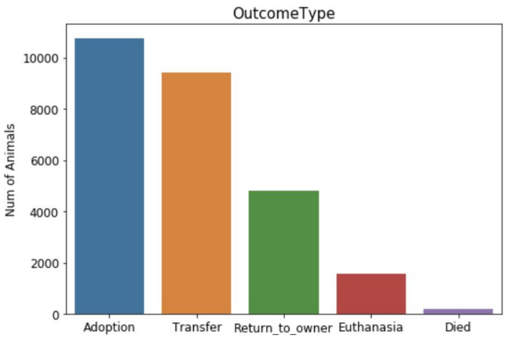
Shelter Animal Predictions

W207 Final Project

Danielle Adler, Alla Hale, John Pette

Project Aim

- Purpose of this study: **“What is the outcome for a shelter animal based on breed, color, sex, and age?”**
- Key input variables include name, date of outcome, animal type, sex upon outcome (fixed vs. intact and gender), and age upon outcome
    - Dataset contained 26,729 animal observations; all variables besides name were present for all observations and no clear outliers were present
- Outcome classes include adoption, dying, euthanasia, transfer, and return to owner; euthanasia and died each had very few observations making the dataset unbalanced



Background

- Learned that **random forests** win over individual decision trees in situations with a large number of features and sparsity among the features (Breiman, Leo)
- Learned about **data imputation** and the merits of oversampling from underrepresented classes or undersampling from overrepresented classes (Galar, M., A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera)
- Learned about **classifiers for sparse datasets** such as gradient-boosted decision trees
  - Gradient boosting constructs new decision tree models that predict the residuals of prior models and then adds them together to make final predictions (Chen, Tianqi and Guestrin, Carlos)
- Learned about shelter animals in that condition of intake and reason for intake informed outcome most with regards to euthanasia, adoption, etc.

Methodology

- **Exploratory Data Analysis:**
  - Dogs account for 60% of the dataset; even male/female split; 75% of animals are fixed
  - Breed is dominated by Domestic Shorthair cats, while color is dominated by Black/White
  - Older animals are more likely to be returned to their owner or euthanized, while younger animals are more likely to be transfered or adopted
  - The adoption and transfer outcomes are highest in the summertime

- **Data Pre-Processing**
  - Removed secondary breeds (and the word mix) as well as secondary colors
  - Created a continuous age variable so that all elements were relative to one another
  - Binarized all discrete variables, which created a matrix of 94% sparsity
  - Initialized columns with zero for train columns not present in the test dataset
  - Changed the order of the test columns to match train for prediction purposes
  - Before modeling, ended with 283 features and 26,729 observations

Methodology

- **Classifiers Evaluated**
  - Dummy classifier (only predicting majority class), Logistic Regression, Decision Trees, Random Forest, Multinomial Naive Bayes, and Gradient Boosted Decision Trees (best for sparse datasets)
  - Tried all algorithms using the StratifiedKFolds function within sci-kit learn, which splits training data into train vs. test with the distributions of the outcome classes

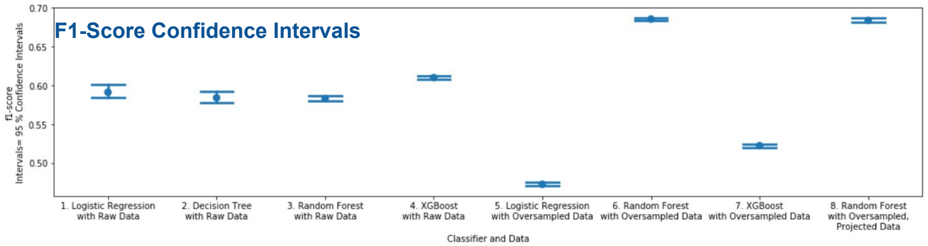
- **Class Balance Correction**
  - Tried two techniques (RandomSampler and SMOTE) to balance the outcome classes and oversample from smaller classes
  - RandomSampler worked more quickly and improved the random forest classifier by 6 percentage points; moved forward with this classifier

- **Principal Component Analysis**
  - Performed PCA for 70 components; 30 components explained ~90% of the variation
  - PCA did not change the f1-score for the oversampled, random forest classifier, but helped wi processing speed

Results

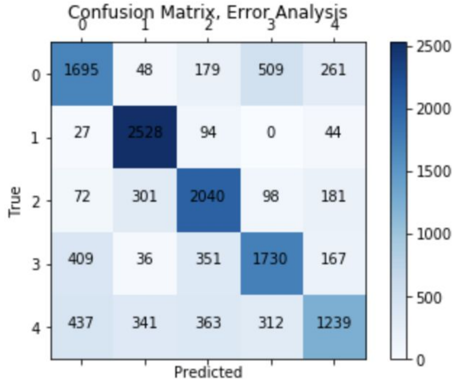
- Chose to focus on **weighted f1-score over accuracy**, as f1-score takes precision and recall into account
- Weighted f1-scores are best in the random forest, oversampled models
- Model **did not generalize well** (training folds f1-scores were higher than testing folds), this improved with the oversampled data
- Log loss is best with XGBoost, but we **must balance that with f1-score, log loss, and and speed of implementation**

	Accuracy	F1 Score	Log Loss
Decision Tree with Raw Data	0.59	0.58	3.43
Dumb	0.40	0.23	20.62
Logistic Regression with Oversampled Data	0.50	0.47	
Logistic Regression with Raw Data	0.63	0.59	1.24
Multinomial Naive Bayes with Oversampled Data	0.39	0.35	
Multinomial Naive Bayes with Raw Data	0.52	0.50	
Random Forest with Oversampled Data	0.69	0.69	
Random Forest with Oversampled, Projected Data	0.69	0.68	
Random Forest with Raw Data	0.60	0.58	1.45
Random Forest with Undersampled Data	0.40	0.40	
XGBoost with Oversampled Data	0.53	0.52	
XGBoost with Raw Data	0.63	0.61	1.10



Results

- **Winning classifier:** Random forest, oversampled, projected data, max_depth=80, min_samples_split=6
- Conducted extensive error analysis to figure out what classes were predicting the wrong label
- Top Issues with class predictions were: **return to owner predicted when the true value is adoption**, followed by adoption being predicted when the true label is return to owner or transfer
    - In all of cases, animals are leaving the shelter, which cause the confusion
- Top **feature issues involved breeds**:
    - Very common breeds, such as domestic shorthairs, were mispredicted as “transfer” or “died”, when the actual label was frequently euthanasia
    - As the domestic shorthair breed appears for both cats and dogs, we evaluated an interaction variable of domestic shorthair * animal type but it did not help



Label Legend
0: Adoption; 1: Died; 2: Euthanasia; 3: Return to owner; 4: Transfer

Conclusion

- Is there a better question than Kaggle's? What if we only had two outcomes?
  - Created another random forest model, oversampled model with PCA with just **two outcomes (positive: adoption, return to owner, transfer) or negative (euthanasia, or died)** to compare
  - With only two outcomes, we generalize better, overfitting is negligible, and **the f1-score is ~0.78 with a log loss of 1.18**
  - As we only had two outcomes, logistic regression performed very well with a marginally better f1-score of ~0.79 and a log loss of 0.49

- What can we do from our analysis?
  - Give shelters clearer understanding of animal breeds and colors that lead to negative outcomes
  - Help shelters understand specific seasons and ages for which certain outcomes are more common or certain animal types are more susceptible to certain outcomes

Questions?