Machine Learning: Animal Shelter Adoption

Team 1-01: Mark, Michael, Isabel, and Len

April 16, 2018

Overview

- Business context background
- Data
  - Original
  - Cleansing
  - Final
- Modeling process
- Cost evaluation
- Summary
- Appendices

Can you find us a home?

The Story: AAC Austin Animal Center

Austin Animal Center runs the largest **no-kill** municipal animal shelter in the United States, providing shelter to more than **18,000 animals each year** and animal protection services to all of Austin and Travis County. They have a large (+ 900 families) **animal foster care program**.

Animals in need of shelter are accepted regardless of age, health, species or breed.

Travis County, Texas
Population: 1.18 Million
Median Age: 33
Median HH Income:$65,269
Source: https://datausa.io

Source: http://www.austintexas.gov/department/aac

The Original Cat Data

Austin Animal Center Shelter Outcomes

Source:
https://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-outcomes-and/data

29,421 observations
37 variables *(columns)*

Data types:

- Numeric
- Date/Time
- String
- Boolean
- Categorical

MISSING VALUES:
Cells with missing value: 85,885

These columns had missing values:

1.	Color:	3,626
2.	Name:	12,774
3.	Outcome_subtype:	10,780
4.	Outcome_type:	3
5.	Breed2:	29,369
6.	Coat_pattern:	10,266
7.	Color2:	19,067

Data Cleansing

1. **Removed 25 variables**
    a. Repetitive or data we wouldn't have known when an animal arrived
2. **Removed 3 observations**
    a. Contained unknown response variable
3. **Relabeled "Name" as binary**
    a. Animal already had a name = 1
4. **Turned "outcome_type" into 2 classes**
    a. Adopt/Return to Owner/RTO-Adopt was "Adopt", everything else "Other"
5. **Turned "period_of_stay" into 2 classes**
    a. <= 30 days is "Short," other is "Medium-Long"
6. **Empty cells assigned "Not-Known" value**
7. 3 **coat_patterns** merged into **"rare_coat_pattern"** since zero variances
8. **Rare "breed"** (<= 50 observations) lumped together
9. **Rare** (<= 10 observations) **"coat"** lumped together
10. **Rare** (<= 18 observations) **"color"** lumped together

Data Cleansing - Columns Removed

- Age_upon_outcome
- Animal_id
- Animal_type
- Date_of_birth
- Date_time
- Monthyear
- Outcome_subtype
- Sex_upon_outcome
- Cat.kitten..outcome
- Count
- Periods
- Period.range
- Outcome_age.years.

- Sex_age_outcome
- Dob_year
- Dob_month
- Dob_monthyear
- Outcome_month
- Outcome_year
- Outcome_weekday
- Outcome_hour
- Breed1
- Breed2
- Color1
- color2

Final Data

29,418 observations of **10** predictors (which were then turned into dummies)

1. **Name:** binary
2. **Sex:** binary
3. **Spay.neuter:** binary
4. **Age_group:** factor 10 levels
5. **CFA_breed:** binary
6. **Domestic_breed:** binary
7. **Coat_pattern:** factor 8 levels
8. **Breed_merged:** factor 10 levels
9. **Coat_merged:** factor 29 levels
10. **Color_merged:** factor 34 levels

*coat_pattern and coat_merged are distinctly different

Response variables:

Outcome_type: Binary

Adopt vs. Other

14,196 vs. 15,222 observations

Period_of_stay: Binary

<= 3 months vs.

> 3 months

15,327 vs. 14,091 observations

Our primary modeling GOAL was to train a classification model that could predict whether a cat would be adopted or not

Costs - Assumptions

“Adopt = no or 0” & “**Other** = yes or 1”

- **Adopt correct cost (TN)**: $20 adoption cost, for 2 months of adoption events on average
- **Other correct cost (TP):** $50 finding foster care and other cost
- **Adopt incorrect guess cost (FN, Type II)**: $70 we show the cat at adoption events but then we also have to find it foster care because it wasn’t adopted
- **Other incorrect guess cost (FP, Type I):** $50 for finding foster care but it’s somehow adopted but we didn’t show it at events so don’t have event costs

Modeling

Packages:

- Caret
- doParallel
- Snow
- stringr

Models

1. SVM linear
2. SVM radial
3. Classification tree
4. Random forest
5. Boosting

Trained and tuned on **8,000 rows** of data subset

Applied **3-folds cross-validation**

Evaluated and compared models on **overall accuracy**

Validated the best model by running on the best subset model on the non-subset data

*Note: the scale of the Y-axes are designed to show relative change and do not start at 0!

Modeling 1 - 1. SVM Linear Kernel

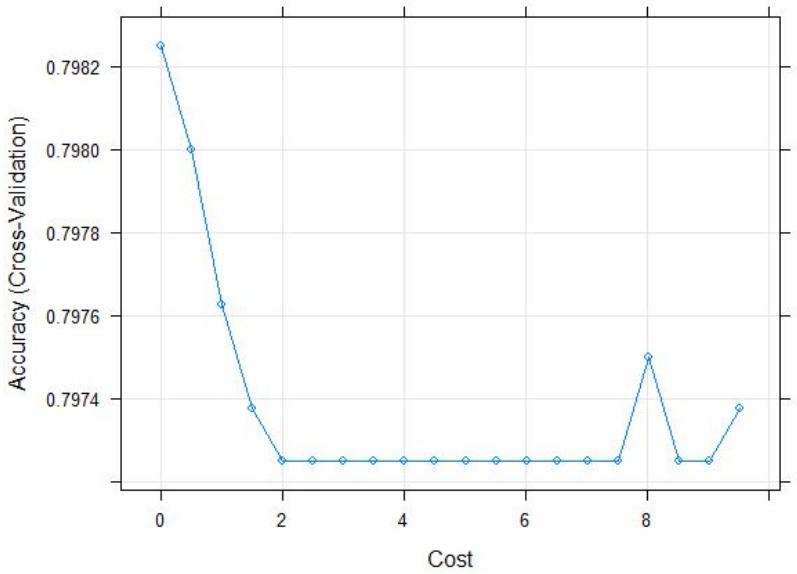
Optimal Model Parameter(s):

Cost: 0.01

Confusion Matrix:

	Prediction	
Actual	Adopt	Other
Adopt	1031.7	275.7
Other	262.3	1097
Accuracy	0.7982	

Learning Curve:



Modeling 1 - 2. SVM Radial Kernel

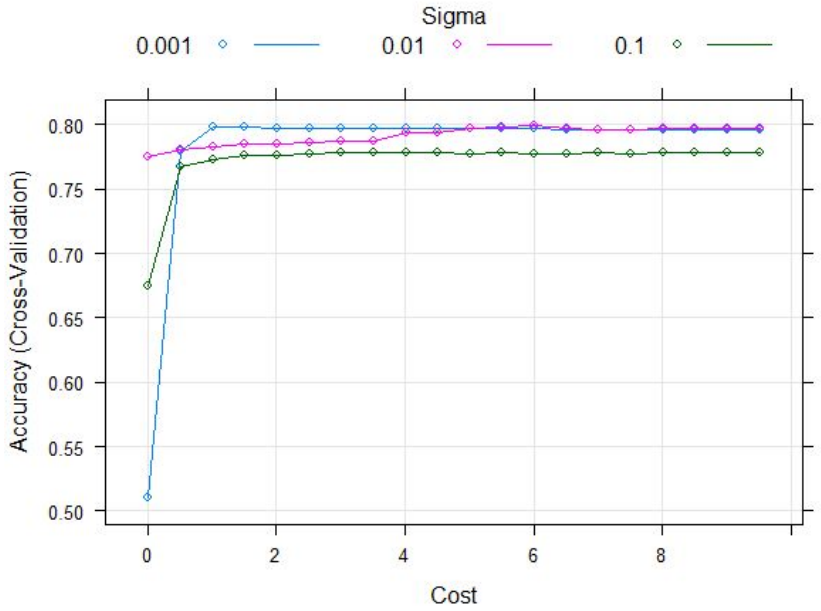
Optimal Model Parameter(s):

Cost: 6.01
Sigma: 0.01

Confusion Matrix:

	Prediction	
Actual	Adopt	Other
Adopt	1098	209.3
Other	326.7	1032.7
Accuracy	0.799	

Learning Curve:



Modeling 1 - 3. Classification Tree

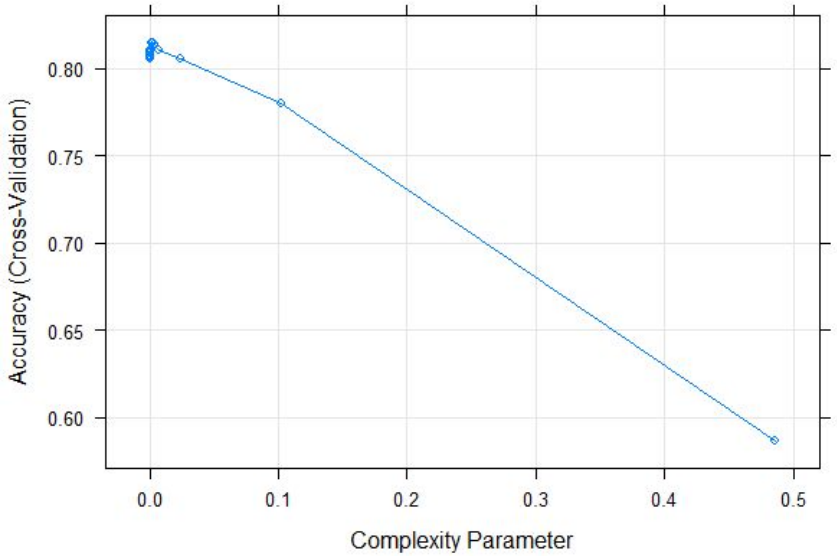
Optimal Model Parameter(s):

Cp: 0.002294748

Confusion Matrix:

	Prediction	
Actual	Adopt	Other
Adopt	1120.7	186.7
Other	308	1051.3
Accuracy	0.8145	

Learning Curve:



Modeling 1 - 4. Random Forest

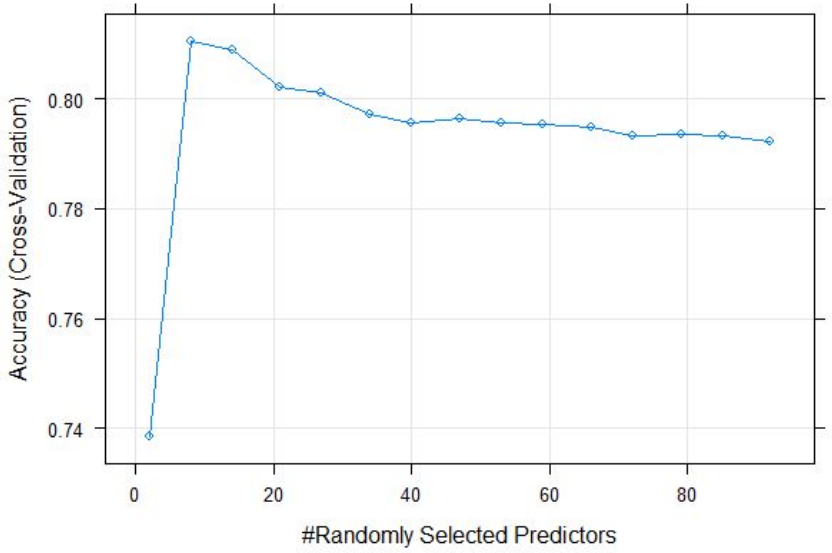
Optimal Model Parameter(s):

Mtry: 8

Confusion Matrix:

	Prediction	
Actual	Adopt	Other
Adopt	1140.7	166.7
Other	337	1022.3
Accuracy	0.8111	

Learning Curve:



Variable Rankings from 4. Random Forest

1. Name (yes)
2. Sex (male)
3. Spay.Neuter (yes)
4. Age_group
5. CFA_breed (true)
6. Domestic_breed (true)
7. Coat_pattern

The shelter should ensure the proper recording of the information of each animal they are taking in, as it's useful in the models and to the future improvements of their planning.

Modeling 1 - 5. Boosting

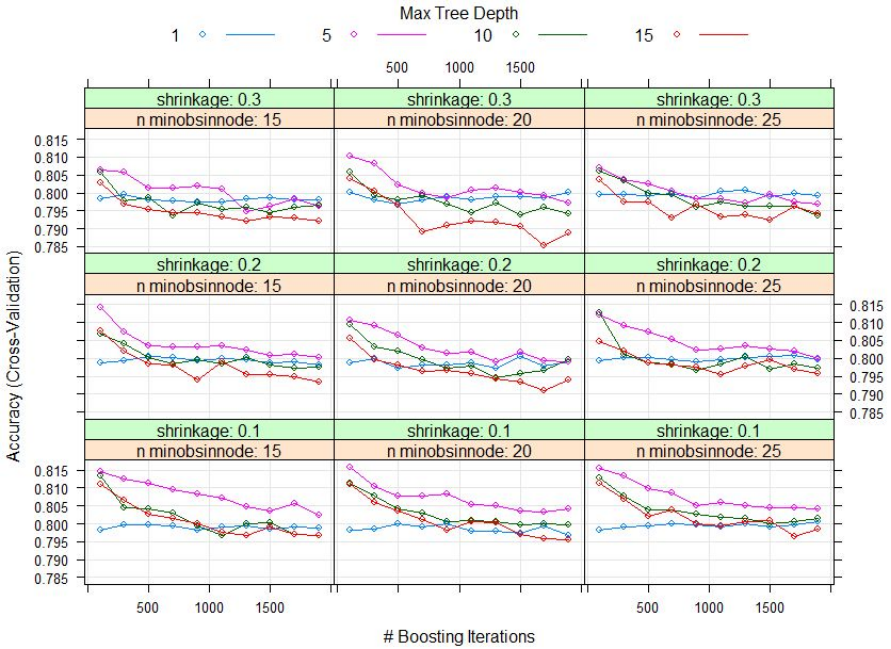
Optimal Model Parameter(s):

N.trees: 100 Shrinkage: 0.1
Interaction.depth: 5 N.minobsinnode: 20

Confusion Matrix:

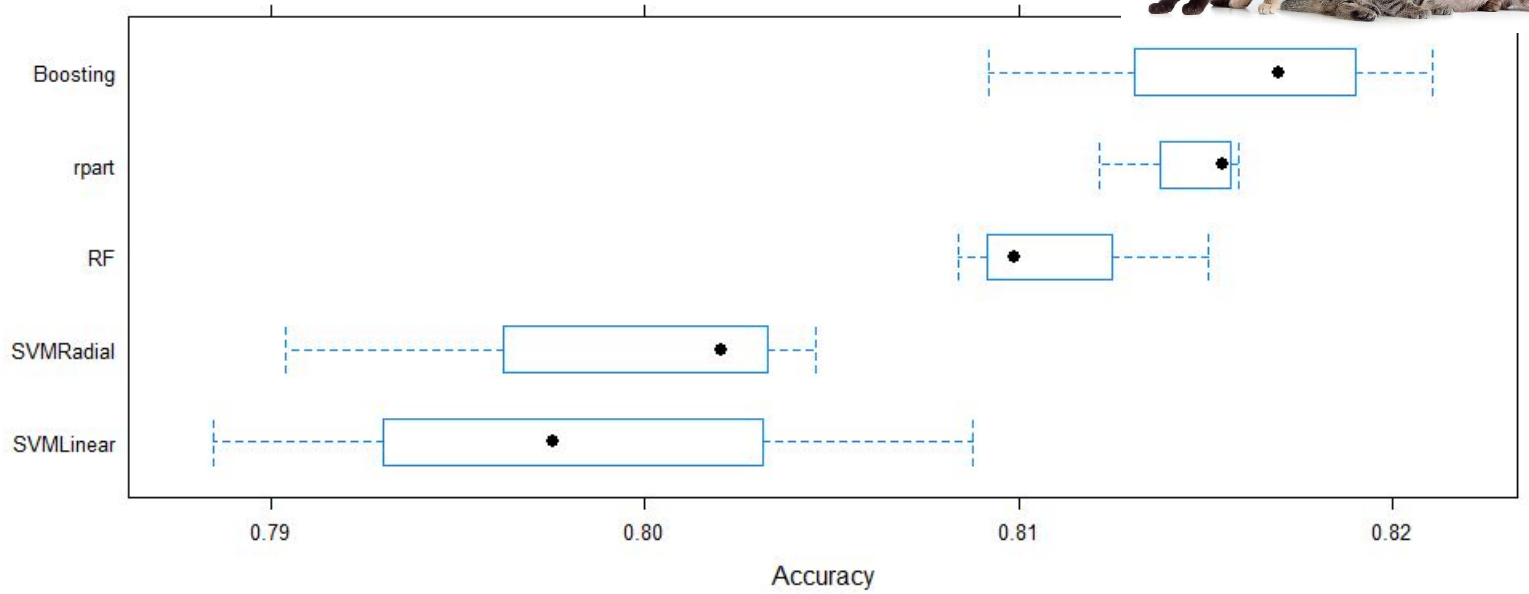
	Prediction	
Actual	Adopt	Other
Adopt	1122.7	184.7
Other	306.7	1052.7
Accuracy	0.8157	

Learning Curve:



Modeling 1 - Accuracy Comparison with C.V.

Visualize CV accuracy spread and compare mean accuracy



Modeling 1 - Parameter Validation (Boosting)

Cross-validate (8 folds) the optimal model with its best parameters on the dataset without the training observations

Compare results of this validation dataset CV with that of the training subset

Tuning Confusion Matrix:

Over training dataset

	Prediction	
Actual	Adopt	Other
Adopt	1120.7	186.7
Other	308	1051.3
Accuracy	0.8145	

V.S.

Validation Confusion Matrix:

Over entire dataset

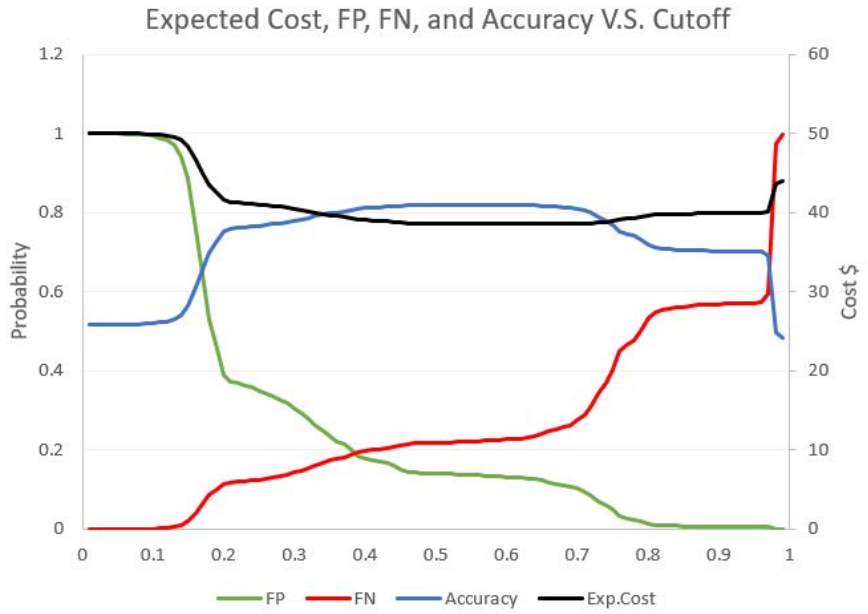
	Prediction	
Actual	Adopt	Other
Adopt	1105.4	178.9
Other	304.1	1088.9
Accuracy	0.8196	

Cost Evaluation - Cutoff Analysis on Full Dataset

Results at the cutoff based on the optimal expected cost:

Cutoff	0.69
Exp.Cost	$38.55
Accuracy	0.8127
Type - I Error	0.1068
Type - II Error	0.2624
True Positive	0.7376
True Negative	0.8932

Expected Cost:



Summary and Conclusions - Business Implications

Our model may be one way to help the shelter plan its capacity better when it receives new animals

- Communication with other shelters
- Knowing if you need more volunteers to run events or socialize the animals
- Getting in contact with potential foster families in a timely manner

The adoption center may want to consider tracking other variables such as:

- Health rating
- Friendliness rating/socialized to living with other animals
- Time in foster care if an animal returns to the shelter

Using our boosting model, it may be possible to help a shelter plan if an animal is likely to be adopted or not adopted

Appendix A:
Secondary Modeling for Period of Stay

Our secondary modeling GOAL was to train a classification model that could predict if a cat would stay less than 3 months or more than 3 months

Modeling 2: Data

We used the same data, methodologies, and cost function as when we modeled outcome_type; however, we excluded “outcome_type” as a predictor variable, just as “period_of_stay” was excluded when we modeled “outcome_type”

Modeling 2 - 1. SVM Linear Kernel

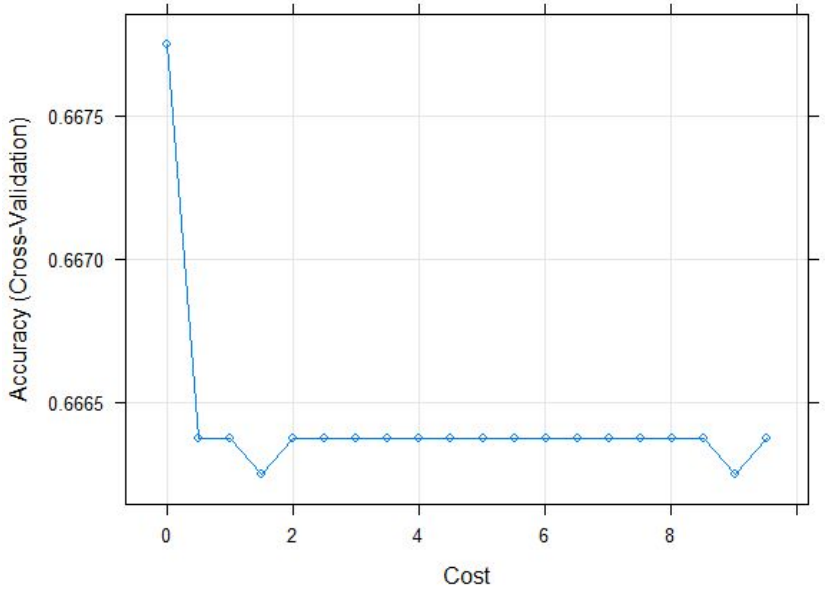
Optimal Model Parameter(s):

Cost: 0.01

Confusion Matrix:

	Prediction	
Actual	Short	Medium-Long
Short	1342	36.7
Medium-Long	849.3	438.7
Accuracy	0.6677	

Learning Curve:



Modeling 2 - 2. SVM Radial Kernel

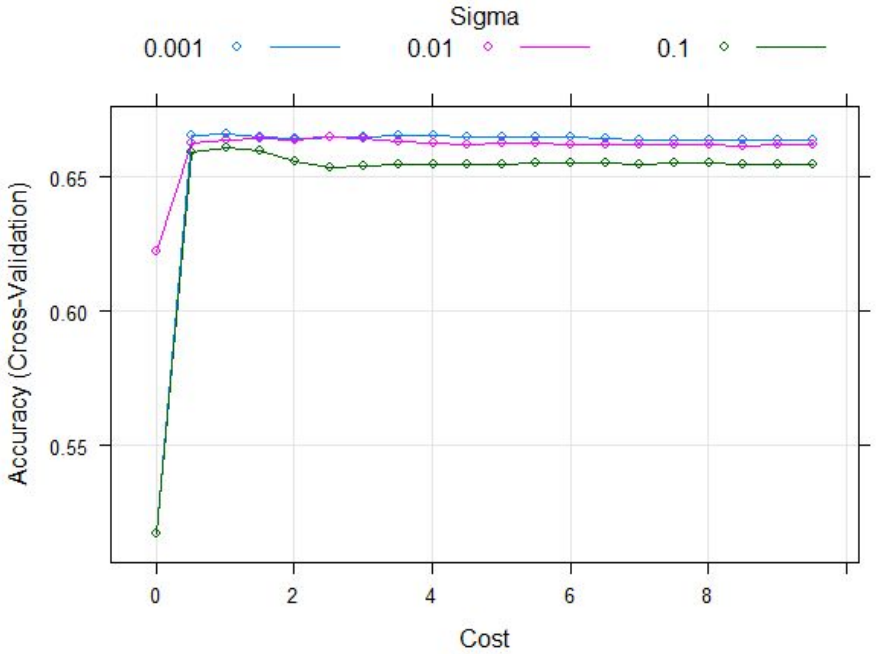
Optimal Model Parameter(s):

Cost: 1.01 Sigma 0.001

Confusion Matrix:

	Prediction	
Actual	Short	Medium-Long
Short	1341.7	37
Medium-Long	854.3	433.7
Accuracy	0.6657	

Learning Curve:



Modeling 2 - 3. Classification Tree

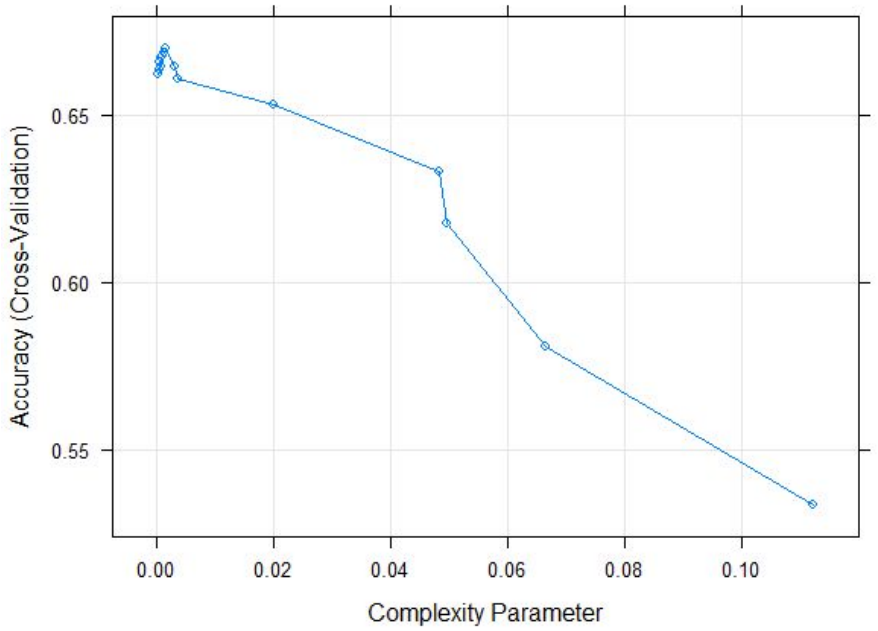
Optimal Model Parameter(s):

CP: 0.001423395

Confusion Matrix:

	Prediction	
Actual	Short	Medium-Long
Short	1235.7	143
Medium-Long	737.7	550.3
Accuracy	0.6697	

Learning Curve:



Modeling 2 - 4. Random Forest

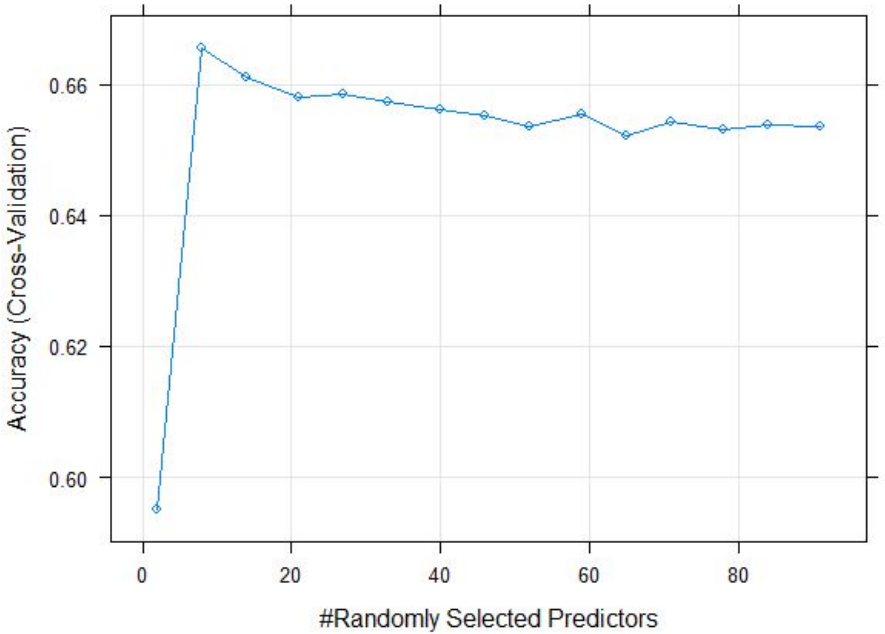
Optimal Model Parameter(s):

Mtry: 8

Confusion Matrix:

	Prediction	
Actual	Short	Medium-Long
Short	1194.3	184.3
Medium-Long	707.7	580.3
Accuracy	0.6655	

Learning Curve:



Modeling 2 - Variable Rankings

1. Name1
2. SexMaleYes
3. Spay.NeuterYes
4. Age_group
5. CFA_breedTrue
6. Domestic_breedTrue
7. Coat_patternNotKnown
8. Coat_patternPoint

Given the weak accuracies of these models, the shelter may consider exploring other variables.

The shelter should ensure the proper recording of the information of each animal they are taking in, as it's useful in the models and to the future improvements of their planning.

Modeling 2 - 5. Boosting

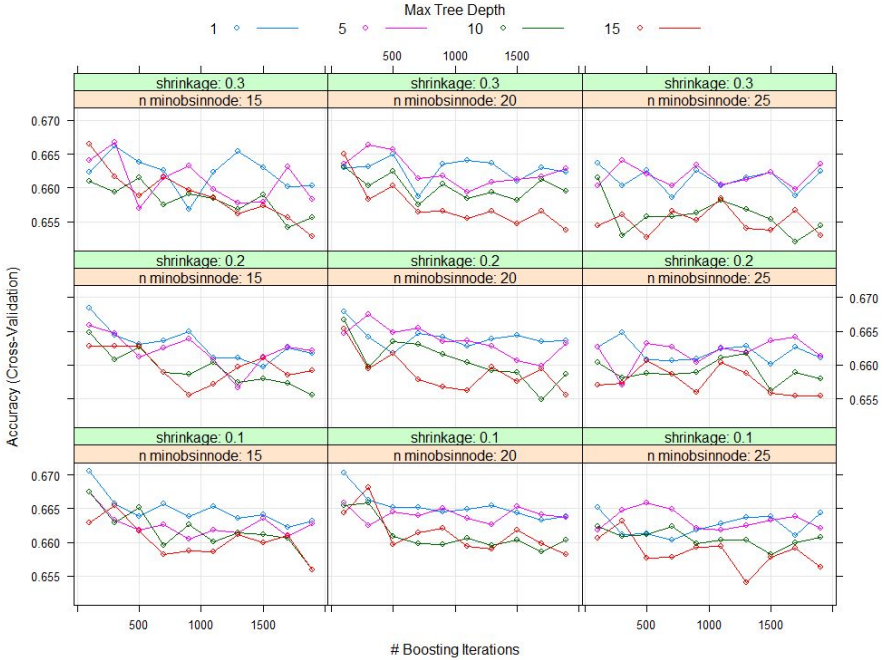
Optimal Model Parameter(s):

N.trees: 100 Shrinkage: 0.1
Interaction.depth: 1 N.minobsinnode: 15

Confusion Matrix:

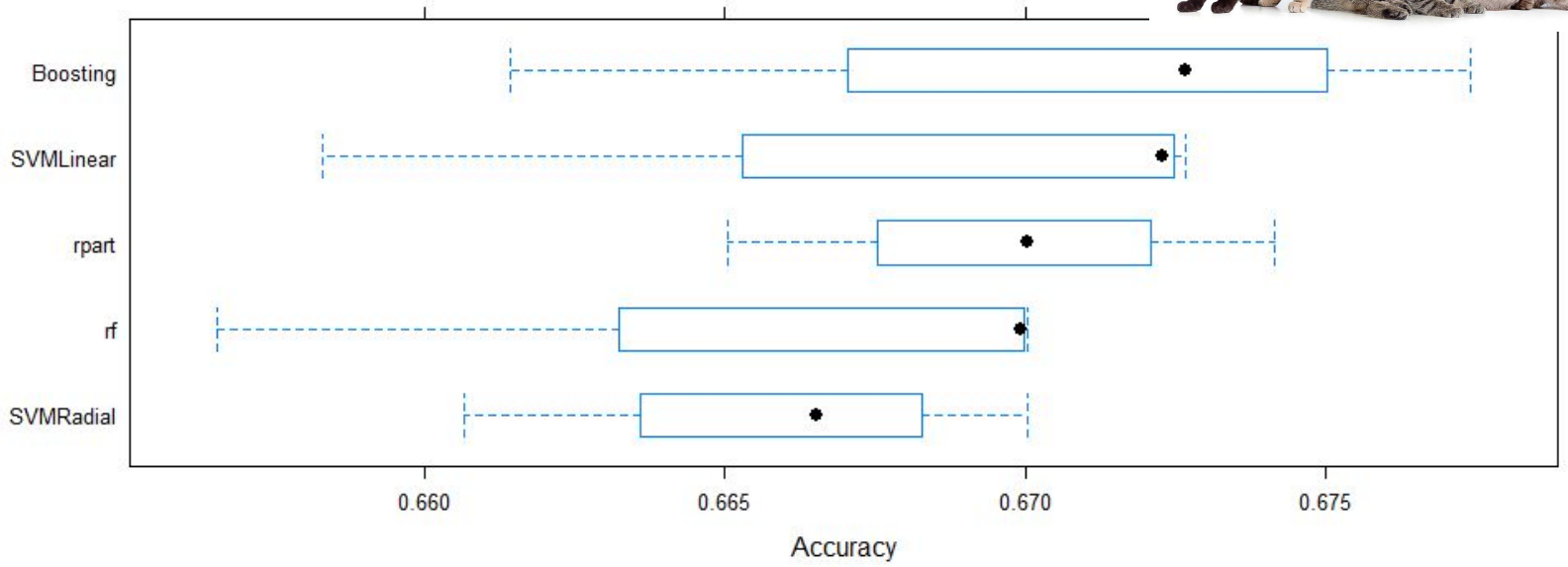
	Prediction	
Actual	Short	Medium-Long
Short	1350	28.7
Medium-Long	850	438
Accuracy	0.6705	

Learning Curve:



Modeling 2 - Accuracy Comparison with C.V.

Visualize CV accuracy spread and compare mean accuracy



Modeling 2 - Parameter Validation (Boosting)

Cross-validate (8 folds) the optimal model with its best parameters on the dataset without the training observations

Compare results of this validation dataset CV with that of training subset

Tuning Confusion Matrix:

Over training dataset

	Prediction	
Actual	Short	Medium-Long
Short	1350	28.7
Medium-Long	850	438
Accuracy	0.6705	

V.S.

Validation Confusion Matrix:

Over entire dataset

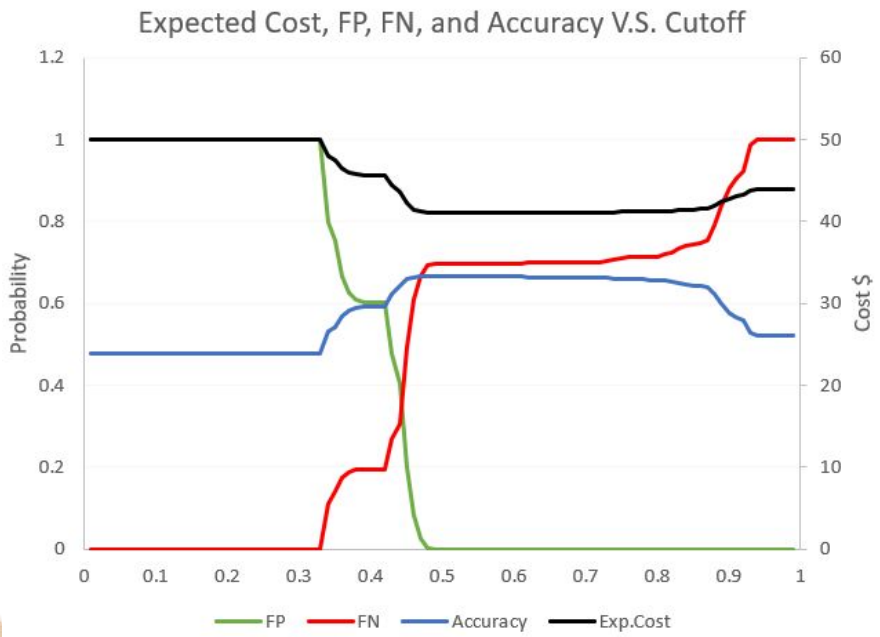
	Prediction	
Actual	Short	Medium-Long
Short	1398.9	0
Medium-Long	894.4	384
Accuracy	0.6659	

Cost Evaluation - Cutoff Analysis on Full Dataset

Results at the cutoff based on the optimal expected cost:

Cutoff	0.49
Exp.Cost	$41.04
Accuracy	0.6665
Type - I Error	0
Type - II Error	0.6962
True Positive	0.3038
True Negative	1

Expected Cost:



Summary and Conclusions - Business Implications

The period_of_stay modeling had weak performance in overall accuracy

- 66.6% accuracy rate

We likely need other data to help understand why animals stay for different periods of time

- Date when animal enters
- Health rating
- Friendliness rating/socialized to living with other animals

Our model may be one way to help the shelter plan its capacity better when it receives new animals

- Streamlining animals for adoption events
- Contacting foster families in a timely manner
- Helping order supplies more efficiently having a general sense of how many animals there will be
- Planning schedules for volunteers

Appendix B:
Raw Model Accuracy Comparisons for Model 1 and Model 2

Modeling 1 - Model Accuracy Comparison

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
SVMLinear	0.7884471	0.7929862	0.7975253	0.7982488	0.8031496	0.8087739
SVMRadial	0.7904012	0.7962130	0.8020247	0.7990007	0.8033004	0.8045761
rpart	0.8121485	0.8138012	0.8154539	0.8145001	0.8156759	0.8158980
RF	0.8083990	0.8091489	0.8098988	0.8111255	0.8124888	0.8150788
Boosting	0.8092204	0.8130873	0.8169542	0.8157516	0.8190173	0.8210803

Modeling 2 - Model Accuracy Comparison

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
SVMLinear	0.6582896	0.6652903	0.6722910	0.6677488	0.6724784	0.6726659
SVMRadial	0.6606674	0.6636045	0.6665416	0.6657501	0.6682914	0.6700412
rpart	0.6650413	0.6675413	0.6700412	0.6697494	0.6721035	0.6741657
rf	0.6565429	0.6632302	0.6699175	0.6655006	0.6699794	0.6700412
Boosting	0.6614173	0.6670416	0.6726659	0.6705009	0.6750426	0.6774194

Shelter Goals: The No Kill Plan

Comprehensive adoption program

Large volunteer foster program

Reduce euthanasia of owned animals

Reduce Feral-Cat Euthanasia

High-Volume, Free and Low-Cost Spay/Neuter

Revise the Mission of Animal Services

Ensure a Compassionate Management and Staff

Public Involvement and Candor

Explore Outsourcing Possibilities

increasing adoption outcomes = saving lives