Dallas Animal Services Statistical Consultation

Zherui Lin and Rob Pruette

Southern Methodist University

Department of Statistical Science

Executive Summary

Dallas Animal Services (DAS) is the only open-admission shelter in Dallas. DAS has an open-door policy for animals, regardless of capacity or the animal's behavior. With such a large intake of animals and microchip technology, DAS collects immense amounts of data. Microchips allow DAS to collect information on each animal that comes through the shelter.

The goal of this consultation was to provide DAS with insight about animals that are repeatedly impounded. Although DAS expects a high number of animal intakes, the organization wants to minimize the number of animals that come through their shelter repeatedly. Therefore, with useful insight into trends surrounding repeat impounds, DAS could make procedural modifications to optimize efficiency.

DAS provided three separate data files. These files separately contained information about impounds, bites, and violations in the Dallas area. Although bites and violations may have been interesting variables when exploring repeat impounds, the datasets were nearly 95% smaller than the impounds data. Therefore, only the impounds data were included in this analysis. Exploratory plots, however, were generated for the other datasets.

A binary outcome variable was created labeling each observation as a single impound or a repeat impound. To predict repeat impounds, the following variables were included: animal's sex, type of animal, longitude of capture or owner's home, latitude of capture or owner's home, type of impound, year, month, and age of animal. Implementing the use of optimal cut points and weighted class variables, three different model types were fit including logistic regression, random forests, and neural networks.

Ultimately, all three model types produced similar results. Four different metrics were used to evaluate the models' performance: accuracy, sensitivity, specificity, and the F1 score. Since repeat impounds were underrepresented in the data, higher levels of sensitivity were desirable since this determines how well the model will predict repeat impounds. The final random forest model achieved the highest accuracy while still maintaining relatively high levels of sensitivity. The logistic regression model achieved the lowest values for all four metrics.

Although the logistic regression model did not outperform the other models, it still achieved an accuracy of 81.3% and a sensitivity of 48.9% (compared to the random forest's 84.0% and 50.1%). The logistic regression model offers the competitive advantage of interpretability in this analysis. Since DAS is more interested in trends among repeat impounds rather than pinpoint accuracy, this model can provide more insight into how different variables are associated with repeat impounds.

Based on the logistic regression model results, animals that are associated with higher probabilities of being repeatedly impounded are animals that are neutered/spayed, dogs, or have been impounded due to an owner surrender or transfer. Furthermore, animals that are associated with higher probabilities of *not* being repeatedly impounded are animals that picked up or located further north in Dallas, or have been impounded as a stray animal.

This analysis is limited by only using the impounds data and not incorporating data regarding bites and violations, which could provide some valuable insight into trends among repeat impounds.

In addition to model results and variable importance plots, exploratory plots were created for all the provided data and R Shiny Apps. These apps allow the user to explore impounds, bites, and violations, as they pertain to location, using Google maps.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Located in West Dallas, Dallas Animal Services (DAS) is the only open-admission animal shelter in the city of Dallas. Open admission means animals are admitted to the shelter "regardless of the current capacity or the health and behavior of the animal" (Dallas Animal Services, 2020). DAS has the "third highest dog and cat intake in the country with more than 39,000 cats and dogs" entering their care for the fiscal year of 2019 (Dallas Animal Services, 2020).

Not only does DAS seek to provide care to the animals in their shelter, but the organization also serve as a resource to the community. DAS provides education about animal care and also offers services such as vaccinations and microchipping.

Data are collected on animals that are brought into the shelter in addition to animals that are picked up by officers in the community. Any animal that comes through the DAS shelter is microchipped if they have not been microchipped previously. The microchip allows each animal to be uniquely identified and therefore, extensive data to be collected.

For this consultation, DAS has provided data related to animals that have been impounded between November 2016 and September 2020. The purpose of this consultation is to identify trends in the data related to all intakes of dogs and cats in addition to trends among repeat impounds. Repeat impounds are impounds in which the animal has been impounded multiple times. DAS is particularly interested in trends among repeat impounds because if animals are repeatedly coming to the shelter, they may need to make procedural changes in order to be more effective. DAS also provided data on bites and violations.

For this consultation project, we used R (R Core Team, 2020) for data preparation, visualization, and analysis.

2. Methods

The following steps were completed for statistical analysis of the data: preprocessing, imputation of missing data, transformation of variables, split data for validation, fitting statistical and machine learning models, and implementation of optimal cut points.

2.1 Data Preprocessing

The raw *DLLS impound repeats.csv* dataset consists of 132,401 observations and 14 attributes. The raw *DLLS BiteStats.csv* dataset consists of 5,631 observations with 28 attributes and the raw *Violation Notices.csv* dataset consists of 26,074 observations with 8 attributes. A data dictionary for the impounds data is given in *Table 1*.

Table 1
Impounds Data Dictionary

Variable Name	Description
Impound	Unique ID for each impound
Animal	Unique animal ID
In_Date	Date animal was impounded
Type	Type of impound
crossing	Address where animal was picked up, where the animal was surrendered, or where the owner lives
ZipCode	Zip code from address
sex	Sex of the animal

animal_type	Type of animal
age_now	Current age of animal
Prev_Intake	Impound ID from previous intake
Prev_InDate	Date from previous intake
Prev_InType	Impound type from previous intake
Prev_ZipCode	Zip code from previous intake
Prev_Addr	Address from previous intake

Variables from each dataset were preprocessed in various ways. These variables include location, date, age, and animal type.

Each dataset contained the animal's location or the location of the owner's home by address: typically the street address, town/city, state, and zip code. To prepare addresses, the *geocode* function from the *ggmap* (Kahle & Hadley, 2013) package was implemented to obtain latitude and longitude values for each address. Addresses were consistently formatted, and a new variable was created with zip codes to replace missing or incorrect zip code values. To iterate through each observation of the dataset and generate new location data, a loop was created. When an address was formatted differently (e.g., "Pennsylvania Ave. and Atlanta St.") we manually adjusted the address to the correct format.

Dates in the datasets were separated into four variables: year, month, day, and the corresponding day of the week.

Animals' ages were recorded using years, months, and weeks, and these variables were separated and recalculated to years using the *stringr* (Wickham H. , 2019) package.

The *impounds repeats.csv* data contain multiple levels for the animal type variable including birds, cats, dogs, livestock, and other. We restricted our analysis to only cats and dogs. This removed 5,901 observations from the data.

2.2 Missing Data

Ultimately, the impounds dataset was the only dataset used for modeling. Therefore, the techniques we used to deal with missing data and those related to variable transformation only pertain to the impounds data.

First, we considered imputation. After reducing the data to only contain cats and dogs, there were 127,310 observations. For the address (crossing) variable there are 1,765 missing values. For the animal age variable, there are 1,827 missing values. Removing all of these observations would result in a total loss of 3,602 observations, which is roughly 2.8% of all the data. This ultimately may not have been a significant amount of data, but we wanted to keep as much information as possible. Imputing values means to replace missing values with values that are "reasonable." A common technique to do this is replacing the missing values with the mean, or average value, if the variable is numeric. If the variable is a factor or categorical, the mode can replace missing values. These techniques are simple and easy to understand but can sometimes introduce bias and underestimate the error. Because only a small percentage of the data were missing, we used imputation. We imputed all the missing values using the *mice* (van Buuren & Karin, 2011) package. To ensure the imputation cover factor and categorical variables, we used the predictive mean matching (ppm) method. Also, to increase the reliability of the imputation, we imputed 10 times with 30 iterations.

2.3 Variable Transformations

The impounds data contain two types of variables: factor and numeric. Factor variables are categorical variables, such as color. However, numbers can be used as factors depending on what they represent. For example, if dog = 1 and cat = 2, then this will still be treated as a factor because of what it represents. A value such as 1.5 would be meaningless in this instance.

There are few numeric variables in this data. These variables take on values that are continuous. Age, for example, ranges from 0 to 30 in this data, and a value of 1.2 is meaningful since an animal can be 1.2 years old.

We centered and scaled all of our numeric variables.[1] We did this for two reasons: First, to avoid collinearity. Second, after fitting a model, we wanted our results to be interpretable. When the variables are on the same scale, we can better interpret results since they will have the same units.

2.4 Data Split

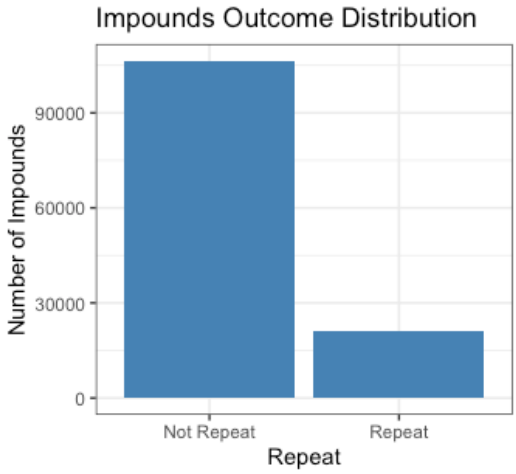
For modeling purposes, we split the data into training and testing sets. The training set contains 70% of the original data while the testing set contains the remaining 30%. When fitting our models, we optimized different parameters using the training set, and then we evaluated how well those models performed on the testing set. In order to ensure that equal proportions of repeat impounds were present in both the testing and training sets, we used the *createDataPartition* function from the *caret* (Kuhn, 2020) package. Both sets were roughly 16.6% repeat impounds.

2.5 Evaluation Metric

When determining how effective a model is, accuracy is a commonly used metric. Accuracy is calculated by dividing the total number of correct predictions, by the total number of predictions. Although intuitive, this metric on its own does not always capture how effective a model is.

Roughly 16.4% of the impounds data are classified as repeated impounds while the remaining 83.6% are classified as not repeated which creates an imbalance of data (*Figure 1*).

Figure 1



[1] https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

If we simply predicted every observation to not be a repeat impound, we could still achieve 83.6% accuracy. Although that may sound like a reasonable accuracy, this model would not offer any insight into how to better predict repeat impounds.

To account for the data imbalance, we incorporated some other metrics to evaluate the performance of our models including sensitivity, specificity, precision, and the F1 score. Sensitivity, also known as recall, will measure how well we are predicting repeat impounds. This value is determined by dividing the number of correctly predicted repeat impounds by the total number of actual repeat impounds. Similarly, specificity will measure how well we are predicting the not repeated observations. We expect this value to be higher since we have more data for the not repeated class. This is calculated by dividing the number of correctly predicting not repeat impounds by the total number of actual not repeat impounds.

Precision will measure the probability of a repeated impound prediction actually being a repeat impound. It is calculated by dividing the number of correctly predicted repeat impounds by the total number of predicted repeat impounds (both correctly and incorrectly predicted).

Lastly, we include the F1 score as another metric for our models. The exact calculation of this metric is slightly more complicated, but it is another way to measure accuracy using both sensitivity and precision.

By incorporating several different evaluation metrics, we can obtain more insight into how well each of our models perform.

2.6 Logistic Regression Model

The first model we implemented was the logistic regression model. It is one of the most commonly used models for a binary outcome. A binary outcome can take only two values. For our data, the binary outcome is "repeated impound" or "not repeated impound."

In logistic regression, we input different variables from the data and the output is a probability between 0 and 1. For our model, the "positive class" is the repeated impound. Therefore, the probability that is returned corresponds to the probability of being a repeat impound. The default cut off point is 0.5. This means if a probability is greater than 0.5, it will be labeled as a predicted repeated impound, otherwise, it will be labeled as a predicted non-repeated impound.

We implemented the *caret* package to fit the logistic regression model. We tuned the *alpha* and *lambda* parameters, allowing penalization in our model (elastic net). We used 10-fold cross validation with 5 repeats. This means our data are split into ten parts. Nine of these parts are used to train the data and the remaining part is used for evaluation. This process is repeated 5 times.

In the *train* function that we used for this model, we also specified the weights of each class. The weights for each class were calculated by dividing 0.5 by the total number of observations in that class. This puts more weight on the minority class, that is, the repeated impounds.

$$Weight_{repeat\ impounds} = \frac{0.5}{total\ number\ of\ repeat\ impounds}$$

2.7 Random Forest Model

The second model we used was a random forest. To better understand a random forest model, we will introduce the concept of the decision tree. In a decision tree, we will use one or more criteria to decide the outcome. For example, if someone tries to apply for a credit card, the credit card company

makes the decision based on criteria such as income, education, Fico score, etc. The process of making that decision is a decision tree, and each criterion is called a node. A random forest model is a model that implements the use of multiple decision trees.

We used the *randomForest* (Liaw & Wiener, 2002) package instead of the *caret* package for this model. Only two parameters can be tuned in this model: *mytry* and *tree*. The *mytry* parameter is the number of variables that are used to make a tree, and the *tree* parameter is the number of trees in the model. We only tuned the mytry parameter and set the tree to 500. Since the *mytry* parameter randomly selects the numbers of variables to build a tree, we did not use cross validation for this model.

2.8 Neural Network Model

The last model we implemented was a neural network. Neural networks work in a way very similar to the human brain. When someone tries to describe an object to us, we need to understand the features of the object in order to determine what it is. To determine what the object is, our brain must obtain information about the object and use that information in a series of calculations. In neural networks, the descriptive features of the object are called "input units" or "neurons", the series of calculations are called "hidden layers" and the object itself is the output unit. For our model, the input units are the different variables contained in the data, the hidden layers and units are the algorithm, and the output unit is whether the observation is a repeated impound.

In the neural network model, we set up two hyper-parameters: *size* and *decay*. The size determines the number of units in the hidden layers and the decay is used to avoid overfitting. We also applied the 10-fold cross validation with 5 repeats for this model. Similar to the logistic regression, we run an additional model with adjusted weights on the minority class to see how the result may differ.

2.9 Optimal Cut Point

As previously mentioned in the logistic model section, the default cut off point to classify an observation is 0.5. This threshold, however, does not always handle imbalance in the data well. We were concerned about imbalance in our data so we implemented the *optimal.cutpoints* function from the *OptimalCutpoints* (Lopez-Raton, Rodriguez-Alvarez, Cadarso Suarez, & Gude Sampedro, 2014) package to adjust this threshold on the training set.

Unfortunately, due to the large dataset and limitations of the package, we could not implement this function using the full dataset for the logistic regression and neural network models. We only used 40,000 rows to determine the optimal cut point.

3. Results

3.1. Exploratory Plots

With access to three different data sets, we wanted to merge all the data and include bites and violation information in our models predicting repeat impounds. However, we would have had to reduce our impounds data from 127,310 observations to, at most, 5,631 observations. There also was not a single variable on which we could easily merge the three data sets. Therefore, we proceeded with the impounds data as our primary data for analysis and examined some exploratory plots for the bites and violations data. These plots will be attached in a separate document for the client.

We also generated some exploratory plots for the impounds data. *Figure 2* shows the total number of impounds between November of 2016 and September of 2020. The different colors distinguish between cats and dogs. *Figure 3* provides the same information for repeat impounds only.

Figure 2

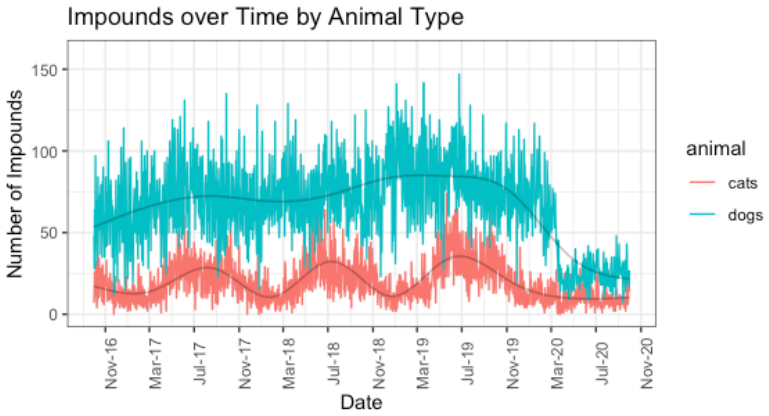


Figure 3

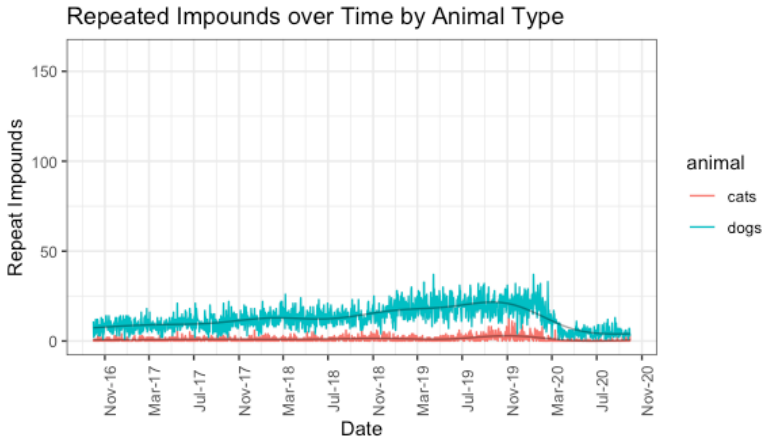


Figure 2 and *Figure 3* show some interesting differences in cyclical patterns over time. According to the client, warmer summer months are often peak season for cat impounds because cats breed in warmer temperatures. In *Figure 2*, the pattern of increase in cat impounds appears especially in the months of July and August. Dog impounds, however, do not appear to follow any sort of cyclical pattern. We also see from these plots that total impounds drop drastically in March of 2020 when the coronavirus pandemic was declared a national emergency. However, in *Figure 3*, we do not see the same behavior with repeat impounds. In *Figure 4* below, we have reduced the limits of the y-axis to see repeat impounds more clearly.

With repeat impounds, there does not seem to be any evidence of the same cyclical pattern seen in cats from total impounds. Repeat dog impounds seem to be steadily increasing until August of 2019. Although there might have been periods of increasing total dog impounds, it does not appear to be as drastic as the increase seen in repeat impounds.

Figure 4

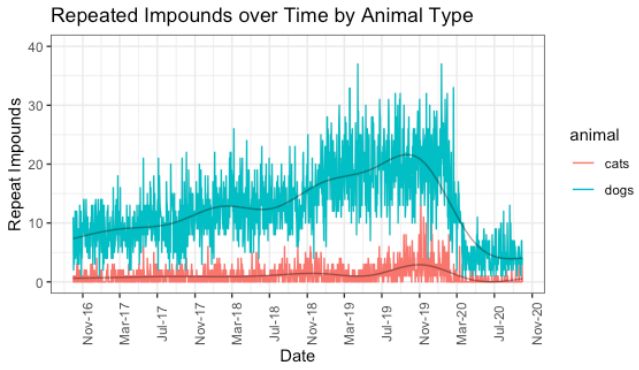


Figure 5 and *Figure 6* show total and repeat impounds for each year. Each year is also separated by impound type. Impound types include the following: confiscate, disposal required, keep safe, owner surrender, stray, and transfer.

Figure 5

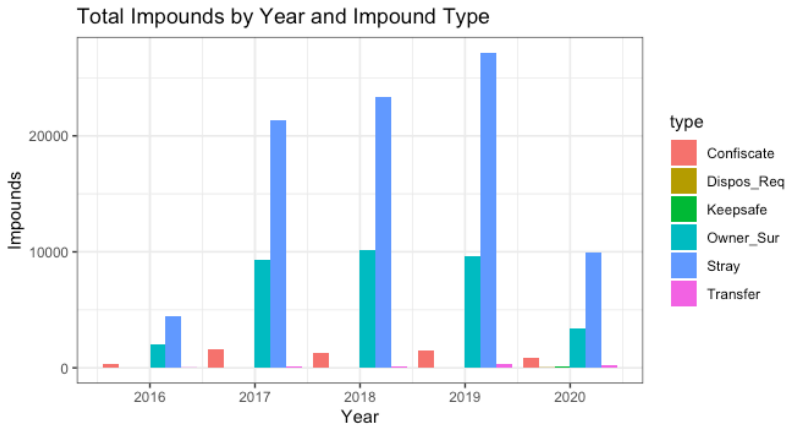


Figure 6

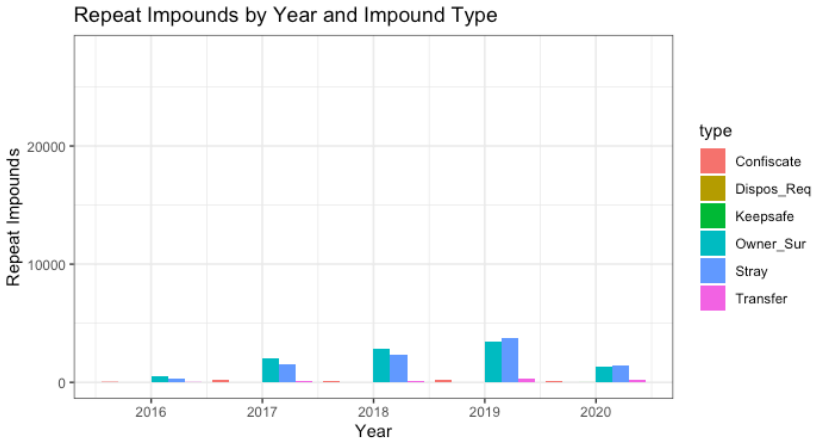


Figure 5 and *Figure 6* offer insight into differences between impound types between total impounds and repeat impounds. For total impounds, nearly all of the impound types are stray and owner surrender. Stray impounds more than double those of owner surrender. However, with repeat impounds,

as shown in *Figure 6*, owner surrender impounds are similar to stray impounds, if not greater. For 2016 to 2018, repeat owner surrender impounds are greater than those of stray impounds. Then in 2019 and 2020 repeat stray impounds are largest. Differences in repeat impounds for strays and own surrenders are not as drastically different as they are for total impounds.

In *Figure 7* and *Figure 8*, total and repeat impounds are separated by the sex of the animals and the two different colors distinguish between dogs and cats. Neutered and spayed refer to animals that have had their reproductive organs removed whereas female and male animals still have their reproductive organs.

Figure 7

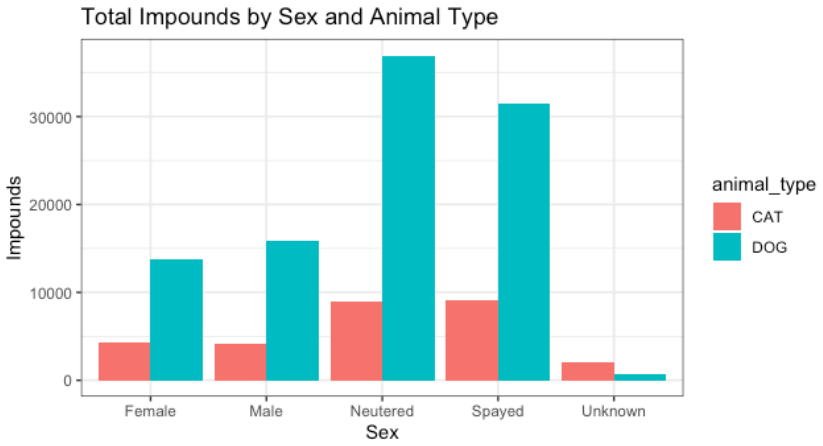


Figure 8

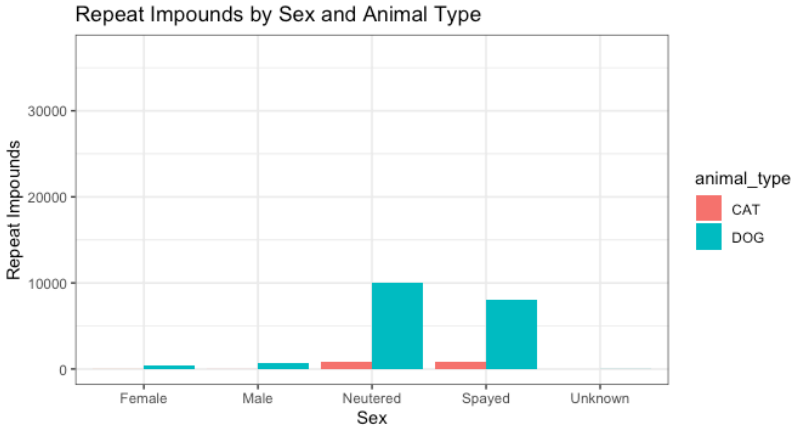


Figure 7 and *Figure 8* highlight differences between sex for total and repeat impounds. For total impounds, *Figure 7* shows that spayed and neutered animals are more often impounded than animals not spayed or neutered. Animals that have not been spayed or neutered, however, still appear to make up a sizable portion of the impounds. This difference between sexes seems to be even more extreme with repeat impounds, as shown in *Figure 8*. We first see that most repeat impounds are dogs, but we also see that a vast majority of the repeat impounds are spayed and neutered. This may not be surprising to our client; however, it was counterintuitive to what we expected concerning the previous care of animals that have been repeatedly impounded.

3.2 Logistic Regression

We first implemented the logistic regression model tuning the parameters *alpha* and *lambda* while also implementing repeated 10-fold cross-validation. The test set contained 38,192 observations, with 6,265 predicted to be repeated impounds and 31,927 predicted to not be repeat impounds. *Table 2* shows predictions compared to the actual labels. The model correctly classifies 751 repeated impounds (out of 6,265) resulting in a sensitivity of 12.0%. On the other hand, the model correctly classifies 31,442 (out of 31927) not repeated impounds which result a specificity 98.5%. Overall, the model achieved an accuracy of 84.3%. We then incorporated adjusted weights and an optimal cut point to improve the model.

Table 2

Logistic Regression Classification Results

		Predicted	
		Repeat	**Not Repeat**
Actual	**Repeat**	751	485
	Not Repeat	5514	31442

As a result, the final logistic regression model did not incorporate weighted classes, it included an optimal cut point of roughly 0.28, and the *alpha* and *lambda* parameters were tuned to 0 and 0.01 respectively. As seen below in *Table 3*, the accuracies and sensitivities are similar for the last three models, but our final model without weighted classes achieved the highest F1 score.

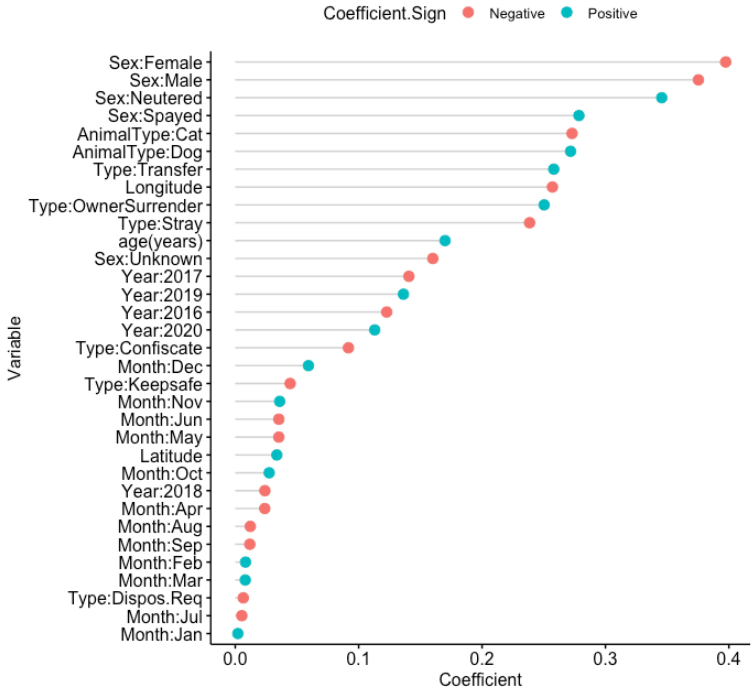
Table 3

Logistic Regression Model Results

Model	**Cut Point**	**Accuracy**	**Sensitivity**	**Specificity**	**F1 Score**
Logistic	0.5	0.843	0.120	0.985	0.200
Logistic with adjusted weights	0.5	0.762	0.531	0.807	0.422
Logistic with optimal cut point	0.279	0.813	0.489	0.877	0.462
Logistic with adjusted weights and optimal cut point	0.562	0.791	0.476	0.852	0.427

Additionally, we generated a variable importance plot for the final model. In *Figure 9*, the variable importance plot displays the absolute values of the coefficients. We added colors to represent which variable coefficients are positive and negative. For factor variables, such as animal type and sex, negative coefficients imply that these levels (or characteristics) reduce the probability of being a repeat impound, while positive coefficients imply an increase in the probability. For numeric variables, such as age and longitude, a negative coefficient implies that as the value of the given variable increases, the probability of being a repeat impound decreases. (Similarly, a positive coefficient implies that as the variable increases, the probability of being a repeat impound also increases.)

Figure 9
Variable Importance, Logistic Regression



3.3 Random Forest

For the random forest, we did not tune the class weights due to limitations of the *randomForest* package. The optimal value for the *mytry* parameter was 4, resulting in relatively high accuracy and sensitivity but a low value for sensitivity. As seen in *Table 4*, the model correctly classifies 1,390 of 6265 repeated impounds resulting in a sensitivity of 22.2%. On the other hand, the model correctly classifies 31,705 of 31,927 not repeated impounds resulting in a specificity 99.3%. Overall, the model achieved an accuracy of 86.7%.

Table 4
Random Forest Classification Results

		Predicted	
		Repeat	**Not Repeat**
Actual	**Repeat**	1390	222
	Not Repeat	4875	31705

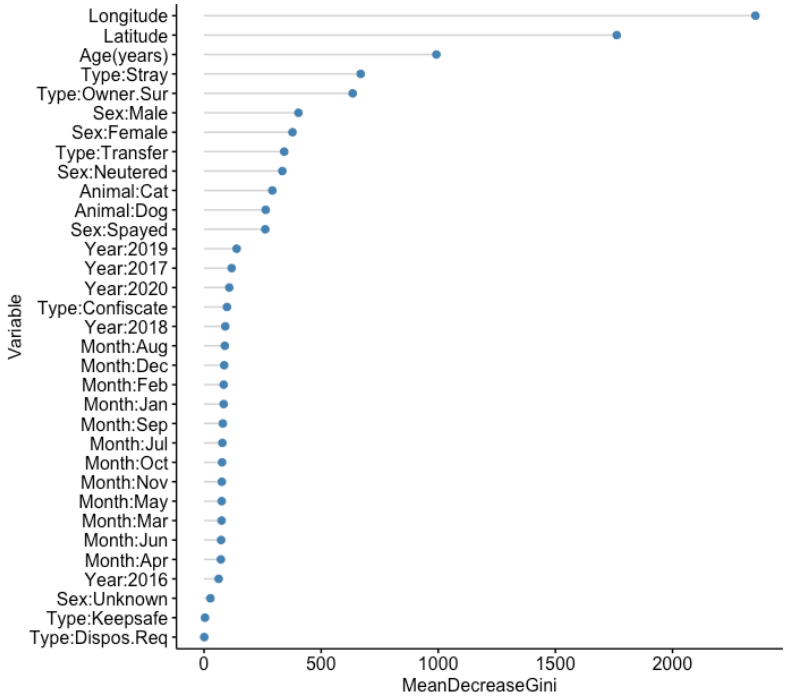
In efforts to improve the specificity, we adjusted the optimal cut point. At a small cost to the accuracy and sensitivity, using the optimal cut point boosted both the specificity and F1 score. The exact values are given in *Table 5* below.

Table 5
Random Forest Results

Model	Cut Point	Accuracy	Sensitivity	Specificity	F1 Score
Random Forest	0.5	0.867	0.222	0.993	0.353
Random Forest with Optimal Cut Point	0.127	0.840	0.501	0.906	0.506

Shown in *Figure 10*, we generated a variable importance plot using the random forest model. Instead of coefficient sizes, this plot displays the mean decrease in Gini. The Gini is special metric for decision trees; a higher mean decrease in Gini indicates higher variable importance.

Figure 10
Variable Importance, Random Forest



3.4 Neural Network

With the neural network models, the *size* and *decay* parameters were tuned to 9 and 1, respectively. Given in *Table 6*, the model correctly classifies 1,574 of 6,265 repeated impounds which resulting in a sensitivity of 26.9%. On the other hand, the model correctly classifies 31,213 of 31,927 not repeated impounds resulting in a specificity 97.8%. Overall, the model achieved an accuracy of 86.2%.

Table 6
Neural Network Classification Results

		Predicted	
		Repeat	**Not Repeat**
Actual	**Repeat**	1574	714
	Not Repeat	4681	31213

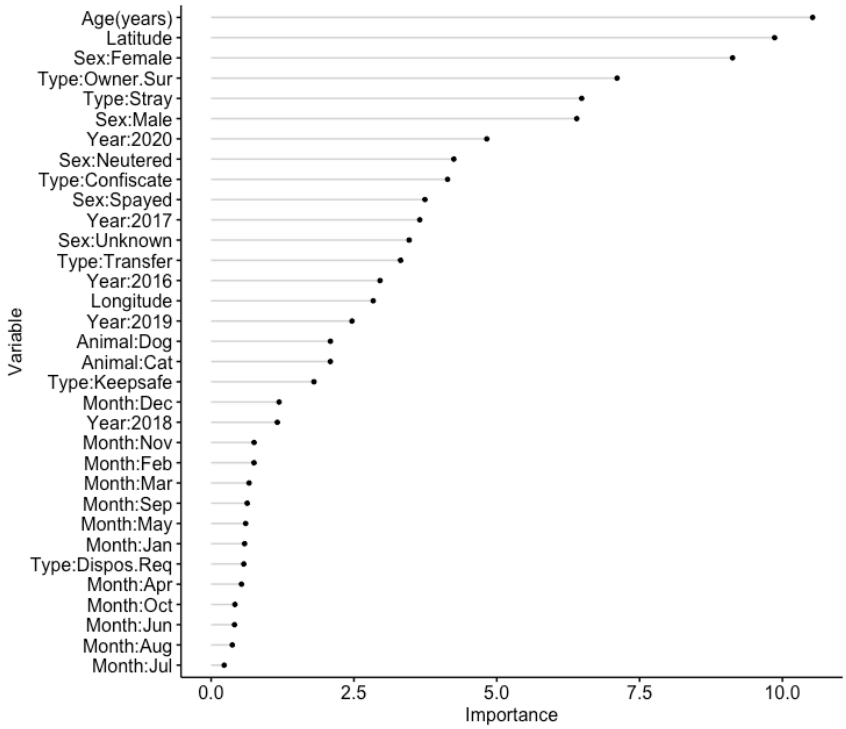
Then, we incorporated adjusted class weights and the optimal cut point, as we did with the logistic regression model, to achieve the best results. *Table 7* shows that the model including an optimal cut point and the model including both a cut point and adjusted weights performed similarly by all the evaluation metrics. Because the model that incorporated the optimal cut point without adjusted weights slightly outperformed the other model, we chose this as our final neural network model.

Table 7
Neural Network Results

Model	Cut Point	Accuracy	Sensitivity	Specificity	F1 Score
Neural Network	0.5	0.862	0.269	0.978	0.389
Neural Network with Adjusted Weights	0.5	0.726	0.785	0.714	0.484
Neural Network with Optimal Cut Point	0.292	0.829	0.885	0.544	0.896
Neural Network with Adjusted Weights and Optimal Cut Point	0.680	0.816	0.868	0.548	0.887

The variable importance plot, given by *Figure 11*, uses a combination of the absolute values of the weights to evaluate the importance. A higher value indicates higher variable importance.

Figure 11
Variable Importance, Neural Network



4. Discussion

4.1 Model Comparison

Between the three final models, the neural network model using an optimal cut point achieved the highest sensitivity and F1 score while the random forest model achieved the highest accuracy and specificity. The logistic regression model did not outperform the other models in any of the four metrics. The specificity of the neural network model was only 54.4%, which is far lower than the other models. This result means that the model is able to predict non-repeat impounds only as well as it can predict repeat impounds. Therefore, we do not think the neural network is the ideal model for this data.

Table 8
Final Model Comparison

Model	Accuracy	Sensitivity	Specificity	F1 Score
Logistic Regression with Optimal Cut Point	0.813	0.489	0.877	0.462
Random Forest with Optimal Cut Point	0.840	0.501	0.906	0.506
Neural Network with Optimal Cut Point	0.829	0.885	0.544	0.896

One of the competitive advantages of the logistic regression model is its interpretability. Since the model provides coefficients for each variable, we know how each variable contributes to the resulting probability of being a repeat impound. For example, the coefficient for the gender variable is -0.397 which means that being a female animal multiplies the odds of being a repeat impound by $e^{-0.397} = 0.672$ (i.e. it decreases the odds by 32.8%). We cannot obtain similar information from the random forest or neural network models. The mean decrease in Gini and importance metrics do not tell us about the size or direction of the change related to the outcome variable.

Although the random forest and neural network models do not provide interpretable coefficients, there is still some valuable insight within the importance plots. *Table 9* shows that the sex of the animal appears as an important variable in all three models. This reinforces some exploratory findings given by *Figure 7* and *Figure 8*: That is, sex is critical in predicting whether an animal is a repeat impound or not.

Table 9
Variable Importance, Top 6

Logistic Regression	Random Forest	Neural Network
Sex:Female	Longitude	Age(years)
Sex:Male	Latitude	Latitude
Sex:Neutered	Age(years)	Sex.Female
Sex:Spayed	Type:Stray	Type:Owner.Sur
Animal:Cat	Type:Owner.Sur	Type:Stray
Animal:Dog	Sex:Male	Sex:Male

An interesting result of these models is that location appears to be an important variable for both the random forest and neural network models, but it is not one of top important variables for the logistic regression model. This may suggest that the relationship between location and repeat impounds is not linear or the relationship involves some interactions between location and other variables.

Because Dallas Animal Services is primarily concerned about identifying trends related to repeat impounds and understanding how different variables affect these repeat impounds, we recommend the logistic regression model using an optimal cut point as a predictive model. This model provides DAS with more interpretable results while still maintaining the ability to make good predictions.

4.2 R Shiny App

In efforts to provide some more insight into the location of the data, we created R shiny apps that use static Google maps to display exact geographic coordinates. Three separate apps were created to display impounds, bites, and violations. Separate apps were created because we plotted large amounts of data and we wanted to reduce the loading time in addition to the clutter of variable inputs. *Table 10* lists which variables can be adjusted for each app.

Table 10
R Shiny Adjustable Variables

Impounds	Bites	Violations
Total or Repeat	Year	Year
Year	Month	Month
Month	Animal Type	Violation Type

Impound Type	Neutered/Spayed	Zip Code
Animal Type	Type of Break	Visibility
Neutered/Spayed	Severity of Bite	
Zip Code	Animal at Large	
Visibility	Zip Code	
	Visibility	

Once the input variables have been adjusted, the app generates two plots using the longitude and latitude of each observation. The first plot displays the raw data. The visibility is automatically set to be slightly transparent, so if multiple impounds occur at a specific point, that point will be darker on the map. The second plot is a density plot (heat map). Instead of displaying specific locations, this plot shows areas that have higher densities of impounds or repeat impounds.

Figure 12 shows the density plot for all impounds, while *Figure 13* shows the density plot for only the repeated impounds. The high-density area west of downtown Dallas is where Dallas Animal Services is located. We expect that this address is used often when animals are surrendered at the animal shelter or when the address is not known. We see that higher density areas for total impounds and repeat impounds are in the south and south western parts of Dallas. We also see a large higher-density area in the south eastern part of Dallas, east of the Trinity River. Repeat impounds do not appear to extend as far north as the total impounds, however, there is much less data for repeat impounds. Even though there do appear to be some higher-density areas, the data still seems to be quite dispersed all around Dallas.

Figure 12
Density Plot, All Impounds

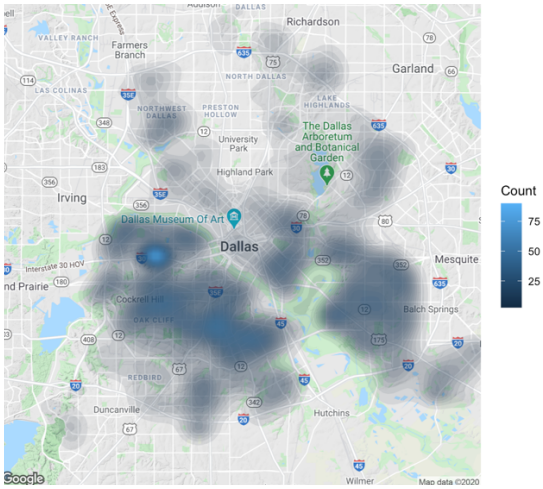


Figure 13
Density Plot, Repeat Impounds

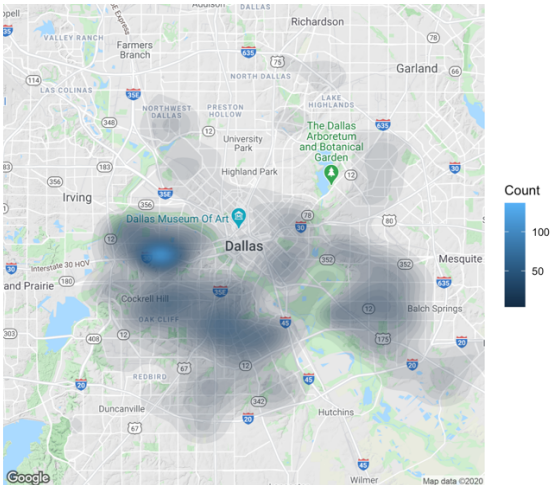


Figure 14 and *Figure 15* show density plots of stray impounds for both total impounds and repeat impounds respectively. For total impounds, the stray impounds density plot looks almost identical to that of all types of impounds (*Figure 12*). For repeat stray impounds, the concentration in south Dallas seems to be much higher than that of other surrounding areas.

Figure 14
Density Plot, All Stray Impounds

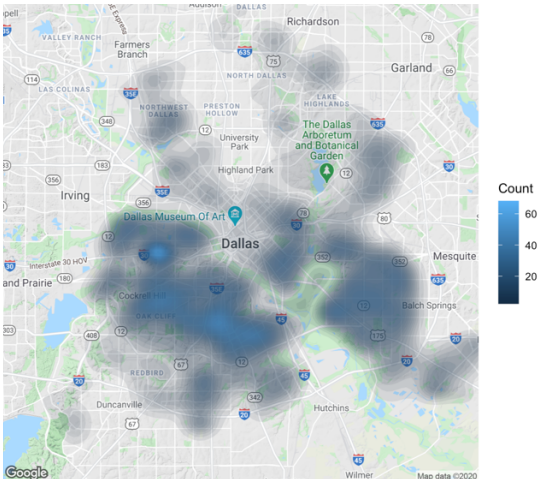
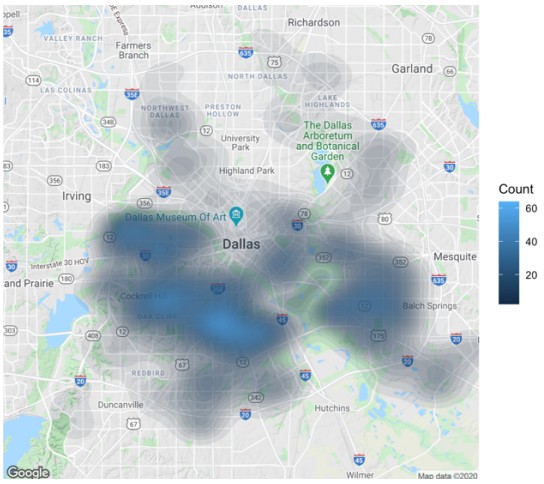


Figure 15
Density Plot, Repeat Stray Impounds



The two following figures show impounds for confiscated animals. *Figure 16* displays the density plot for all impounds, while *Figure 17* displays only repeat impounds. These plots are interesting in that they show another high-density area around downtown Dallas. The downtown area for total impounds seems to have a higher density of confiscated animals than the downtown area for repeat impounds. However, the repeat impounds have a very distinguishable high-density area in south and south-western Dallas.

Figure 16
Density Plot, Total Impounds: Confiscate

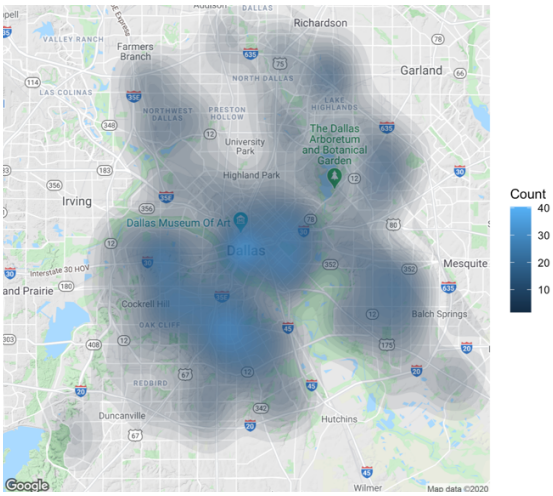
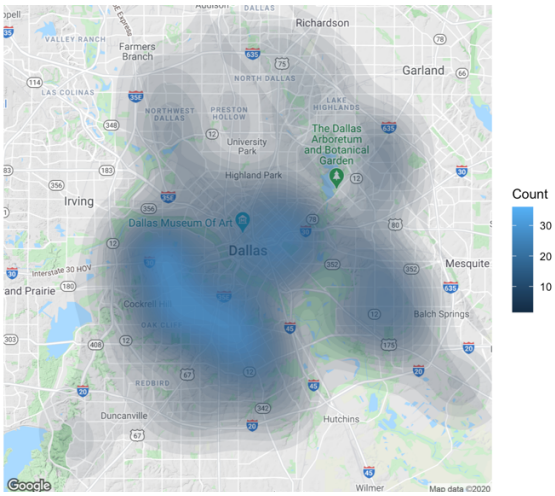


Figure 17
Density Plot, Repeat Impounds: Confiscate



Lastly, *Figure 18* and *Figure 19* show raw data plots specifically for the south western part of Dallas. When examining this kind of plot, it might be more difficult to understand trends, however, it allows the user to identify locations more precisely. *Figure 18* displays all impounds while *Figure 19* displays only repeat impounds. In both figures, we can observe high density areas with more precision. For example, just north of I-30, we see that the neighborhoods behind Singleton Blvd. have higher concentrations of impounds for both total impounds and repeat impounds (especially on the western end).

Figure 18
Total Impounds, South Dallas

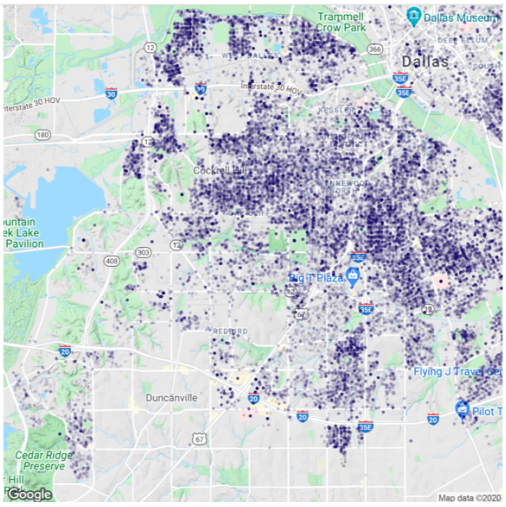
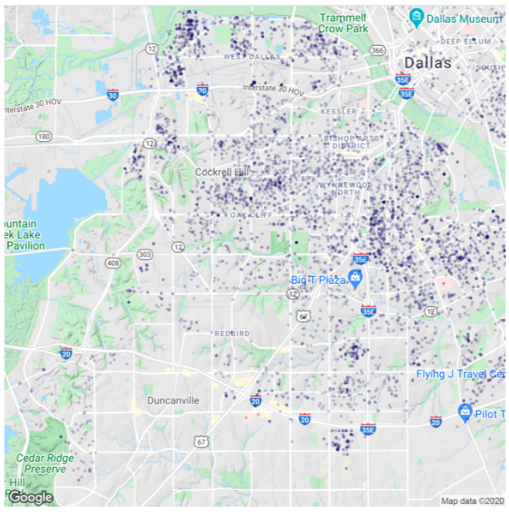


Figure 19
Repeat Impounds, South Dallas



4.3 Conclusion

Using predictive models and generating maps, we were able to identify some trends regarding repeat impounds. The logistic regression model shows that the sex and type of an animal contribute largely to the probability of being repeatedly impounded. Dogs are more likely to be repeatedly impounded as well as animals that have been spayed or neutered. Even if these trends were already known or expected, other models suggest that location may also be an important factor in predicting repeat impounds. Our models suggest that northern areas of Dallas are less likely to be associated with repeat impounds and this can be visually confirmed using the R Shiny app. We hope that these insights can better help DAS to understand the data they have collected and inform decisions as they determine how to best allocate their time and resources to reduce the number of animals that are repeatedly impounded.

Works Cited

Dallas Animal Services. (2020, 1 21). *Volunteer Handbook and Guidelines.* Retrieved from Dallas Animal Services: https://dallascityhall.com/departments/dallas-animal-services/DCH%20Documents/DAS-PRO-114.pdf

Kahle, D., & Hadley, W. (2013). ggmap: Spatial Visualization with ggplot2. *The R Journal, 5*, 144-161.

Kuhn, M. (2020). *The Comprehensive R Archive Network*. Retrieved from caret: Classification and Regression Training: https://CRAN.R-project.org/package=caret

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*, 18-22.

Lopez-Raton, M., Rodriguez-Alvarez, M., Cadarso Suarez, C., & Gude Sampedro, F. (2014). {OptimalCutpoints}: An {R} Package for Selecting Optimal Cutpoints in Diagnostic Tests. *Journal of Statistical Software*, 1-36.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

van Buuren, S., & Karin, G.-O. (2011). {mice}: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*, 1-67.

Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. Retrieved from A Comprehensive R Archive Network: https://CRAN.R-project.org/package=stringr