Your final report should have the following components and follow the APA style as much as possible.

* **Research problem (10pts)**: Describe the task you want to achieve. What is the outcome of interest? What are you trying to predict? Why is it important? What are the potential benefits of having a predictive model for this outcome? Discuss potential applications of such a model.

As people stayed home for most of the early days of the coronavirus pandemic, the demand for many goods and services increased. Some of the increased demand was not surprising and likely expected; items such as home workout equipment, trampolines, and lumber. However, some of the demand took the industry by surprise: yeast for baking, and pets. In fact, the interest in pet adoption increased so much that shelters were regularly reporting empty kennels and sifting through dozens of adoption applications for a single puppy.

Unfortunately, as vaccines were rolled out and people began returning to work and school, shelters and foster groups filled up with animals that were no longer compatible with people’s lifestyles. The decrease in demand for dogs means that people can be more selective in the kind of dog they adopt. However, the information provided by animal shelters and rescue groups are often based on a short period of time with the animal and the animal’s appearance. One shelter’s Border collie mix might be another’s spaniel or shepherd mix. This best guess breed identification can have significant impacts on a dog’s future and could be the difference between adoption and euthanasia. Objectively identifying the impact of a dog’s listed breed on the length of stay in an animal shelter could provide shelters with the information needed to shift away from listing a dog’s breed as the primary information for potential adopters and toward a more holistic evaluation of a dog’s temperament and future needs.

* **Description of the data (15pts):** Describe core features of the data, any additional features you produced from existing features and how, basic descriptive statistics about these features, and any missing data analysis you conduct. The description should be sufficiently clear that the instructor understands all the variables included in your modeling.

The Austin Animal Center in Texas is the largest no-kill shelter in the US. The shelter maintains data on the intake and outcomes of animals beginning from October 2013 to present. This data set was obtained from kaggle: <https://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-intakes-and-outcomes?select=aac_intakes_outcomes.csv>. The data include information about the intake and outcome of the animal, and details on the type and condition of the animal. A brief examination of the data revealed that the animal shelter takes in animals in addition to typical domestic pets (cats and dogs). For the purposes of this study, the following types of animals were excluded: cats, birds, and animals that were classified as other, including rabbits, bats, snakes, raccoons, ferrets, reptiles, and other wild animals that live in close proximity to humans. Additionally, dog breeds with sample sizes less than 20 were excluded as this small sample made it difficult to accurately model the length of stay for the breed. The final data set included variables of the animal (breed, age on intake, sex, condition of the animal), circumstances of the animal arriving at the shelter (type of intake, month of intake), and specifics of the outcome of the animal (outcome, month of outcome, time spent in the shelter, measured in days).

Initial data visualization was performed to understand the data available. This included visualizing the number of animals taken into the shelter each month and further exploring the number of each type of animal taken in each month.

\*\*Figure 1.\*\* \_Shelter Animals Taken in Each Month\_

\*\*Table 1.\*\* \_Type of Shelter Animal Taken in Each Month\_

It is also of interest to explore the relationship between the average length of stay in the shelter of each breed with the frequency of that breed present in the shelter. Table 2 provides the average length of stay in the shelter for the 25 most frequently taken in dog breeds.

\*\*Table 2.\*\* \_Average Length of Stay (in days) for Most Common Dog Breeds in Shelter\_

* **Description of the models (15pts):** List at least three different modeling approaches you apply to this dataset. Describe each model, why the given model was selected, which hyperparameters to be optimized and how. Also, discuss how you plan to evaluate model performance.

In order to predict the length of stay of shelter dogs, three types of modeling approaches were explored: linear regression, linear regression with ridge penalty, and bagged trees. These models were chosen for their increasing complexity to determine the extent to which the increasing complexity added value to or impacted the predictions and importance of variables used in the predictions. All models were fit with 10-fold cross validation for comparison purposes. The performances of the models will be compared using the values of R-squared, MAE, and RMSE.

* **Model fit (20pts):** Provide the results of your model evaluation. Compare and contrasts results from different fits, including a discussion of model performance. Discuss your final model selection and the evidence that led you to this selection. If it is a classification problem, how did you choose a cut-off point for binary predictions? Did you consider different cut-off points?

The linear regression without regularization produced an RMSE of 37.66, an MAE of 16.77 and an r-squared of .11. Even though I used 10-fold cross-validation while training the dataset, the r-squared of .11 on the testing data was much smaller than the r-squared on the training data of .65, suggesting the model is overfitted to the training data. This drop in performance means the linear regression model without regularization is not a realistic method for predicting the length of stay of shelter dogs. Figure 2 illustrates the underperformance of this model.

\*\*Figure 2.\*\* \_Linear Regression without Regularization Model Performance\_

The linear regression with ridge penalty produced an RMSE of 37.66, MAE of 16.57, and r-squared of .11. These results are very similar to those of the linear regression without regularization. The model also revealed the top ten predictors of length of stay for shelter dogs, which included eight specific breeds (Table 2).

\*\*Table 2.\*\* \_Top Ten Predictors of Shelter Length of Stay from Linear Regression with Ridge Penalty Model\_

The bagged tree model had a similar RMSE as the linear regression without regularization and the linear regression with ridge penalty (34.45), a lower MAE of 12.36, and a higher r-squared at .26. Due to the similar RMSE and MAE of all three models, I would choose the bagged tree model due to the higher r-squared value. The three models are compared in Table 3.

\*\*Table 3.\*\* \_Model Comparison\_

* **Data visualization (5pts):** Include at least two plots (or more) to help communicate your findings. The plots may be of initial data explorations, fits of individual models, and plots displaying the performance of competing models.
* **Discussion/Conclusion (25pts):** Discuss and summarize what you learned. Which variables were the most important in predicting your outcome? Was this expected or surprising? Were different models close in performance, or were there significant gaps in performance from different modeling approaches? Are there practical/applied findings that could help the field of your interest based on your work? If yes, what are they?

Discussion

As Table 2 illustrated, the breed of the dog is highly predictive of the amount of time a dog spends in an animal shelter. Although there are a number of large breed dogs, there is not an obvious category of dog that is more predictive of the length of stay. For example, there are representatives of a variety of AKC dog groups: the non-sporting group (Bulldog, Tibetan Spaniel), the herding group (Collie), hound (Coonhound); sizes: the Terrier and the Tibetan Spaniel being small dogs, and the Retriever and Bulldog being larger dogs, and temperament. This suggests there may be some prejudice or unfamiliarity of some breeds.

The three models performed similarly with respect to RMSE. This was not surprising based on the examples we have worked with in class that have performed similarly as well. The difference between the r-squared values of the random forest model to the linear regression models was surprising, even though all values were low.

Conclusion

This wide variety of dog breeds in the top ten category is surprising. I would have thoughts that the larger dogs and bully-type dogs would occupy the top spots. I was not, however, surprised that breed was a top predictor of the length of stay in the shelter. The breed of the dog is often the best predictor of temperament of a dog, however if the breed listed is only based on a dog’s appearance and not genetics or the dog’s history, then breed is less powerful of a predictor. Shelters may find they have better success matching shelter dogs with their forever families if they provide a more thorough behavioral evaluation of the dog and assessment of the dog’s future needs to potential adopters.

* **Reproducibility (10pts):** Provide a link to the GitHub repo at the beginning of your report as a note.

The final report should be no longer than 10,000 words (approximately 15 single-spaced pages), excluding the tables, figures, and references.