Final Paper on Animal Shelter Data

R Statistical Programming

Fall 2019

Xiaoxuan Wang, Yi Wan

Demetra Koumatos, Jiang Du

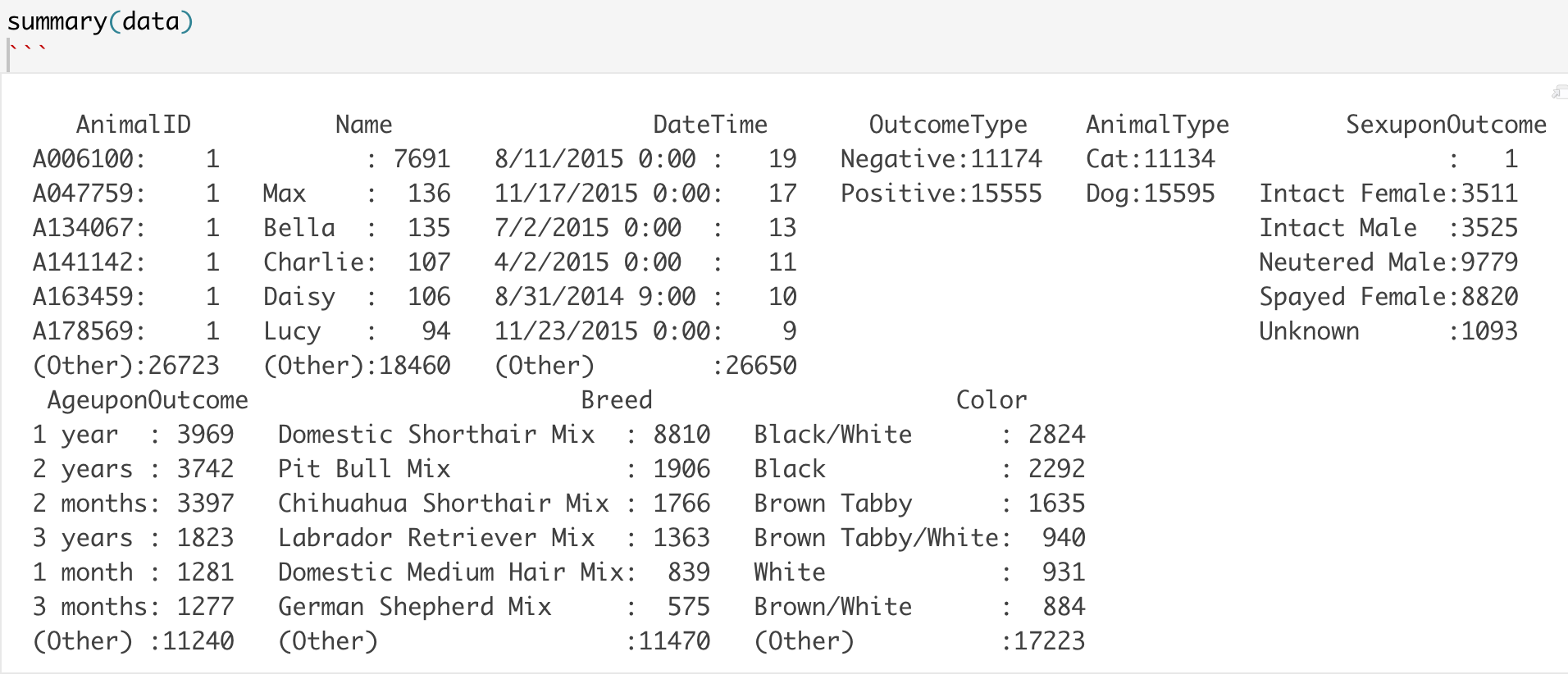
**Intro**

For our final project, we analyze animal shelter outcome data to better understand animal adoption trends. Specifically, we want to predict whether a shelter animal has a positive (adoption, returned to the original owner) or negative (euthanized, transferred to other shelter) outcome based on information on the animal such as breed, gender, and age. We intend to solve this problem by fitting part of the data to a classification model, and use the rest of data to test the model performance.

**Pre-Analysis**

The original Animal Shelter dataset contains 27,629 records of 9 fields. The 9 fields are listed below along with a description of each field.

|  |  |
| --- | --- |
| **Field** | **Description** |
| AnimalID | Animal identifier |
| Name | Name of animal |
| DateTime | Date and time of outcome |
| OutcomeType | Outcome for animal flag |
| AnimalType | Animal type (dog or cat) |
| SexuponOutcome | Gender and neuter/spay info |
| AgeuponOutcome | Age of animal |
| Breed | Breed of animal |
| Color | Color of animal |

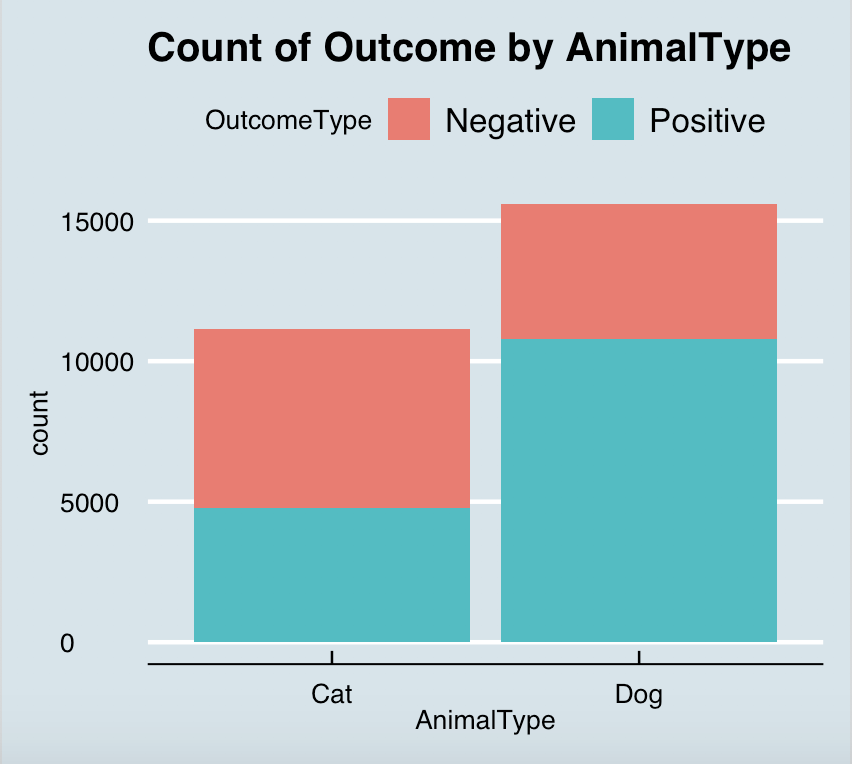
Summary statistics:The output of the summary statistics by group shows the count by category for each of the 9 variables. One major point from this output is that our target field **OutcomeType** contains a breakdown of 40% Negative / 60% Positive outcomes, meaning that in our dataset, animals are more likely to have a Positive outcome than not. Additionally, the **AnimalType** field shows us that the data contains a breakdown of 40% Cats / 60% Dogs.

**Feature Development**

Before we could use our data to create a model, we had to do some feature development on the variables/fields. We changed the following predictor variables to make them have more impact on our model.

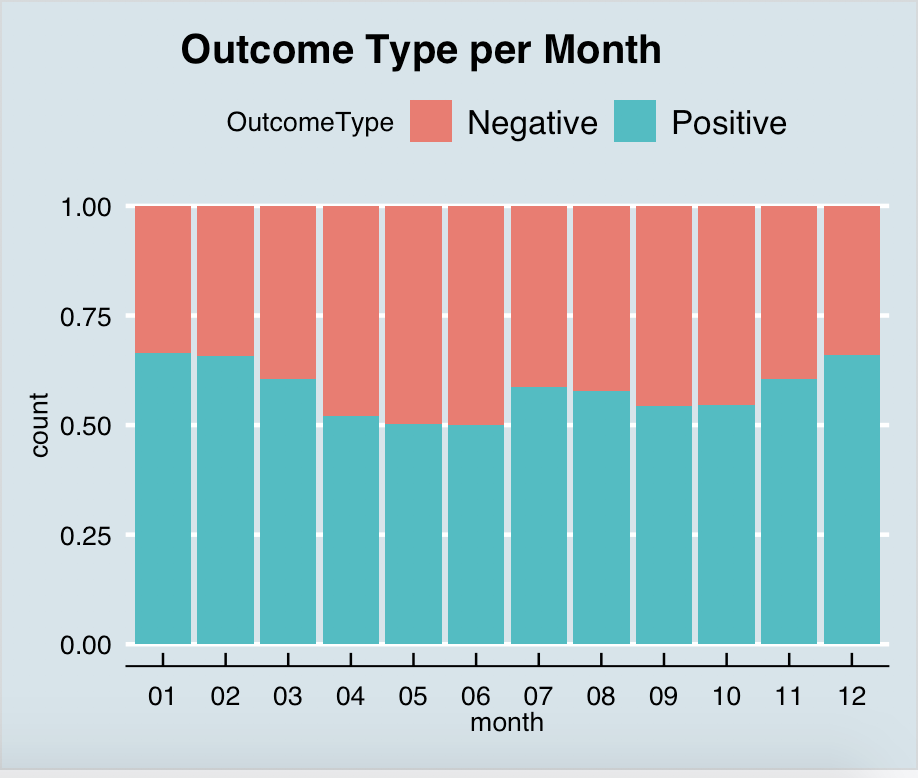
The field **DateTime** was split into four fields (Year, Month, Day, DayofWeek). We split this field because we felt that the year, month, day, and day of the week might give us some important insights on their own such as seasonality of adoption. For example, during the Christmas season, there might be an increase in adoption rates. We also created a new column Weekday for each day to see if there are any trends within a week. For instance, during the weekends, people might be more likely to come to the animal shelter to adopt animals. The field **SexuponOutcome** originally contains 5 levels: Intact Female, Intact Male, Neutered Male, Spayed Female and unknown. We split them into two fields (Neutered and Gender) and then further simplified the Neutered field into two categories: Intact and Non-Intact (Neutered/Spayed), and save them as a new field **Intact**. The field **AgeuponOutcome** was originally in units of days, weeks, months and years, so we used these values to calculate the age of the animals all in the same units. We engineered this field into a field called Age in units of years. The field **Breed** was summarized into two categories (BreedType and Purebred), and saved as a new field **BreedType**. We split this field because we felt whether the animal is a mix or a purebred may have an effect on the outcome of adoption and therefore should be considered separately.

**Initial Plots**

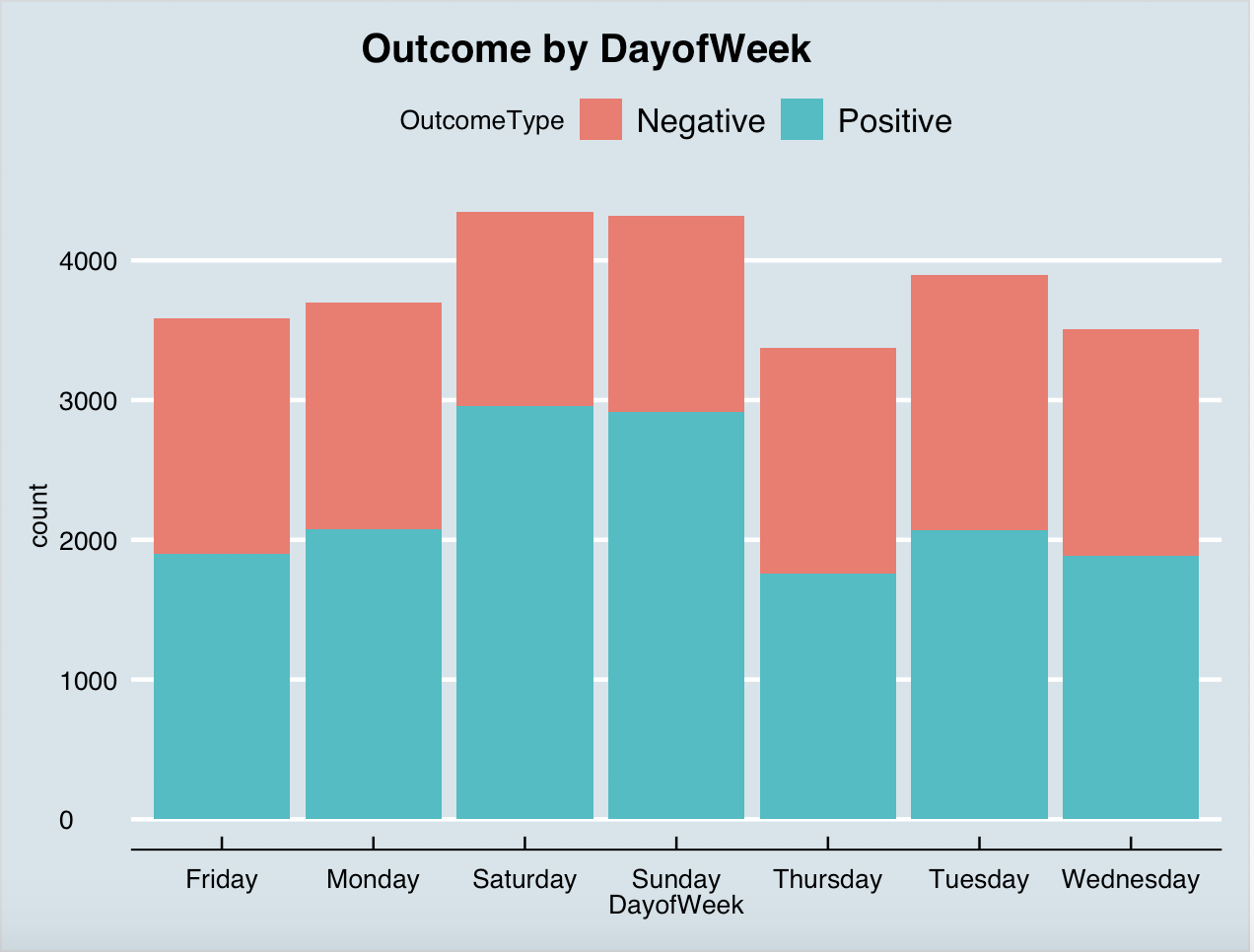


|  |  |  |
| --- | --- | --- |
|  | Cat | Dog |
| Negative | 57.14% (6362) | 30.86% (4812) |
| Positive | 42.86% (4772) | 69.14% (10783) |
| Total | 11134 | 15595 |

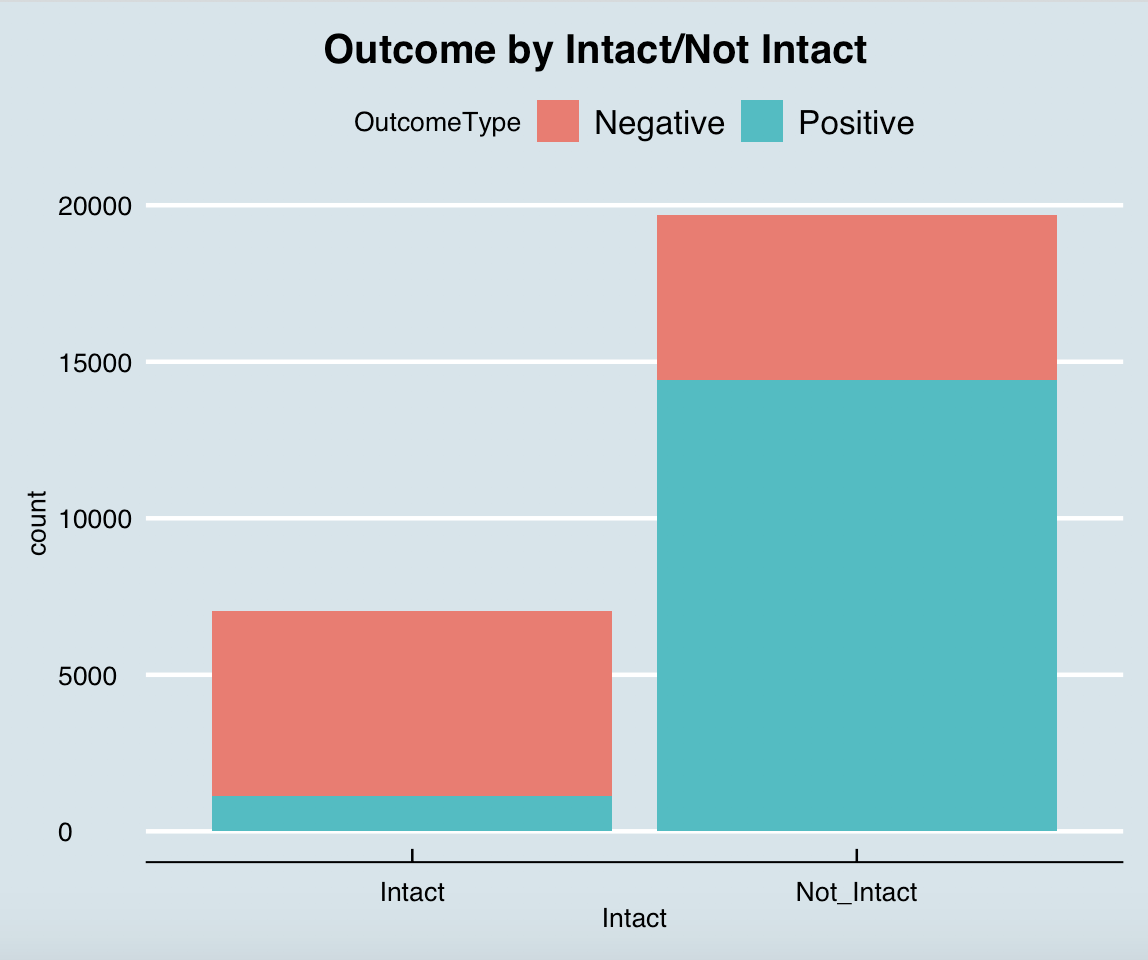
The first plot we used to analyze our data was a count of **OutcomeType** by **AnimalType**. Visually we can see that Dogs are more likely to have a positive outcome whereas Cats are more likely to have a negative outcome. More specifically, a Dog has about a 69.14% chance of a positive outcome whereas a Cat only has about a 42.86% chance to have a positive outcome.



The next plot we used to analyze our data was of **OutcomeType** per **Month**. Visually we can see a slight increase in positive outcomes in the months of December, January, July and August, hinting that the holiday season might be correlated with higher adoption rates, but since correlation does not equal causation, we are careful to note that this is just an observation at this point. We will keep this slight trend in mind as we analyze our data further, but we do not expect the seasonality to be a very important predictor in our model.

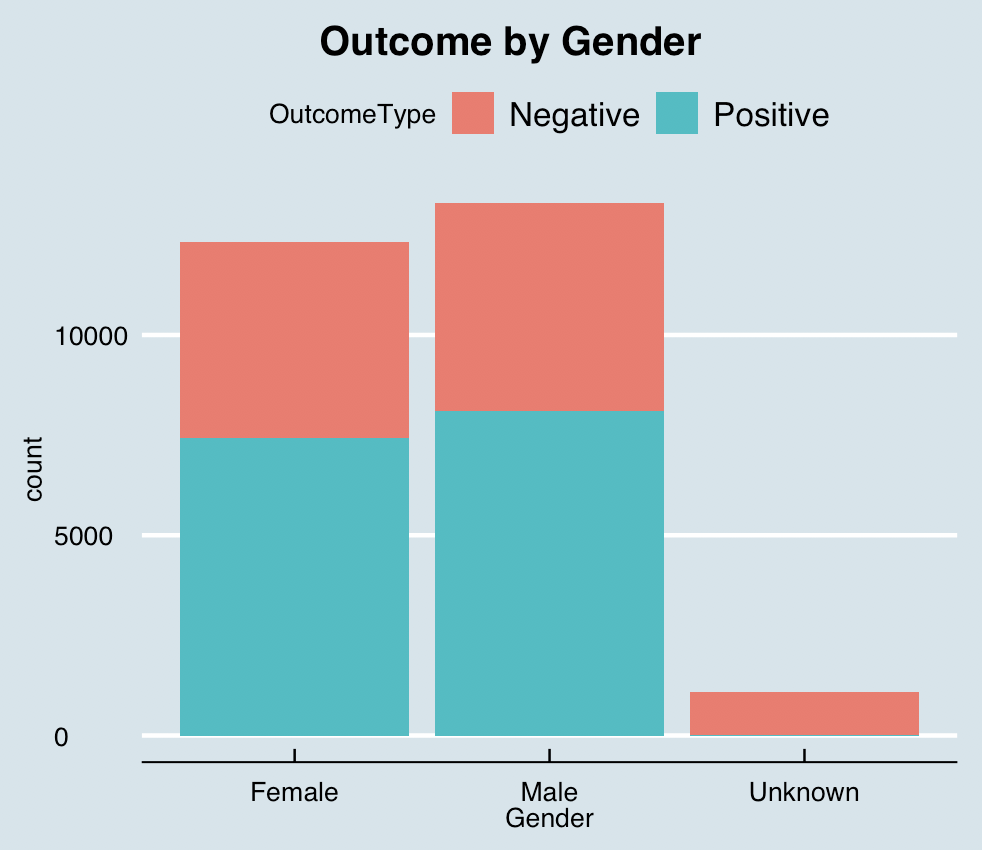


Another plot we used to analyze our data was Outcome by Day of the Week to see if certain days of the week lead to more positive outcomes. We can see that Saturday and Sunday are the most popular days of the week for adoption.



|  |  |  |
| --- | --- | --- |
|  | Intact | Not\_Intact |
| Negative | 83.81% (5897) | 26.80% (5277) |
| Positive | 16.19% (1139) | 73.20% (14416) |
| Total | 7036 | 19693 |

We also used a plot to see if the fact that an animal was intact or not intact(neutered/spayed) was related to outcomes. We can see that the overall number of Intact animals is much smaller than Not\_Intact animals. Within these categories, the animals who were not intact have about a 73.2% positive outcome rate whereas the Intact animals have only about an 16.19% positive outcome rate.

The final plot we used to explore our data was Outcome by Gender. It shows all the animals with unknown gender had negative outcomes; while gender being female or male didn’t make much difference to the outcome.

**Initial Hypothesis**

Based on our exploratory analysis, we predict that the most important factors in determining whether an animal gets adopted will be **AnimalType**, **Intact and Gender.**

**Modeling**

We split the data set into a training set and a testing set, containing 75% and 25% of the total observations respectively. We used the logistic, ridge logistic and lasso logistic regression model because the response was a qualitative variable with two levels, positive and negative. We chose AnimalType, Gender, month, Age, BreedType, DayofWeek, and Intact as independent variables.The p-values tell us which independent variables are significant at a certain confidence level.

Table 1: significant predictors

|  |  |  |  |
| --- | --- | --- | --- |
| Significant  Predictors | Logistic (95% CI) | Ridge Logistic | Lasso Logistic |
| **AnimalTypeDog** | **AnimalTypeDog** | **AnimalTypeDog** |
|  | GenderMale |  |
| **GenderUnknown** | **GenderUnknown** | **GenderUnknown** |
|  | month02 |  |
| month3 | month03 |  |
| month4 | month04 |  |
| month5 | month05 |  |
|  | month06 |  |
| **month7** | **month07** | **month7** |
|  | month08 |  |
|  | month09 |  |
| month10 | month10 |  |
| month11 | month11 |  |
|  | month12 |  |
| **Age** | **Age** | **Age** |
|  | BreedTypePurebred |  |
|  | DayofWeekMonday |  |
| **DayofWeekSaturday** | **DayofWeekSaturday** | **DayofWeekSaturday** |
| **DayofWeekSunday** | **DayofWeekSunday** | **DayofWeekSunday** |
|  | DayofWeekThursday | DayofWeekThursday |
|  | DayofWeekTuesday |  |
|  | DayofWeekWednesday |  |
| **IntactNot\_Intact** | **IntactNot\_Intact** | **IntactNot\_Intact** |

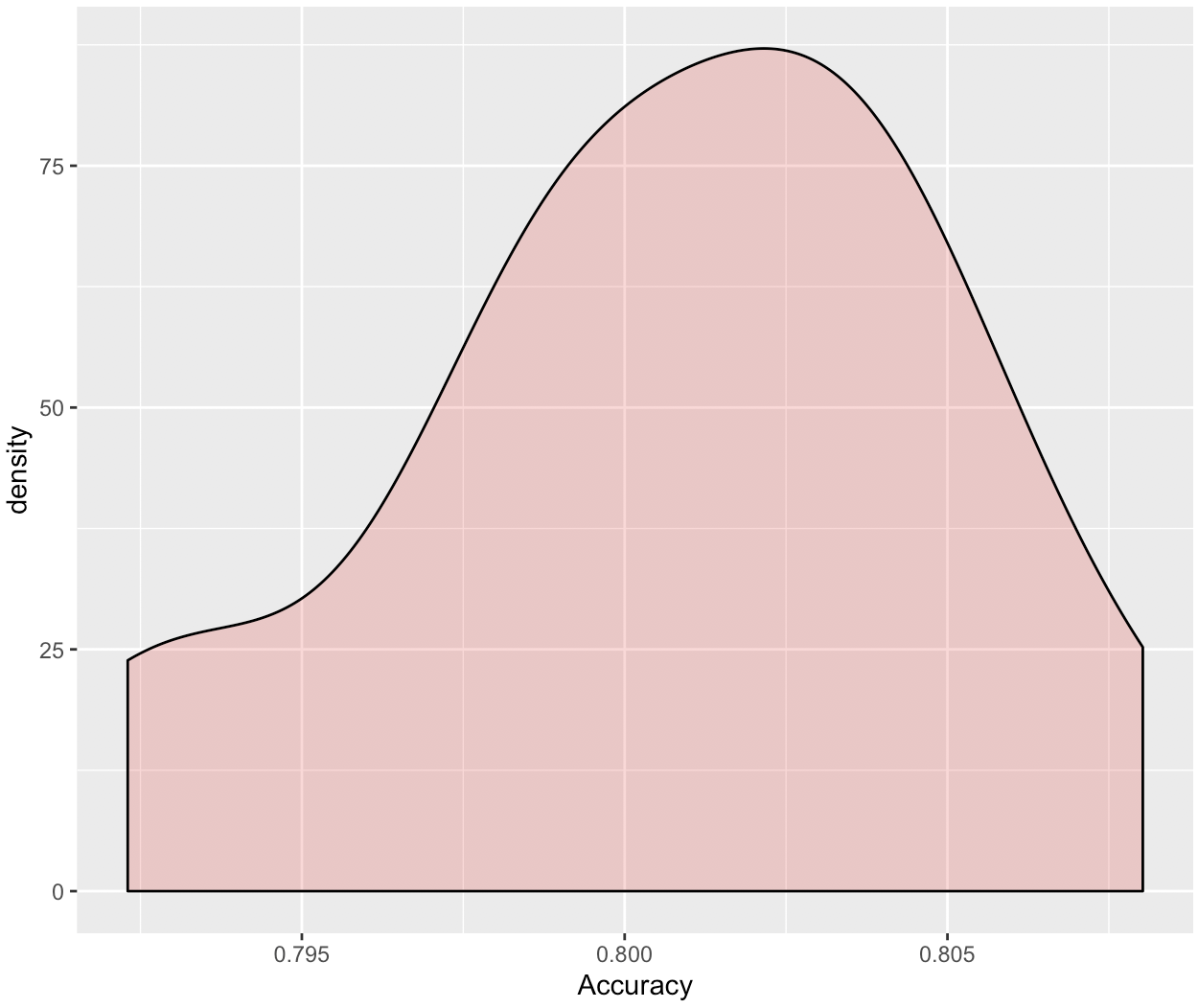
AnimalType, Gender, month, Age, DayofWeek and Intact are predictors that have an influence on shelter outcome in all three models. A positive coefficient of a quantitative predictor means the log odds of the outcome will increase if the predictor increases. A positive coefficient of a qualitative predictor means the log odds of the shelter outcome will increase if the predictor is of the corresponding type. Vice versa for the negative coefficients. For instance, in the lasso model, the log odds of the outcome will decrease 0.009 as the age of the animal goes up by 1, holding other variables constant; if the animal is a dog, then the log odds of the outcome will be 0.555 higher than if it’s a cat, keeping other variables constant.

Table 2: Coefficients of Significant Predictors

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic | Ridge | Lasso |
| **AnimalTypeDog** | **0.685** | **0.605** | **0.555** |
| **GenderUnknown** | **-5.161** | **-2.871** | **-3.767** |
| **month7** | **0.172** | **0.156** | **0.105** |
| **Age** | **-0.031** | **-0.013** | **-0.009** |
| **DayofWeekSaturday** | **0.675** | **0.454** | **0.472** |
| **DayofWeekSunday** | **0.558** | **0.382** | **0.369** |
| **IntactNot\_Intact** | **2.830** | **2.273** | **2.662** |

The table above indicates that the results we get from all three models are consistent. It can be seen that for animal type, dogs tend to have positive outcomes which confirms our initial hypothesis. GenderUnknown, month7 and age have negative impacts on shelter outcome with GenderUnknown having the biggest impact. On weekends, it is more likely for animals to get adopted. Not intact animals tend to have positive outcomes which also aligns with our initial guess. This is reasonable since neutered/spayed animals are more docile and less susceptible to diseases.

For the logistic regression, we used 5 repeated 3-fold cross-validation to build the logistic model and plotted the accuracy distribution shown below. The average training accuracy was 80.08%.

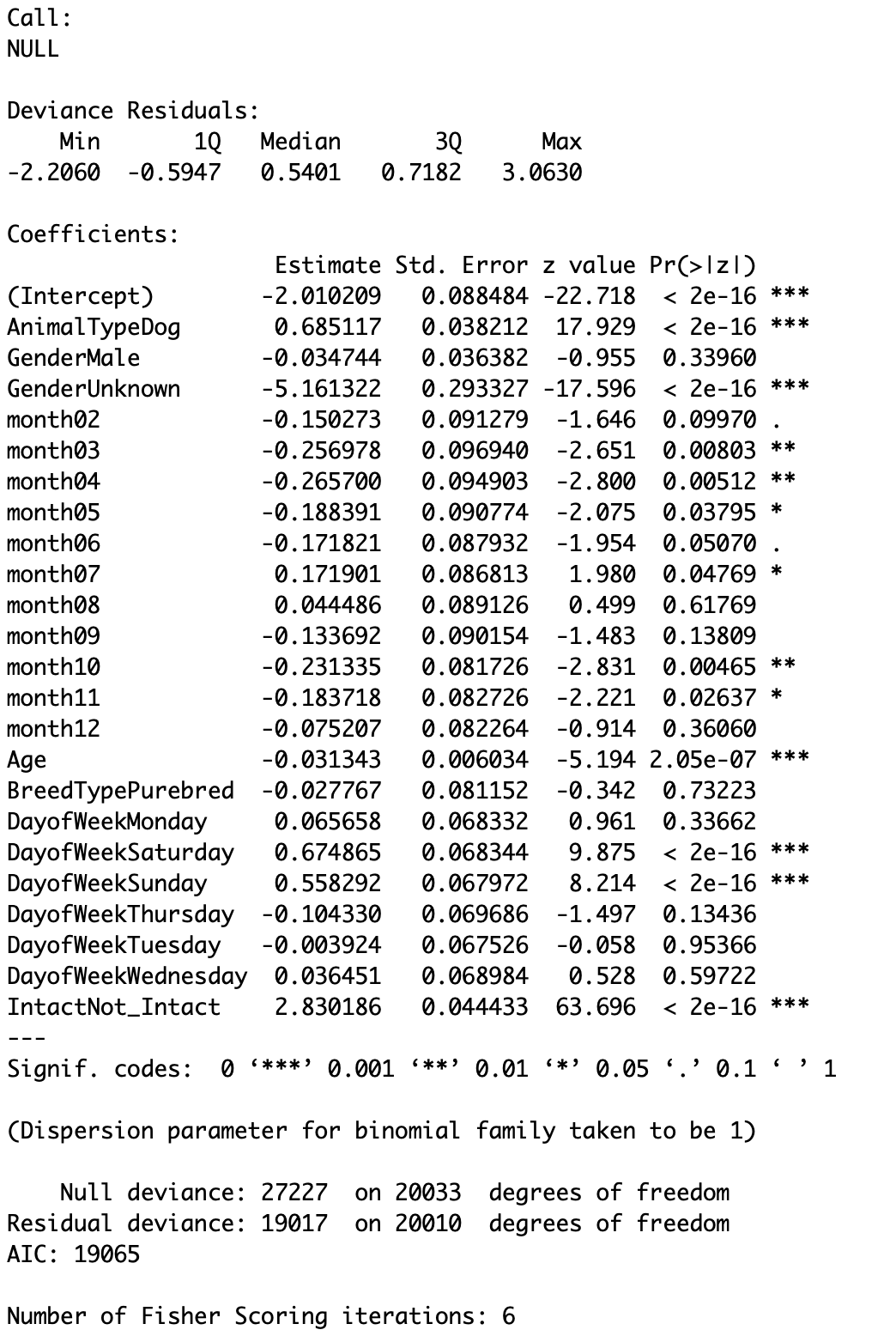


Comparison of Training Results

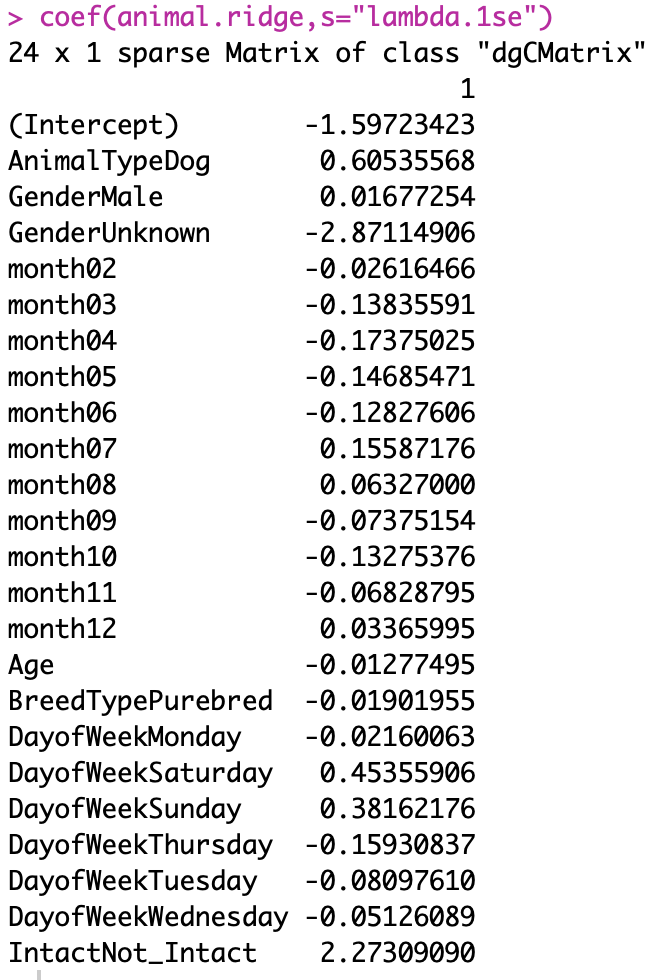
|  |  |  |
| --- | --- | --- |
|  | Accuracy | AUC |
| Logistic | 80.08% | 0.824 |
| Ridge | 79.31% | 0.8195 |
| Lasso | 80.07% | 0.8198 |

**Model Output**

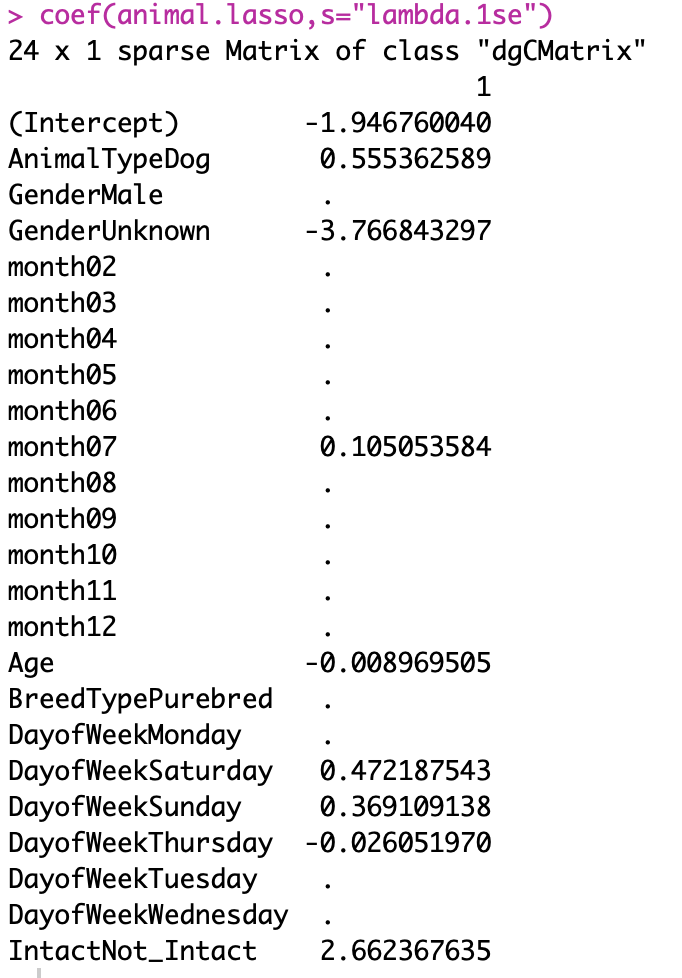
Logistic Regression:



Ridge Regression:

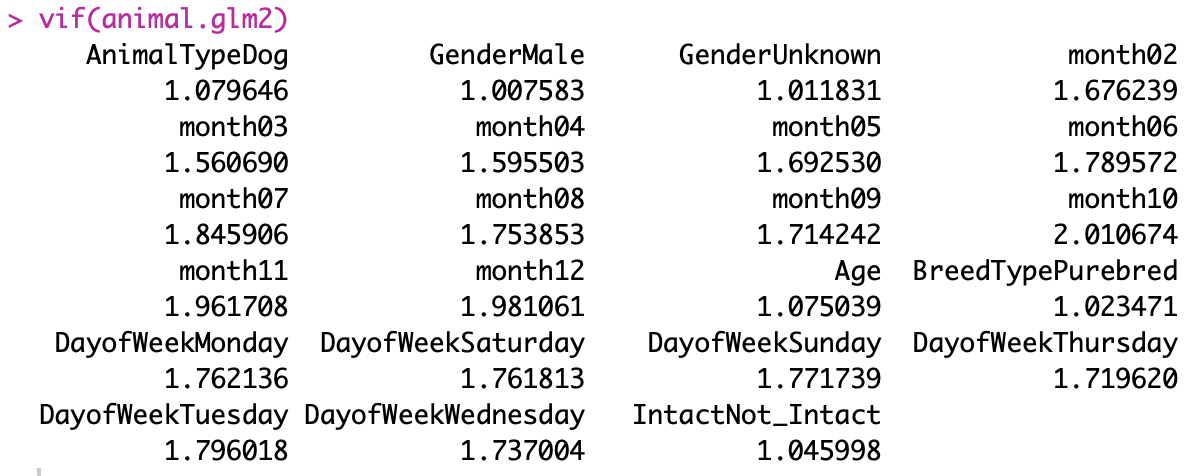


Lasso Regression:



**Model Assumptions**

Our model meets all 4 assumptions for the logistic regression. The response is binary with two classes, positive and negative. The observations are independent, since different adoption cases usually do not affect each other. Since VIF for all independent variables are below 5, there is no strong multicollinearity relationship between independent variables. Finally, the sample size is quite large, with 20,048 observations.



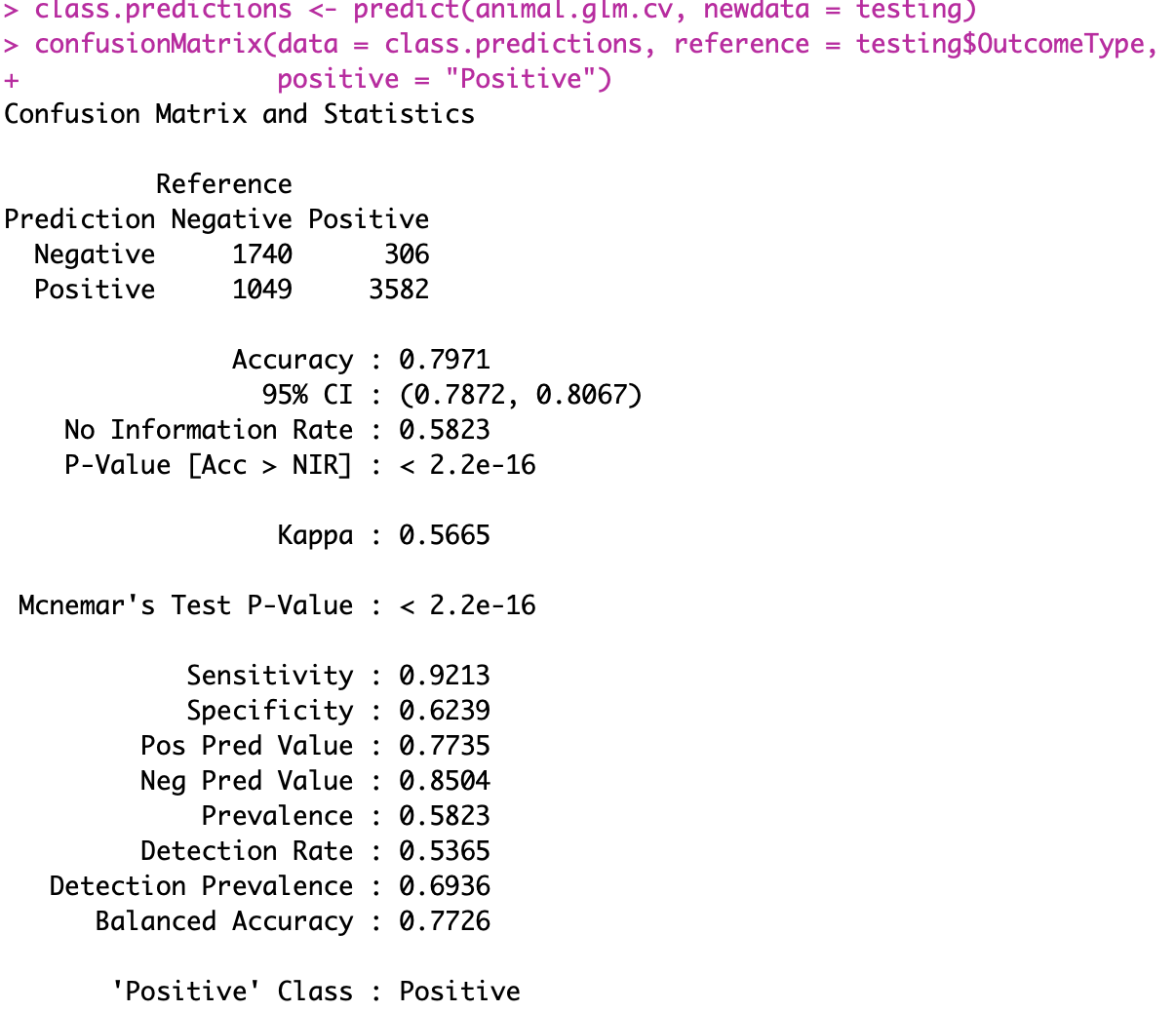
Mo

**Model Validation (Testing Accuracy)**

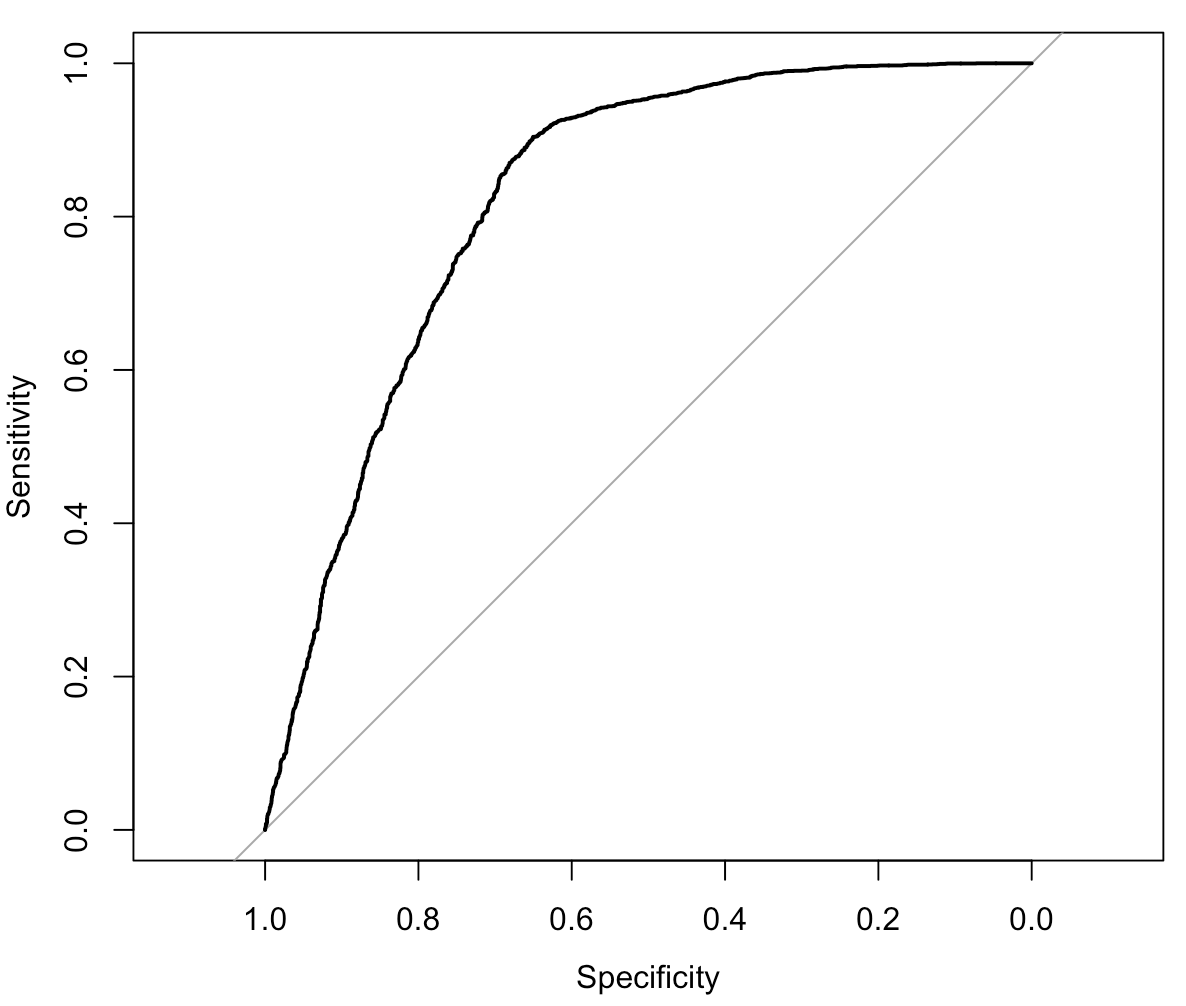
|  |  |  |
| --- | --- | --- |
|  | Accuracy | AUC |
| Logistic | 79.71% | 0.8239 |
| Ridge | 79.06% | 0.8204 |
| Lasso | 79.72% | 0.82 |

The above table shows that the testing results from Logistic, Ridge, and Lasso models are very close to each other, with accuracy being around 79% and AUC being around 0.82.

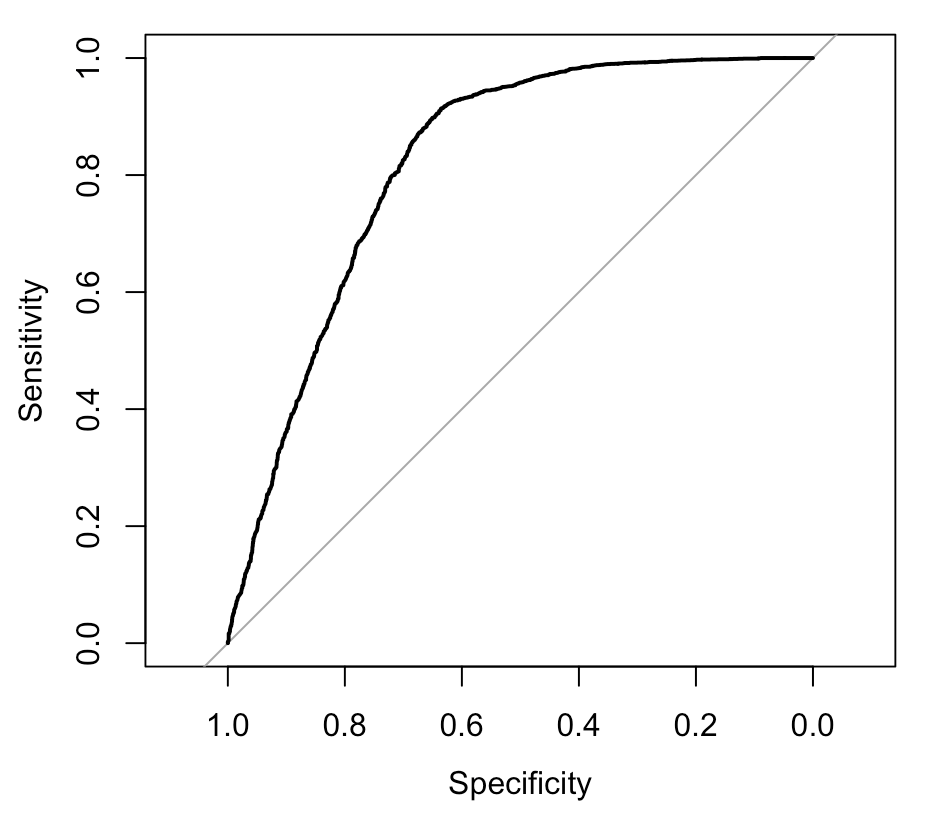
Logistic Regression:



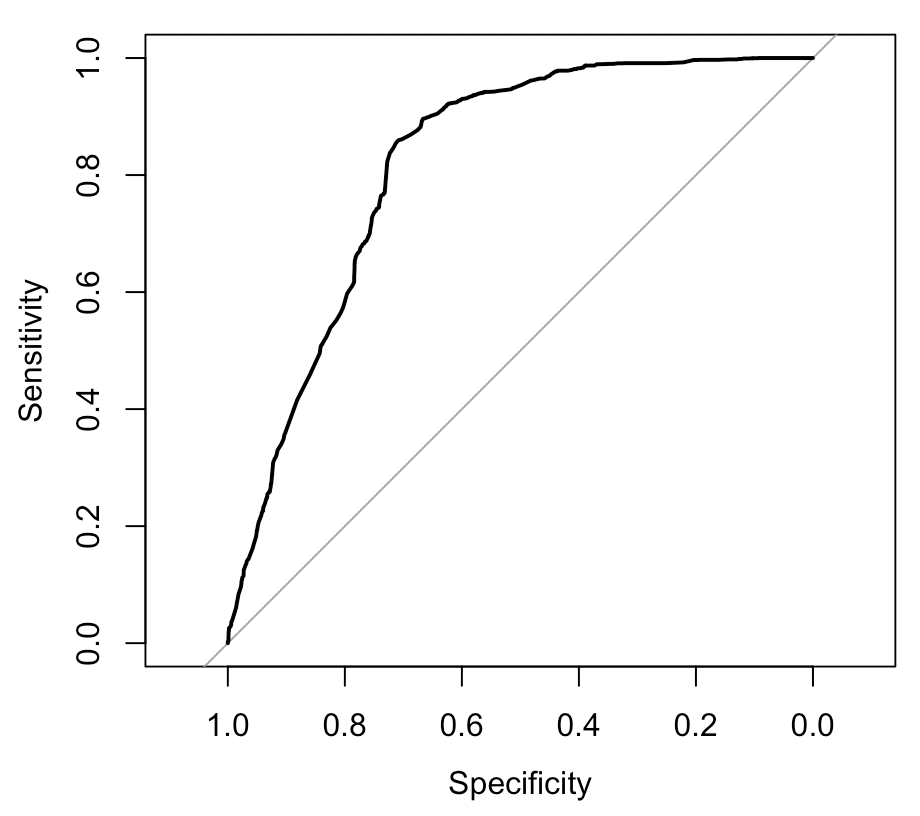
The AUC is 0.8239.



The Ridge logistic regression model had a Testing accuracy of 79.06% and Testing AUC is 0.8204.



The Lasso logistic regression had a Testing accuracy of 79.72% and Testing AUC of 0.82.

****

**Overfitting**

The testing accuracy was only slightly lower than the training accuracy in all three models. The testing AUC actually improved in the ridge and lasso model. The models performed well on the testing set and there was no overfitting.

**Conclusion**

From our analysis, we found AnimalType, Gender, month, Age, DayofWeek and Intact are predictors that have an influence on shelter outcome. Age and unknown gender have a negative impact on the shelter outcome. If the animal is a dog and is not intact, it is more likely to be adopted. Also, adoption rates are higher on weekends and in June. Outputs from Logistic, Ridge, and Lasso are consistent and the accuracy is about 80% which might indicate that the models fit the data well.

**Future Improvements**

One thing we could do to improve our analysis is by applying further transformations to extract information from the independent variables such as creating dummy variables for the most popular breeds to see if certain breed types have higher chances of getting adopted. Additionally we could look further into the fact that animals with unknown gender all had negative outcomes to see what else about this group may have caused them to have negative outcomes. It could be that people forgot to record, resulting in bad data, not that this variable actually had an impact on the outcome. We might as well just delete those observations. We could also improve our model by collecting additional data on the animals such as their length of stay from when they were admitted to the shelter to when they got adopted.