Ensemble Learning for Determining Shelter Animal Outcomes

Arjun Harish
Jie Yan
Yikun Li

Research Problem

- Kaggle competition dataset with 26729 rows of data.
- Animal outcomes include Adoption, Died, Euthanasia, Return to the owner, and Transfer.
- 2.7 million dogs and cats are euthanized in the US every year.

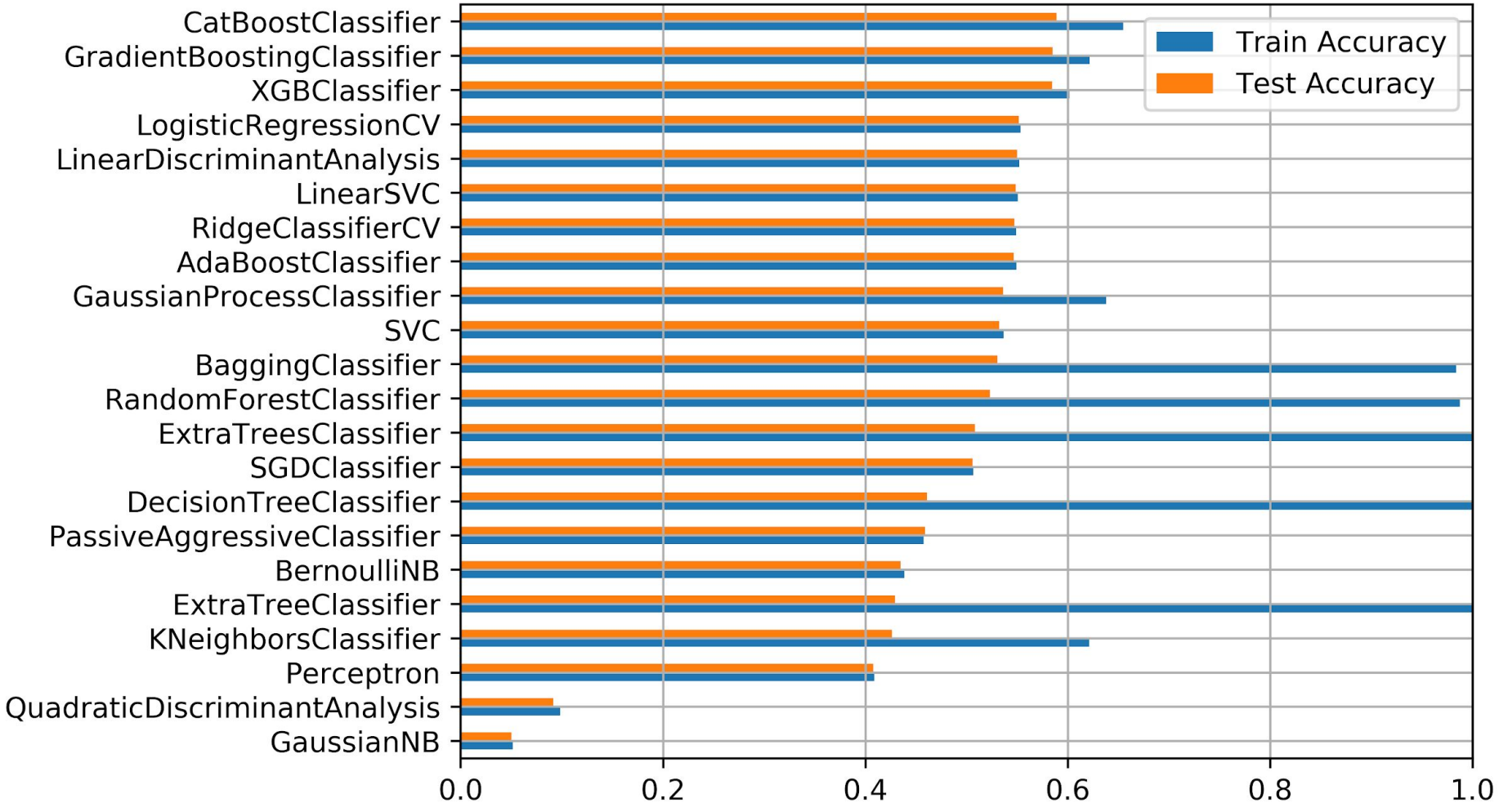
Feature Engineering

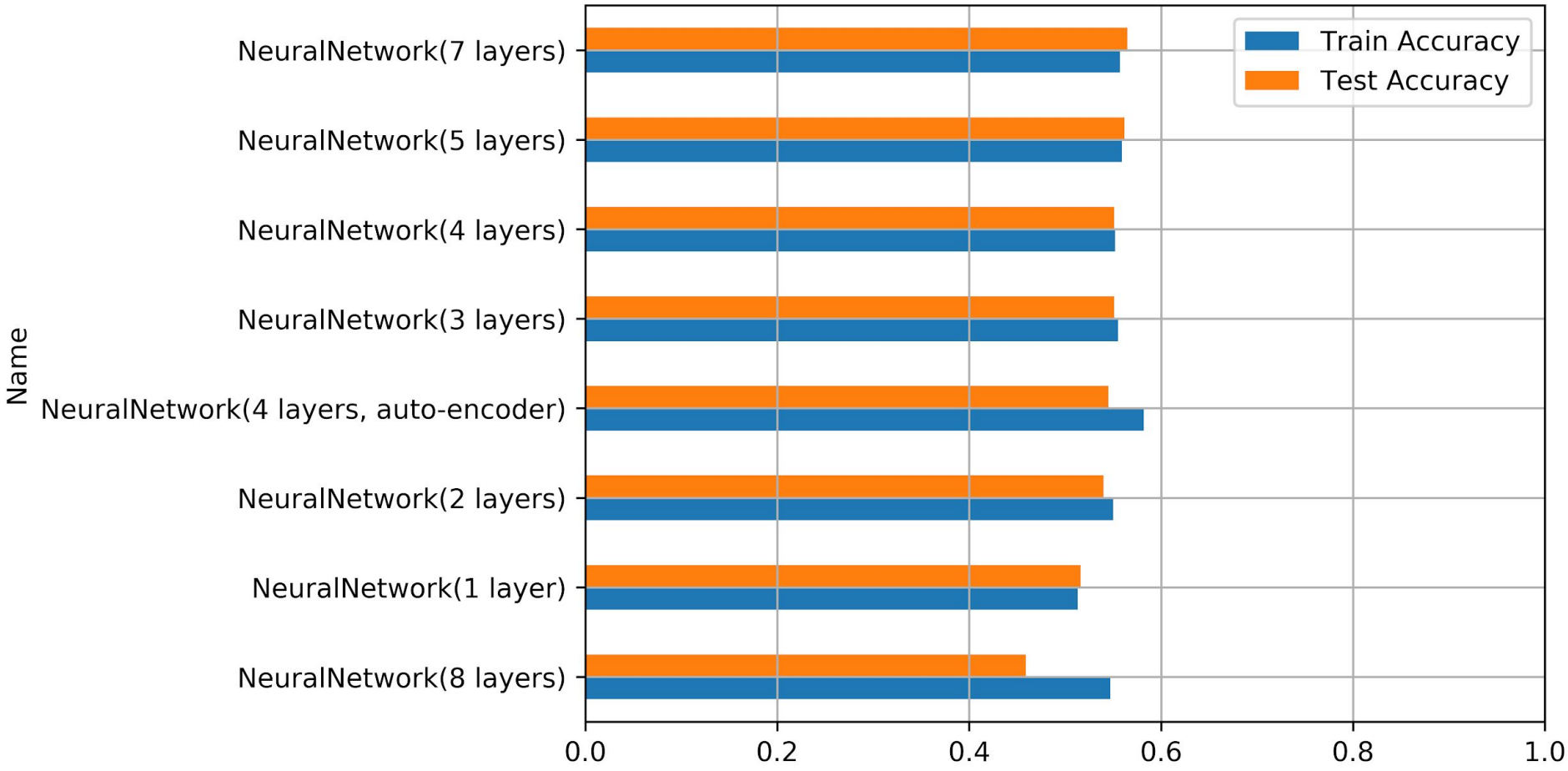
- Name - length of the name; whether name is presented or not; occurrence rate of the name
- Date time - year; month; day; hour; weekday
- Outcome - 5 types of outcomes
- Gender - female or male; neutered or not
- Age - squared root and normalized

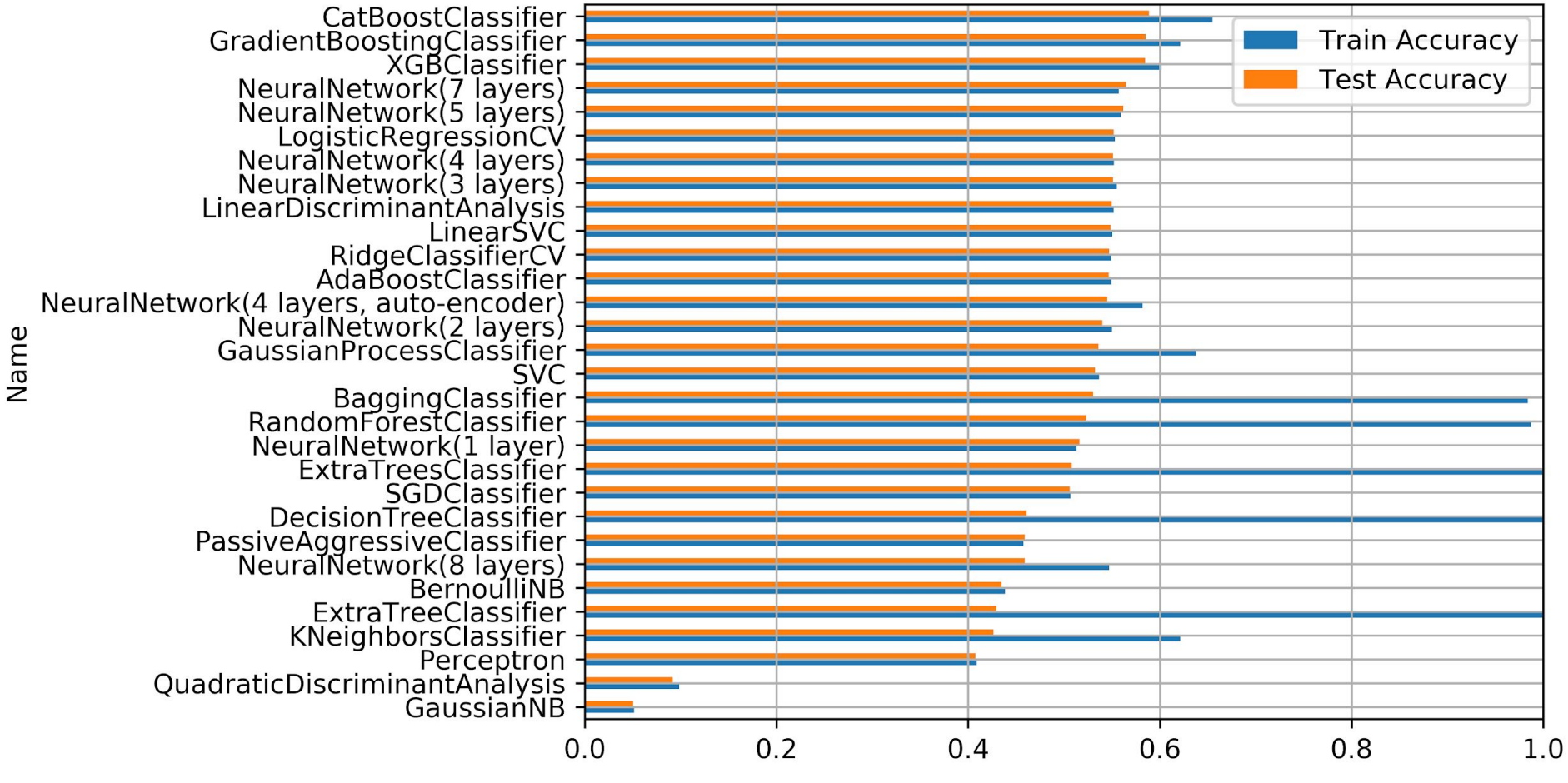
Feature Engineering

- Breed - by dog groups
  - Herding
  - Hound
  - Non-Sporting
  - Sporting
  - Terrier
  - Toy
  - Working
  - Mix
  - Unknown
- Color - by color theory
  - Dark shade
  - Medium shade
  - Light shade
  - Warm (red yellow)
  - Medium (grey, white, black)
  - Cold (blue)

In total 29 features.


Train Accuracy
Test Accuracy
Name
CatBoostClassifier
GradientBoostingClassifier
XGBClassifier
LogisticRegressionCV
LinearDiscriminantAnalysis
LinearSVC
RidgeClassifierCV
AdaBoostClassifier
GaussianProcessClassifier
SVC
BaggingClassifier
RandomForestClassifier
ExtraTreesClassifier
SGDClassifier
DecisionTreeClassifier
PassiveAggressiveClassifier
BernoulliNB
ExtraTreeClassifier
KNeighborsClassifier
Perceptron
QuadraticDiscriminantAnalysis
GaussianNB
0.0
0.2
0.4
0.6
0.8
1.0


Train Accuracy
Test Accuracy
NeuralNetwork(7 layers)
NeuralNetwork(5 layers)
NeuralNetwork(4 layers)
NeuralNetwork(3 layers)
NeuralNetwork(4 layers, auto-encoder)
NeuralNetwork(2 layers)
NeuralNetwork(1 layer)
NeuralNetwork(8 layers)
Name
0.0
0.2
0.4
0.6
0.8
1.0


Train Accuracy
Test Accuracy
Name
CatBoostClassifier
GradientBoostingClassifier
XGBClassifier
NeuralNetwork(7 layers)
NeuralNetwork(5 layers)
LogisticRegressionCV
NeuralNetwork(4 layers)
NeuralNetwork(3 layers)
LinearDiscriminantAnalysis
LinearSVC
RidgeClassifierCV
AdaBoostClassifier
NeuralNetwork(4 layers, auto-encoder)
NeuralNetwork(2 layers)
GaussianProcessClassifier
SVC
BaggingClassifier
RandomForestClassifier
NeuralNetwork(1 layer)
ExtraTreesClassifier
SGDClassifier
DecisionTreeClassifier
PassiveAggressiveClassifier
NeuralNetwork(8 layers)
BernoulliNB
ExtraTreeClassifier
KNeighborsClassifier
Perceptron
QuadraticDiscriminantAnalysis
GaussianNB
0.0
0.2
0.4
0.6
0.8
1.0

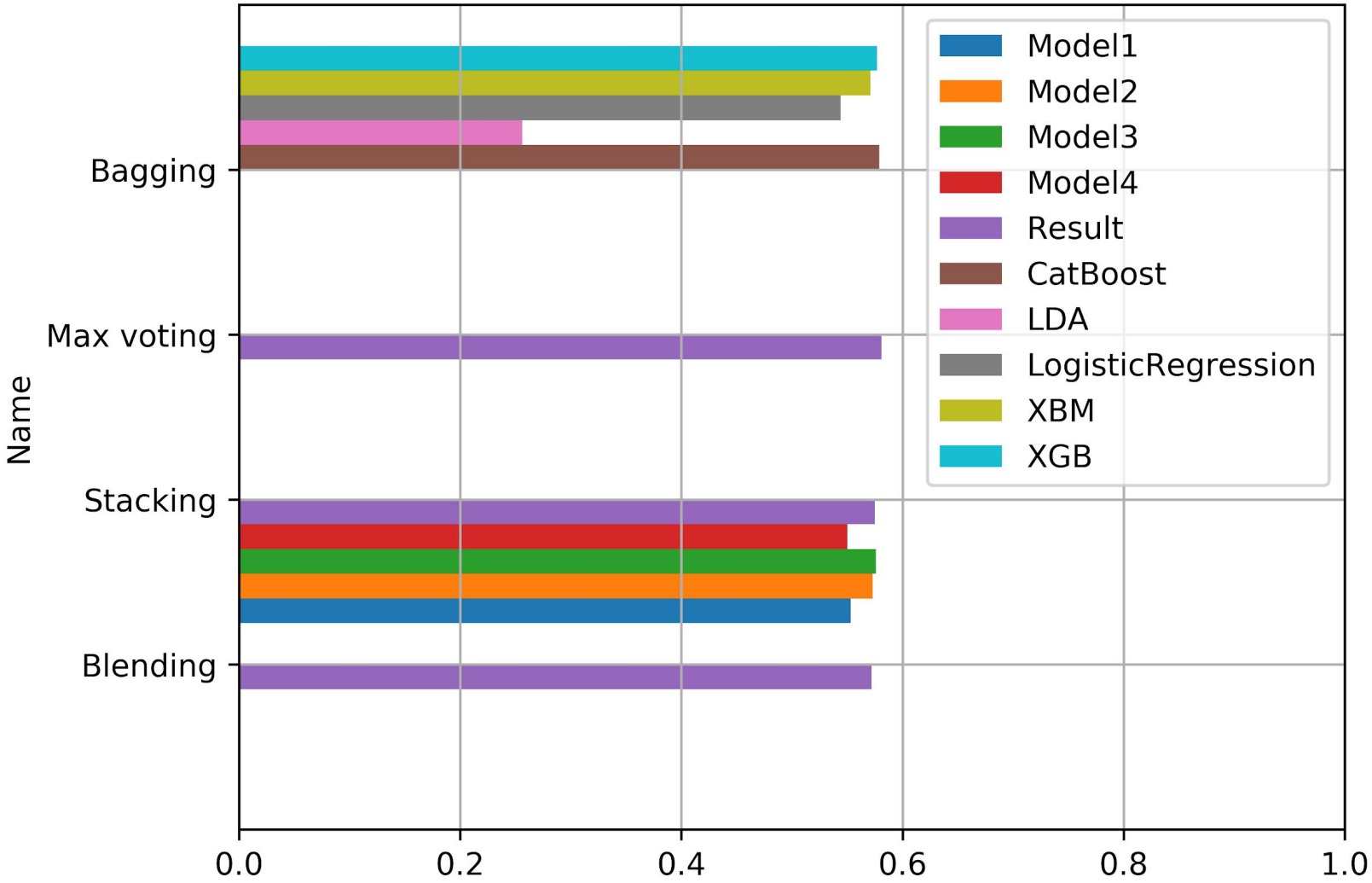
Parameter Tuning

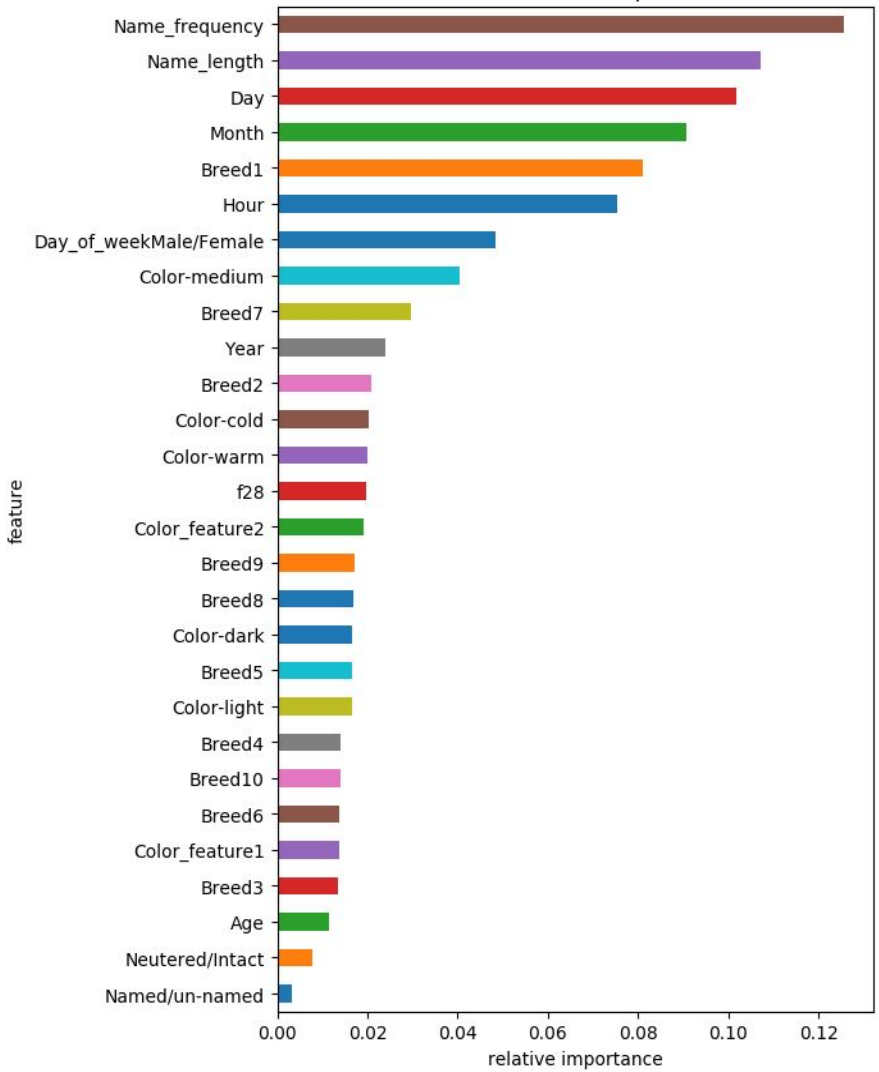
- Catboost:(iterations=500, depth=10,learning_rate=0.1,loss_function='MultiClass')
- XGB:('colsample_bytree': 0.38, 'learning_rate': 0.137, 'max_depth': 13, 'n_estimators': 72, 'nthread': 8, 'objective': 'multi:softprob', 'reg_alpha': 1.778, 'subsample': 0.98)
- XBM:(n_estimators:100, learning_rate:1.0, max_depth:1, random_state:0)
- LR:('penalty': 'l1')
- LDA:('solver': 'lsqr', 'shrinkage': 'auto', 'n_components': 0)

Algorithm	Before	After
Catboost	**0.58**	**0.59**
XGBoost	**0.58**	**0.583**
XBM	**0.58**	**0.58**
Logistic Regression	**0.54**	**0.55**
Linear Discriminant Analysis	**0.54**	**0.55**

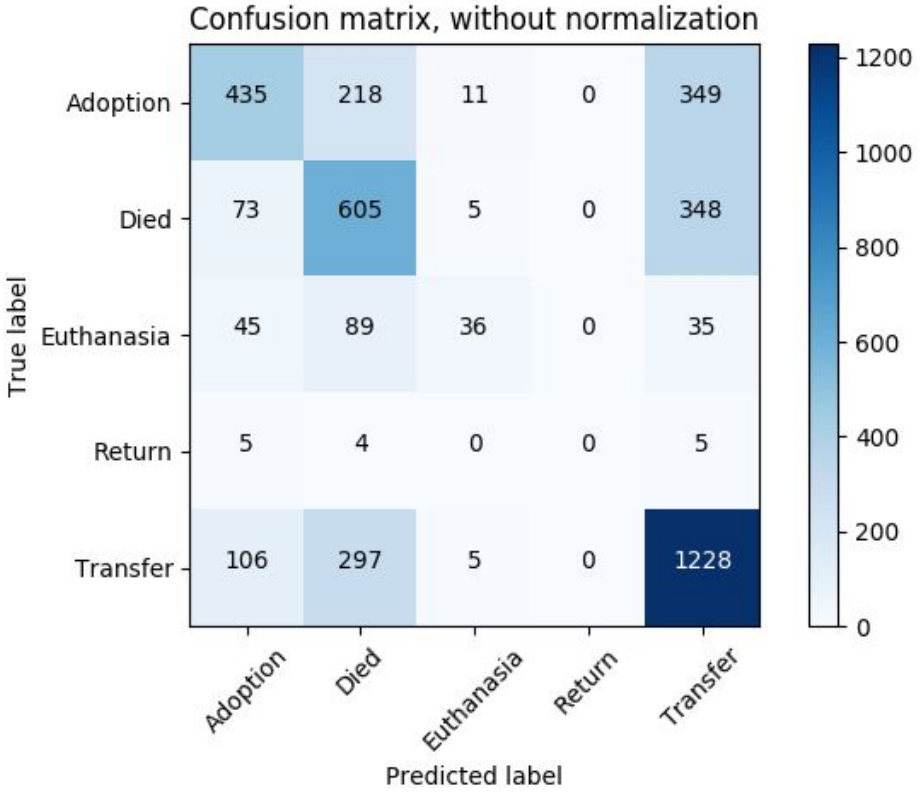
Ensemble Technique Exploration

- Stacking
- Bagging
- Blending
- Boosting
- Max Voting


Model1
Model2
Model3
Model4
Result
CatBoost
LDA
LogisticRegression
XBM
XGB
Bagging
Max voting
Stacking
Blending
Name
0.0
0.2
0.4
0.6
0.8
1.0


Name_frequency
Name_length
Day
Month
Breed1
Hour
Day_of_weekMale/Female
Color-medium
Breed7
Year
Breed2
Color-cold
Color-warm
f28
Color_feature2
Breed9
Breed8
Color-dark
Breed5
Color-light
Breed4
Breed10
Breed6
Color_feature1
Breed3
Age
Neutered/Intact
Named/un-named
feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
relative importance

Findings : Confusion Matrix



THANKS!

Any questions?