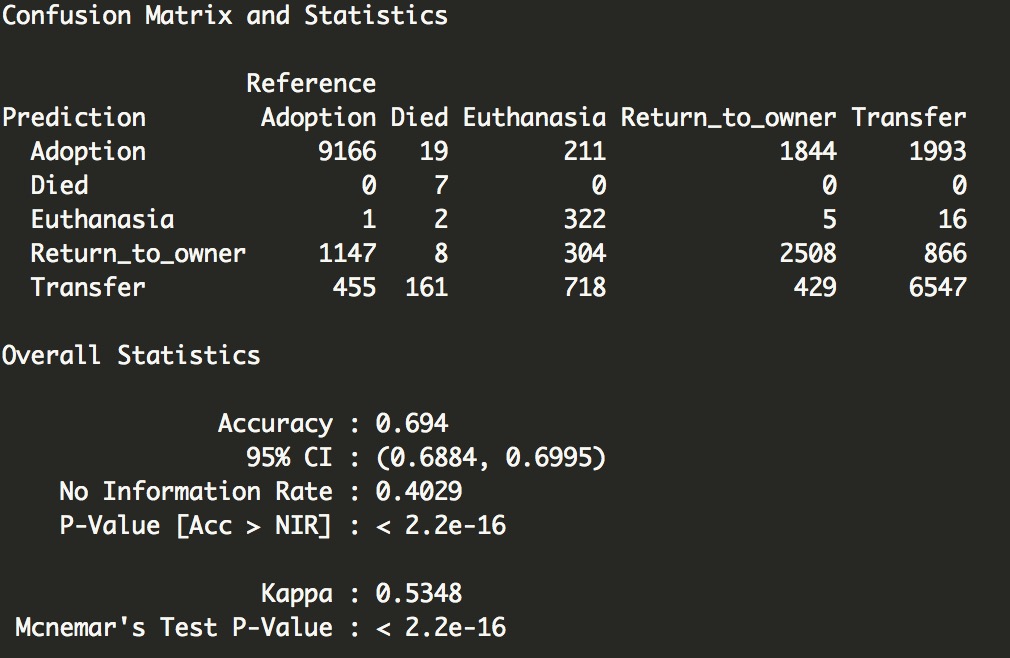
Kaggle Score

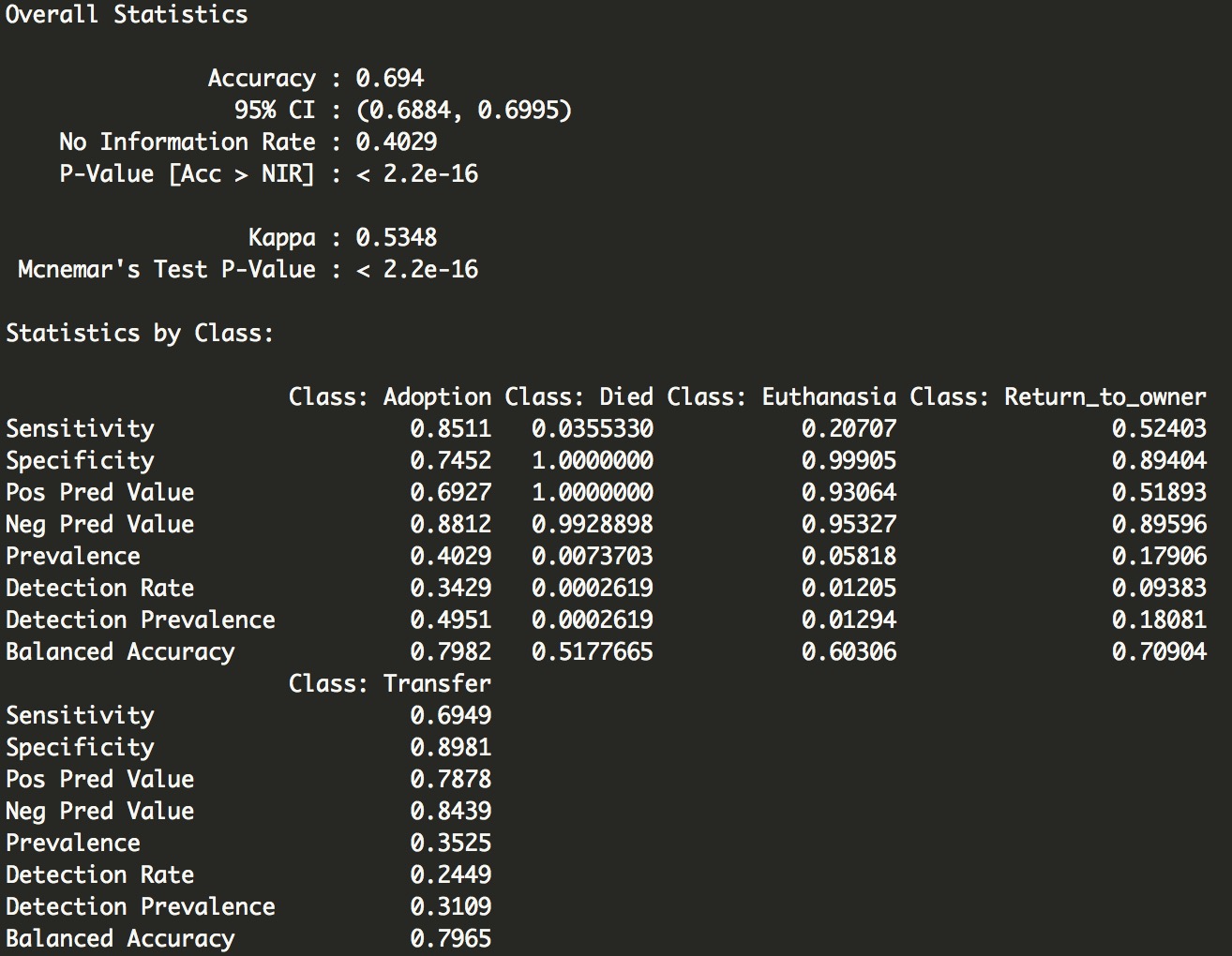


1. Random Forest Model 2

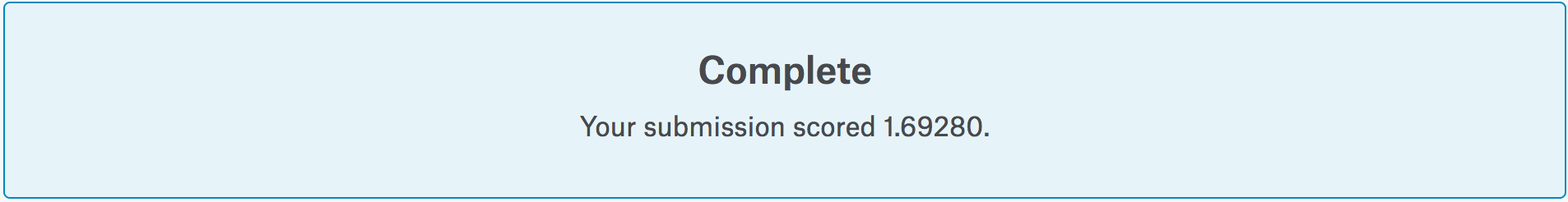
In our second RandomForest Model, we aggregated *DateTimeHour* into a new variable *TimeofDay* (namely morning, midday, lateday and night). Similar aggregation was implemented on *DateTimeMonth,* whichgenerated a new variable named as *Season* (spring, summer, fall and winter). Additionally, we divided *AgeDay* into two groups: baby ( less than 365 days ) and adult ( equal to or more than 365 days). We conjectured that the aggregation may further improve the accuracy of prediction. However, as the confusion matrix in Figure 16 shows, the accuracy rate, sensitivity and specificity of this model are surprisingly much lower than the statistics in the models without these aggregated variables, which are opposite to our assumption.

*Figure 16 Confusion Matrix of RandomForest Model 2*





Kaggle Score

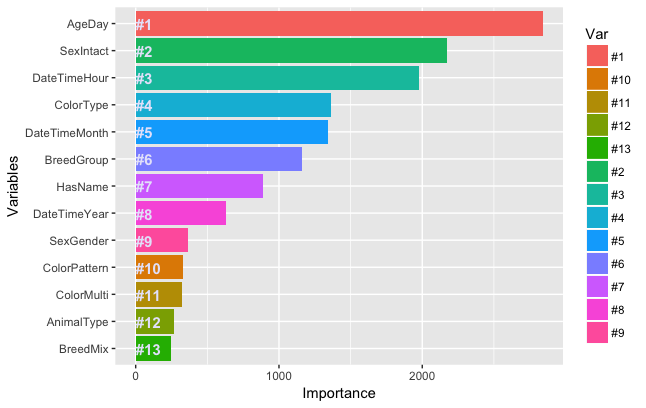


**Limitations**

After building models, we found that grouping dogs’ breed based on the information from American Kennel Club might be not intuitive to most people, and thus this grouping variable appeared not so important as we thought at the beginning. Future exploration can be done if we can have more information on dogs’ breed grouping in terms of intuitive grouping such as charismatic type and non-charismatic type. Meanwhile, we did not spend too much efforts on processing color type and just grab the first color as the main color type. If we could have color information based on RGB values, we might further group the color values as light, medium and dark shades.

**Conclusion**

*Figure 17 Variable Importance in Random Forest Model1*

****

Based on the plot of variable importance in the Random Forest Model1, we can infer that the age of an animal is the most important factor in his/her fate. Backing to the previous data exploration section, we can clearly see that animals aged less than one year old are more likely to be adopted. Among animals aged more than one year old, 1/3 were returned to owner, which shows quite a part of adult animals in shelter have owners before. Secondly, it seems that animals without intactness have more chances of getting adopted. The *DateTimeHour* variable also contributes to animals’ fate, though we didn’t have clear idea about what does the date value stand for. From Figure 18, we can see that dogs with *DateTimeHour* value between 7 to 10 are more likely to get euthanatized than dogs with other hour values. Adoption is concentered on dogs with hour value between 5 to 7 and 18-21. Similarly, adoption occurred most on animals with *DateTimeMonth* value between 7 and 8.

Meanwhile, we can conclude that whether an animal is mixed or not does not greatly impact his chance of getting adopted as we thought it might be. Both dogs and cats in this shelter have almost equal chance of getting adopted. In addition, the pattern of animals’ coat and its multicolor does not play a big role in their fates either.

*Figure 18*

