Shelter Animal Outcomes

W207 Final Project

Sean, John, Subhashini, Talieh

PROBLEM DESCRIPTION

Goal: Help improve outcomes for shelter animals

- Using a dataset of information from Austin Animal Center, Kagglers have been asked to predict the outcome for each animal.
- ~26,000 training samples and ~11,000 test samples.
- Supervised classification problem.
- Plenty of scope for feature engineering.

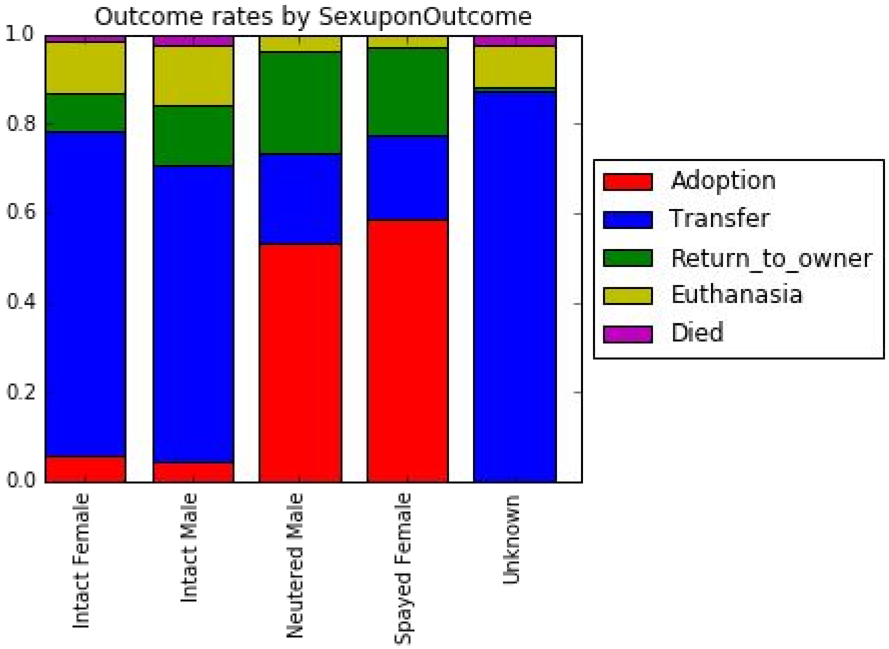
Feature set

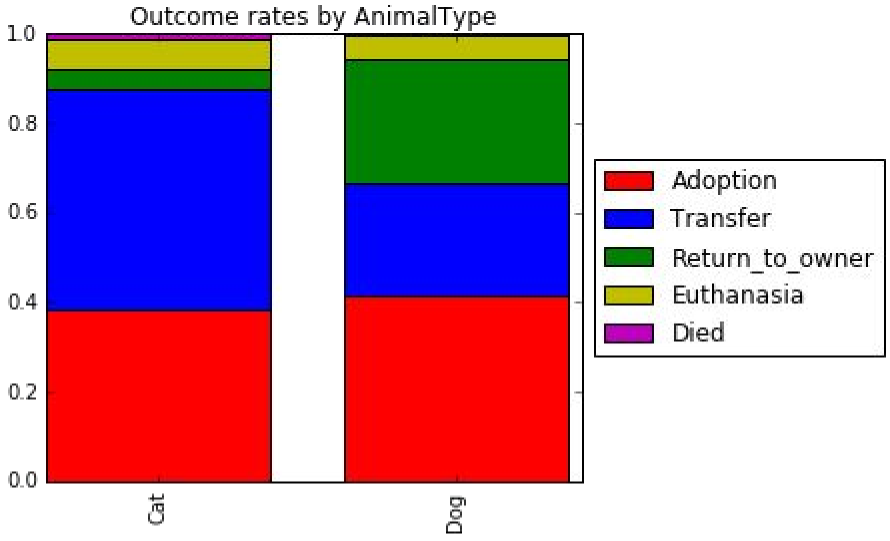
- **Name, DateTime of Outcome, AgeUponOutcome**
- **AnimalType:** Dog or Cat
- **SexuponOutcome:** Male/Female + neutered/spayed
- **Breed:** E.g "Dachshund/Beagle", "Domestic Shorthair Mix"
- **Color:** E.g "Tan", "Brown/White", "Orange Tabby", "Black/White Point”
- **OutcomeType (train only)**: Adoption, transfer, return to owner, died, euthanasia.

FEATURE SUMMARY

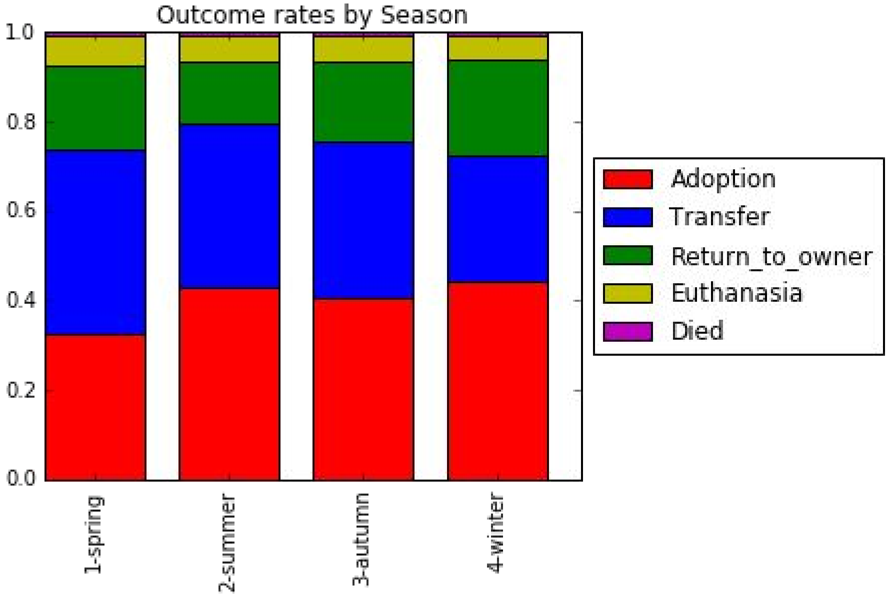
	AnimalID	Name	DateTime	OutcomeType	OutcomeSubtype	AnimalType	SexuponOutcome	AgeuponOutcome	Breed	Color
count	26729	19038	26729	26729	13117	26729	26728	26711	26729	26729
unique	26729	6374	22918	5	16	2	5	44	1380	366
top	A705677	Max	2015-08-11 00:00:00	Adoption	Partner	Dog	Neutered Male	1 year	Domestic Shorthair Mix	Black/White
freq	1	136	19	10769	7816	15595	9779	3969	8810	2824
first	NaN	NaN	2013-10-01 09:31:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
last	NaN	NaN	2016-02-21 19:17:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN

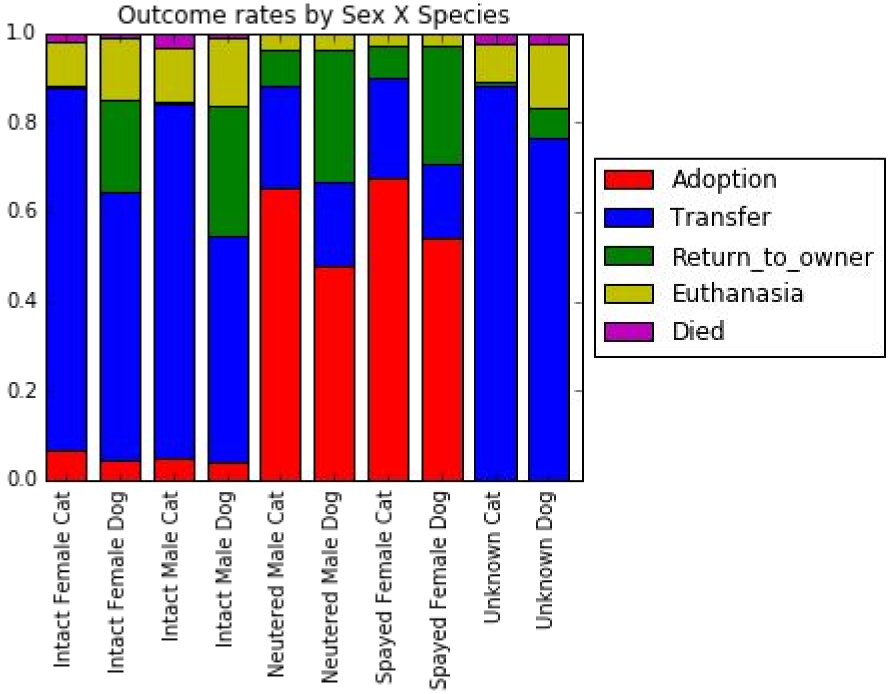
Exploratory Data analysis



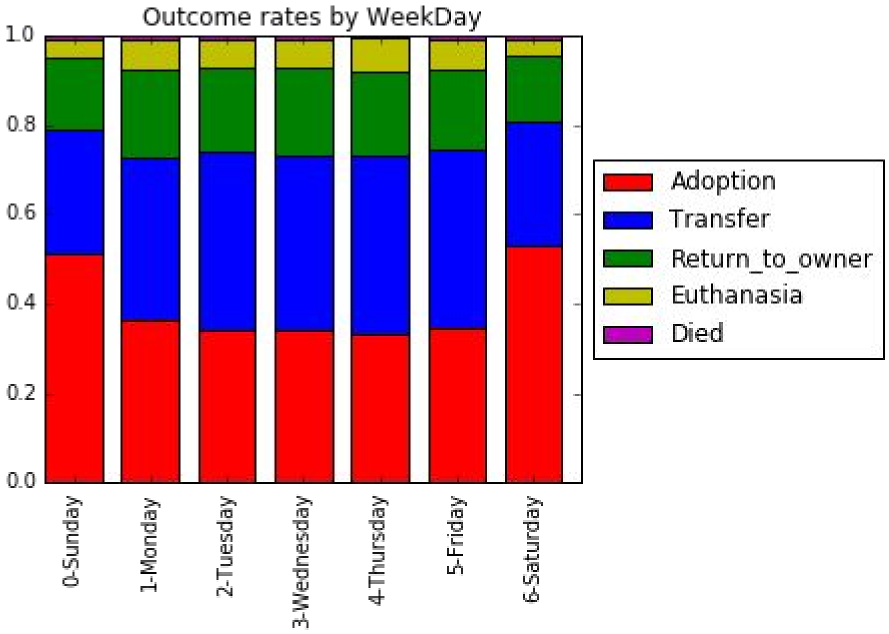


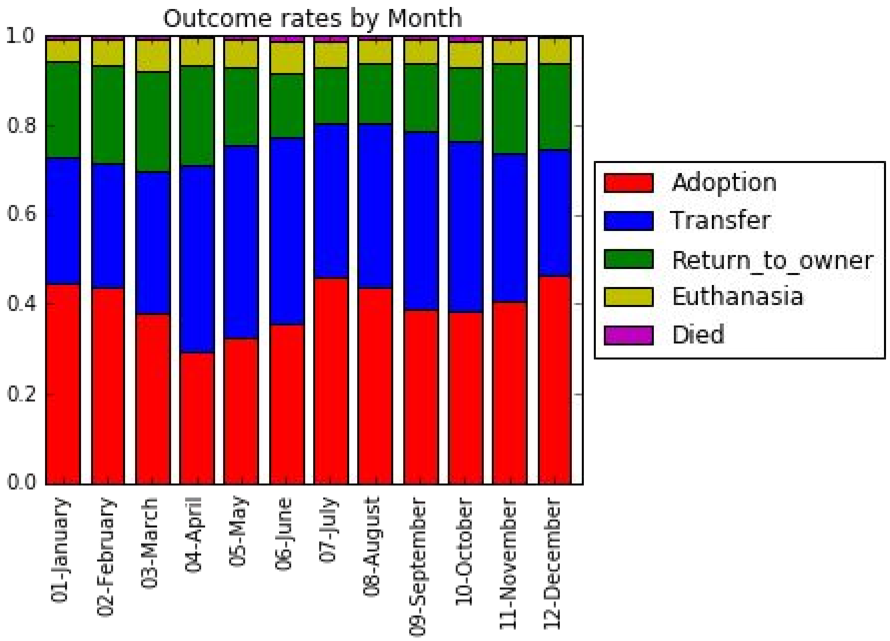
Exploratory Data analysis





Exploratory Data analysis





FEATURE ENGINEERING NOTES

- Combined training and submission data for data exploration and feature engineering.
  - In real life this should not be done because it will adversely impact the generalizability of the data.
  - The Kaggle competition is already different from real life in many ways.
  - The train and test sets include some variables that can only be known after the outcome has already been determined: E.g Outcome datetime, SexUponOutcome, AgeUponOutcome
  - Jupyter notebooks are more manageable if most of the work is done in a global scope.
- Handling Null values: Coded as Unknown category

FEATURE SELECTION

Original Field	FSV1	FSV2	FSV3
Name	Ignored	Binary if name exists.	Binary if name exists.
DateTime	Dummy coded as season.	Split into Season, Month, WeekOfYear, WeekDay, Hour, AmPM and dummy coded.	Split into Season, Month, WeekOfYear, WeekDay, Hour, AmPM and dummy coded.
AnimalType	Binary	Binary	Binary
AgeuponOutcome	Transformed into years and split into quartiles.	Split into separate categories for dogs/cats.	Split into separate categories for dogs/cats.
SexuponOutcome	Split into gender and fixed or not, both binary	Dummy coded	Dummy Coded

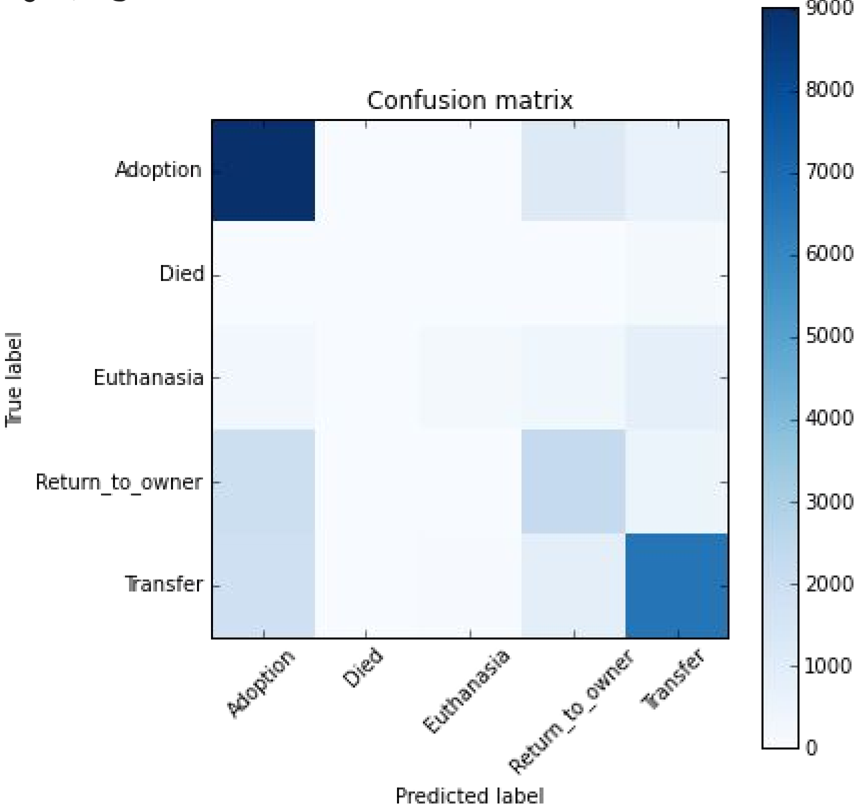
Original Field	FSV1	FSV2	FSV3
Breed	Split into primary and secondary breed then dummy coded.	Split into purebreed, mixed, etc. then dummy coded using countvectorizer.	Split into primary and secondary breed and then dummy coded. Whether mix or not extracted as binary.
Color	Split into primary and secondary color and dummy coded.	Dummy coded using countvectorizer.	Color modifiers extracted, merle/tick/tabby/brindle/point, along with primary and secondary color and dummy coded.

Classification

- Experimented with a number of classifiers.
- Quickly narrowed down to just logistic regression with L1 regularization, and random forests.
- Logistic regression informed much of initial exploration and feature engineering.
- Settled on random forests for Kaggle submissions because of the better classification accuracy.

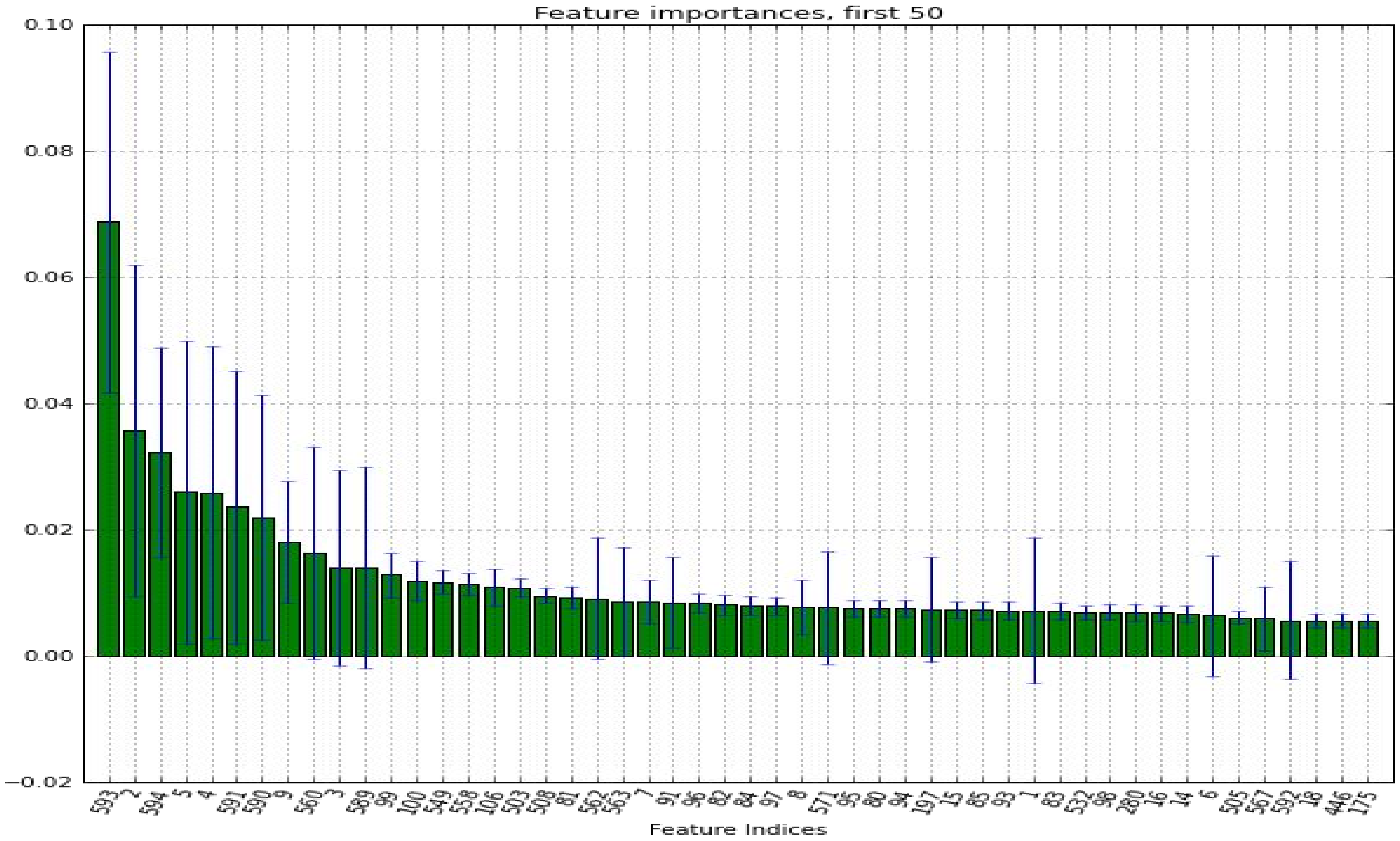
Model Evaluation - Random FOREST

- Evaluated stability and accuracy of the model.
- Use K-Fold CV (K=6)
- Log Loss (each iteration) : 0.78907, 0.79916,0.78273, 0.80660, 0.78045, 0.79239
- Standard deviation: 0.00973
- Mean Log Loss: 0.792



Final Random Forest Classification Report

	Precision	Recall	F1-score	Support
Avg / Total	0.68	0.68	0.66	26729


Feature importances, first 50
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
593 2 594 5 4 591 590 9 560 3 589 99 100 549 558 106 503 508 81 562 563 7 91 96 82 84 97 8 571 95 80 94 197 15 85 93 1 83 532 98 280 16 14 6 505 567 592 18 446 175
Feature Indices

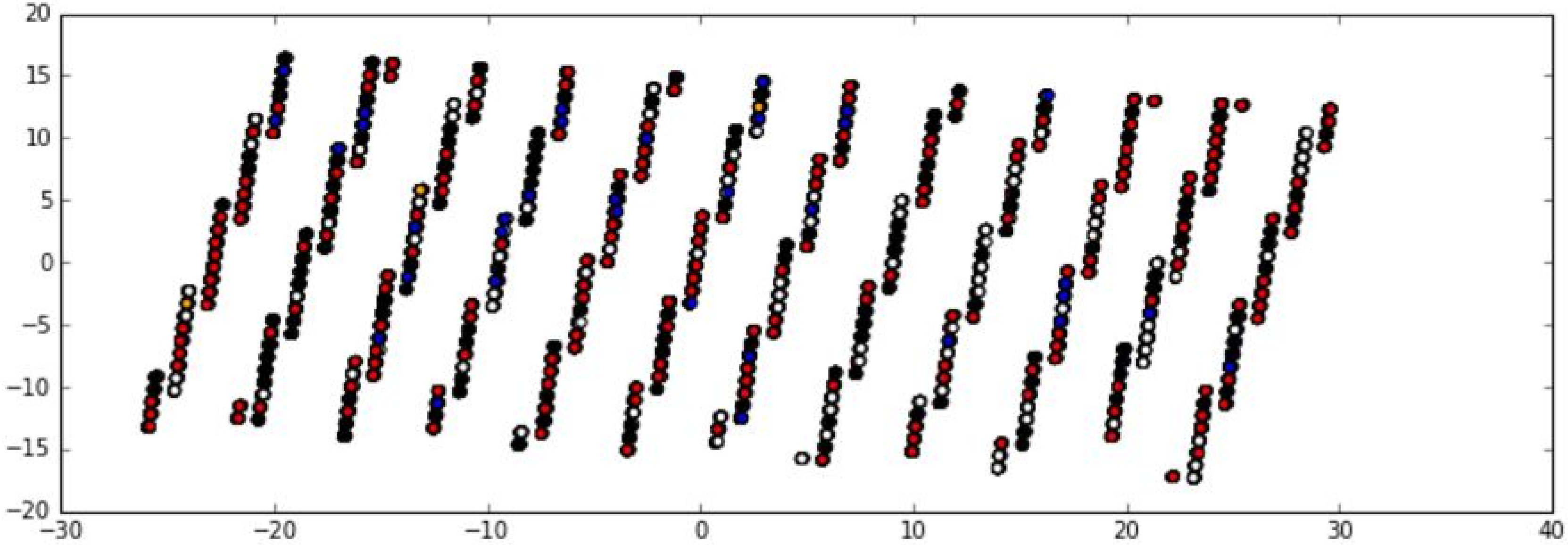
Feature Importance

1. Sex[T.
NeuteredMale]:IsMix
[T.True]
2. AgeCategory[T.baby]
3. AnimalType[T.Dog]:
HasName[T.True]
4. Sex[T.Intact Male]
5. Sex[T.Intact Male]:
IsMix[T.True]
6. Hour[T.17]
7. Hour[T.18]
8. Color2[T.None]
9. Color2[T.White]
10. AmPm[T.PM]
11. Color1[T.Black]
12. Color1[T.Brown]

Other Algorithms

- **Support Vector Machines**:
  - RBF Kernel showed some promise (results similar to LR)
  - Excessive training times
  - No good decision boundary to the data
- **Single decision trees:** Showed initially promising results, but quickly moved to ensembles of trees.
- **Gaussian Mixture Models**: Also dropped due to poor initial performance.
- **AdaBoost with trees**: Performance was highly variable but generally poor compared to random forests.
- **AdaBoost with logistic regression**: Surprisingly didn't perform particularly well compared to logistic regression, and training times were relatively long.

BACKUP - PCA in 2D



Red=Adoption; Orange=Died; Blue-Euthanasia; White=returnToOwner; Black=transfer

DASHBOARD

97	↓13	UTARDM2016- Team10	0.76576	21	Sun, 17 Apr 2016 10:57:18 (-17.3h)
98	↓12	blandard01	0.76778	2	Tue, 29 Mar 2016 16:00:22
99	↑18	ssdf93	0.76852	12	Mon, 25 Apr 2016 12:52:45 (-3d)
100	↑2	dimka	0.76904	4	Thu, 21 Apr 2016 21:21:42
101	↑99	GAVOILLEGuillaume	0.77025	2	Wed, 20 Apr 2016 22:05:51
102	↓15	Michael Semeniuk	0.77142	1	Sun, 17 Apr 2016 05:07:54
103	**↑53**	**Cats and Dogs**	**0.77293**	**14**	**Mon, 25 Apr 2016 04:35:50**
104	↑95	DSM	0.77476	9	Mon, 25 Apr 2016 15:18:49
105	↓17	sagar verma	0.77682	16	Sat, 16 Apr 2016 10:26:11 (-2.8h)
106	↑41	Chris Muenzer	0.77823	9	Mon, 25 Apr 2016 20:53:31

CONCLUSION

- ❑ The model that gives the best Kaggle score is not necessarily the best model when predicting the outcome of real world data.
- ❑ Feature engineering is more important than choice of algorithms.
- ❑ Merging different approaches towards feature engineering was not an easy task.