Group Project

Animal Adoptability Prediction

By

Dhaval Shah (802973719)

Pritesh Pimpale (803000983)

Dipika Mahashabde (803004274)

CPSC 483

Professor: Kenytt Avery

Department of Computer Science

California State University, Fullerton

Spring 2016

CONTENTS

1 INTRODUCTION

Kaggle Competition it is a platform where different users take part, analyze the data and perform different algorithms, the prediction produced by the algorithm is submitted on the online site and the scores are produced by the competition. The results have ranks and scores, the scores are important of all if you have a score near to 0 it's a really good prediction produced by your algorithm. They have rules and regulation before submitting the predicted outputs.

We had to choose a competition which was active on Kaggle. Therefore we looked at different competitions and found one which was on Shelter Animals (https://www.kaggle.com/c/shelter-animal-outcomes).

Shelter Animal outcomes, the data for this competition is taken from Austin Animal Center. Every year many animals are given up by owners or lost or even go through cruelty situation. And leaving this there are many animals which get euthanized. Therefore the Austin animal center wanted to predict the outcome if the animals will go through transfer to other shelter or return to owner or die of natural causes or get euthanized or get adopted.

In simple words we have classify if the animal will get

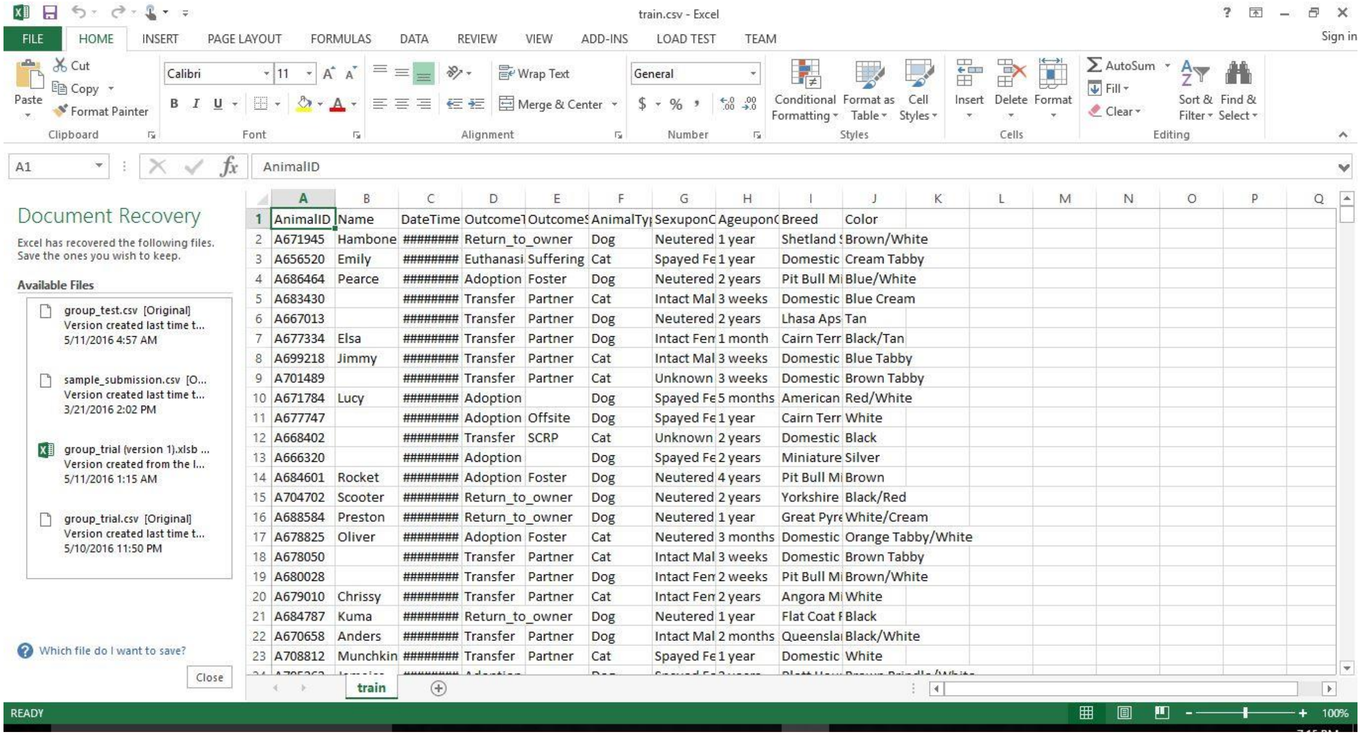
1. Transferred
2. Return to owner
3. Died
4. Adopted
5. Euthanized

Data in data files

1. **AnimalType** : Types of animals (Cat or Dog)
2. **Name** : Name of the animal(pets)
3. **DateTime** : Date and time at which the outcome was measured
4. **OutcomeType** : The result which we have to predict on test data
    a. Transferred
    b. Return to owner
    c. Died
    d. Adopted
    e. Euthanized
5. **OutcomeSubType** : the reason for the outcome mentioned in the above column
6. **SexuponOutcome** : Gender of the animal
    a. Intact Female
    b. Spayed Female
    c. Intact Male
    d. Neutered Male
    e. Unknown
7. **AgeuponOutcome** : age of the animal(in days/week/months/years)
8. **Breed** : Breed of particular animals

9. Color **:** Color of the animals(may one or two colors are given)

Data:



Document Recovery

Excel has recovered the following files. Save the ones you wish to keep.

Available Files

group_test.csv [Original]
Version created last time t...
5/11/2016 4:57 AM

sample_submission.csv [O...
Version created last time t...
3/21/2016 2:02 PM

group_trial (version 1).xlsb ...
Version created from the l...
5/11/2016 1:15 AM

group_trial.csv [Original]
Version created last time t...
5/10/2016 11:50 PM

Which file do I want to save?

Close

	A	B	C	D	E	F	G	H	I	J
1	AnimalID	Name	DateTime	OutcomeT	OutcomeS	AnimalTyp	SexuponO	AgeuponO	Breed	Color
2	A671945	Hambone	########	Return_to_owner		Dog	Neutered	1 year	Shetland S	Brown/White
3	A656520	Emily	########	Euthanasi	Suffering	Cat	Spayed Fe	1 year	Domestic	Cream Tabby
4	A686464	Pearce	########	Adoption	Foster	Dog	Neutered	2 years	Pit Bull Mi	Blue/White
5	A683430		########	Transfer	Partner	Cat	Intact Mal	3 weeks	Domestic	Blue Cream
6	A667013		########	Transfer	Partner	Dog	Neutered	2 years	Lhasa Aps	Tan
7	A677334	Elsa	########	Transfer	Partner	Dog	Intact Fem	1 month	Cairn Terr	Black/Tan
8	A699218	Jimmy	########	Transfer	Partner	Cat	Intact Mal	3 weeks	Domestic	Blue Tabby
9	A701489		########	Transfer	Partner	Cat	Unknown	3 weeks	Domestic	Brown Tabby
10	A671784	Lucy	########	Adoption		Dog	Spayed Fe	5 months	American	Red/White
11	A677747		########	Adoption	Offsite	Dog	Spayed Fe	1 year	Cairn Terr	White
12	A668402		########	Transfer	SCRP	Cat	Unknown	2 years	Domestic	Black
13	A666320		########	Adoption		Dog	Spayed Fe	2 years	Miniature	Silver
14	A684601	Rocket	########	Adoption	Foster	Dog	Neutered	4 years	Pit Bull Mi	Brown
15	A704702	Scooter	########	Return_to_owner		Dog	Neutered	2 years	Yorkshire	Black/Red
16	A688584	Preston	########	Return_to_owner		Dog	Neutered	1 year	Great Pyre	White/Cream
17	A678825	Oliver	########	Adoption	Foster	Cat	Neutered	3 months	Domestic	Orange Tabby/White
18	A678050		########	Transfer	Partner	Cat	Intact Mal	3 weeks	Domestic	Brown Tabby
19	A680028		########	Transfer	Partner	Dog	Intact Fem	2 weeks	Pit Bull Mi	Brown/White
20	A679010	Chrissy	########	Transfer	Partner	Cat	Intact Fem	2 years	Angora Mi	White
21	A684787	Kuma	########	Return_to_owner		Dog	Neutered	1 year	Flat Coat R	Black
22	A670658	Anders	########	Transfer	Partner	Dog	Intact Mal	2 months	Queenslai	Black/White
23	A708812	Munchkin	########	Transfer	Partner	Cat	Spayed Fe	1 year	Domestic	White

2 PREPROCESSING OF DATA

Preprocessing of data is necessary to run the data mining algorithms on the given data. As many of the columns have data which gives the multiple values of the attribute. e.g. Color column has more than one color for the mixed colored cats and dogs. So this type of data is necessary to be preprocessed before attempting any algorithm on it.

- **Name**
  Animals have some name given to them. This name can be a feature for the adoption of the pet by new owner. There are many animals who don't have names so we have kept it blank.
- **Animal Type**
  This attribute has value either 'cat' or 'dog'.
- **sex upon outcome**
  This attribute stores the sex of the pet animal. This attribute has values
  a. Intact Female
  b. Spayed Female
  c. Intact Male
  d. Neutered Male
  e. Unknown
- **Name Frequency**
  We have also calculated the frequency of all the names which were registered
- **Has Name**
  We also created a new data which stores the value 'Yes' if the name is present and 'No' if the name is not present. This parameter is much more significant in case of cats. For cats the adoption rate is much higher if a cat has name.
- **Date Time**
  Date Time attribute stores the time of the outcome of the train dataset. This attribute we have divided into many columns which are
  1. Year
  2. Month
  3. Day
  4. Hour
  5. Minutes
- **Age upon Outcome**
  The Age was in all different forms like days weeks months and years, which would make it difficult to analyze the data. So we convert the in format which is we convert the data into weeks.
- **Age Groups**
  Age is one of the factor for the decision of the owner to adopt a pet. So instead of keeping the data as it is we converted age into different age ranges. So we were able to group the age into ranges which provided an overview of the data.
- **Breed**
  We have different types of breed for each animals so we have checked if the animals (dog/cat) are of pure breed or mixture of breeds.

- **Mix Breed**
  The breed attribute has the value Mix if the if the breed is mixed. We extracted this value into new column to set if the breed is mix or not. If the animal is of pure breed we assign it ‘normal’ and if it is mixture we would assign ‘mix’.
- **Color**
  We have different colors for all the animals so to see if it has mixture of two colors or a single color. All values of the color are stored in the color column.
- **Mix color**
  We have added one more column to see if it has two colors or no. If multiple colors are there for animal we set the Mix color attribute to true and false if it is single color.

After processing the data we have converted the data into numeric form to use it in the algorithms.

e.g.

For Animal Type

- Dog = 0
- Cat = 1

3 Data Analysis

3.1 Animal type distribution

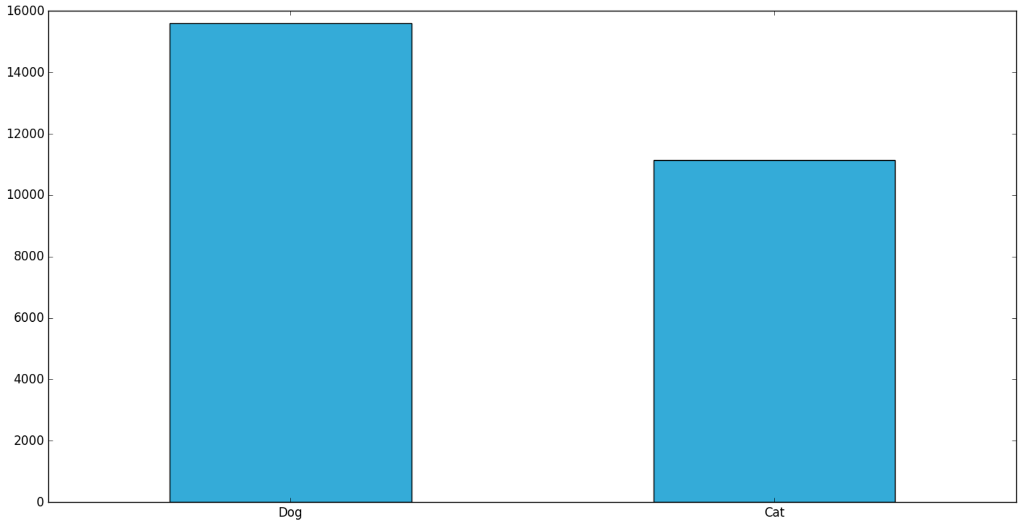


Fig. Cat and Dog

After seeing the graph we get to know that in the train data we have more dogs than cats.

3.2 Outcome Distribution

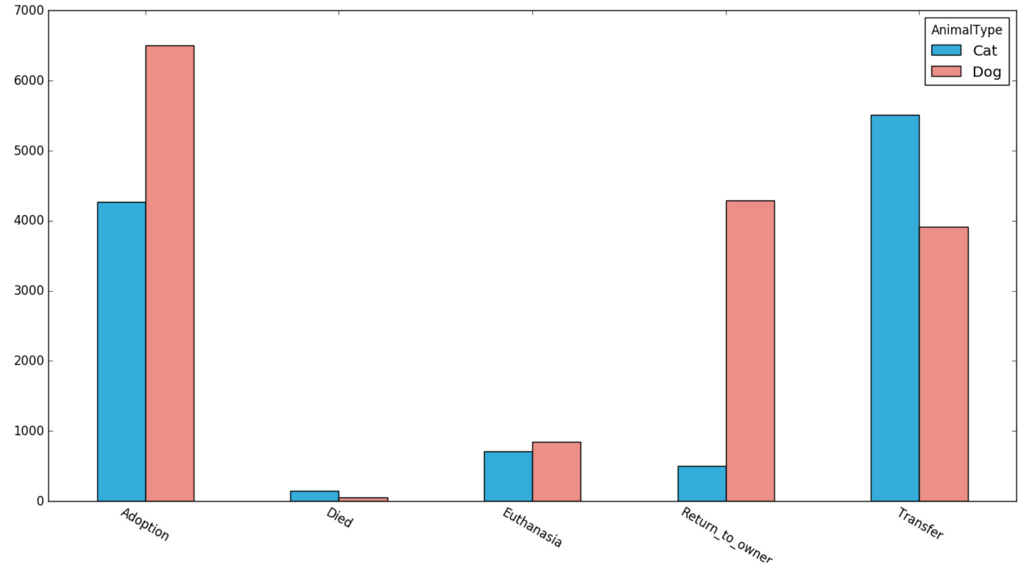


Fig. Animal type vs outcome

This shows that Adoption of animals is more than other outcome. Also dog has more probability of getting adopted or going return to owner than the cats.

3.3 Sex vs Animal Type

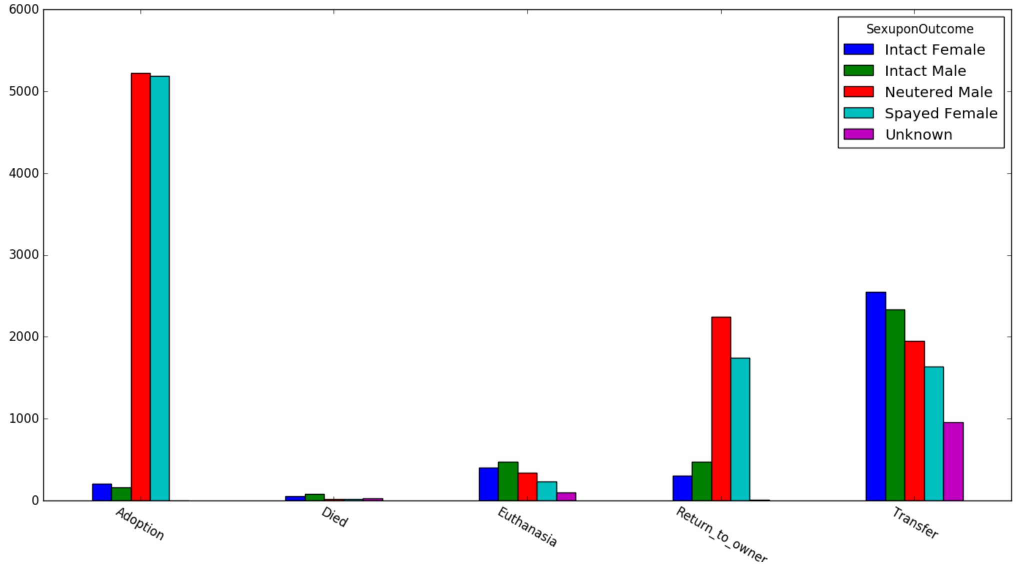


Fig. sex vs animal type

This data shows that adoption of neutered male and spayed female is more likely to happen than other types of the sex. Also same trend is seen for the return to owner outcome. For the transfers the sex doesn't seem to matter much.

- **Adoption:** Neutered Male and Spayed Female are most of time adopted.
- **Died:** sex doesn't matter much for the outcome of the pet.
- **Euthanasia:** intact female and intact male are more euthanized more than neutered or spayed. But still this difference is not much and doesn't provide any significance.
- **Return_to_owner:** Neutered male and spayed female are returned more than others.
- **Transfer:** Intact male and female is more than others. But data doesn't show much significant difference between them.

3.4 AGE GROUPS VS OUTCOME TYPE

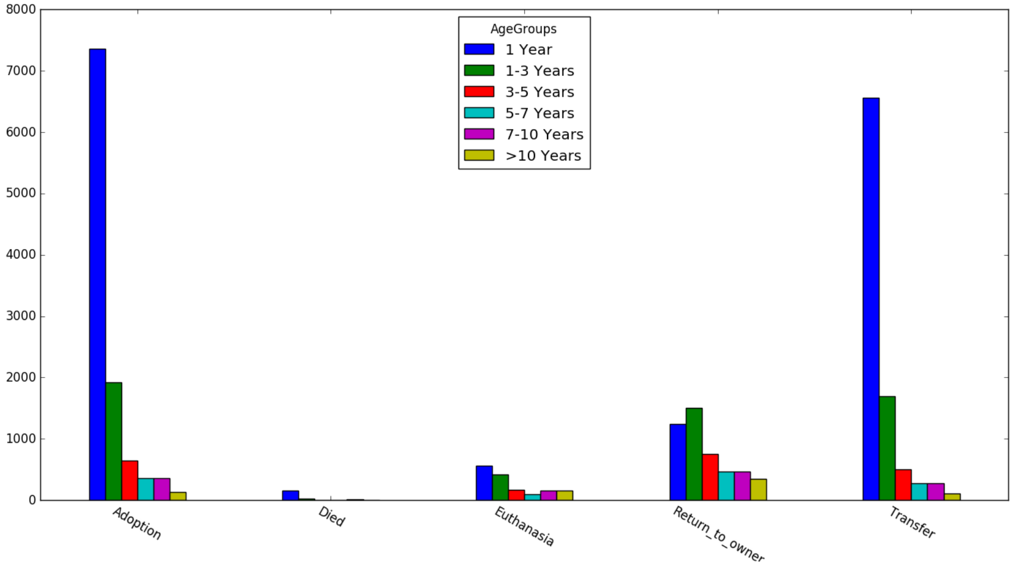


Fig. age group vs outcome Type

Adoption and transfer of the animals is more for the animals aged less than 1 year than for the animals who are old.

3.5 MIX COLOR VS OUTCOME TYPE

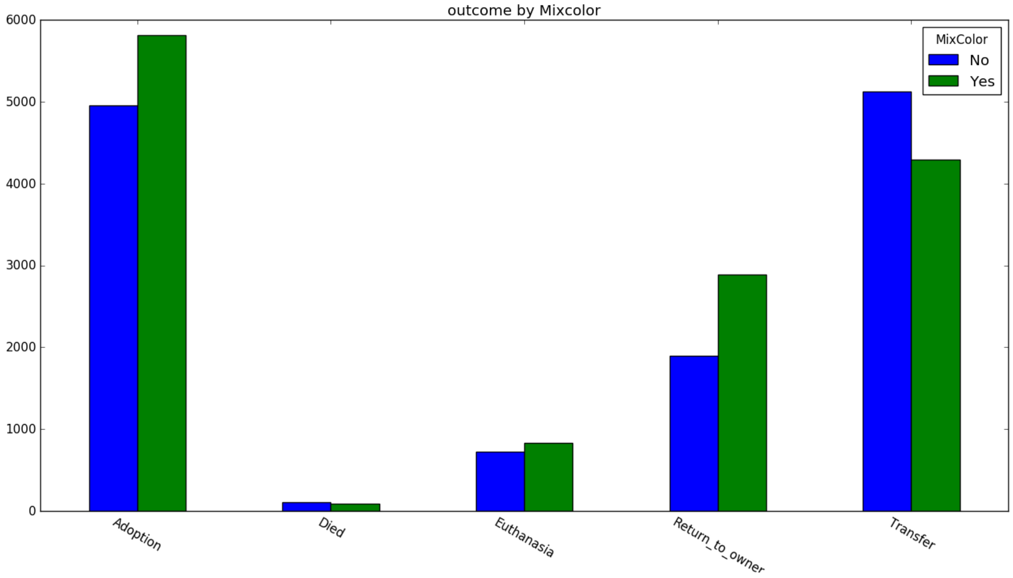


Fig. Mix Color vs Outcome Type

This graph tells us that how many animals have mix color and how many have only one color. This graph doesn’t give any significant information as for all the outcome data is pretty similar.

3.6 HAS NAME VS OUTCOME TYPE

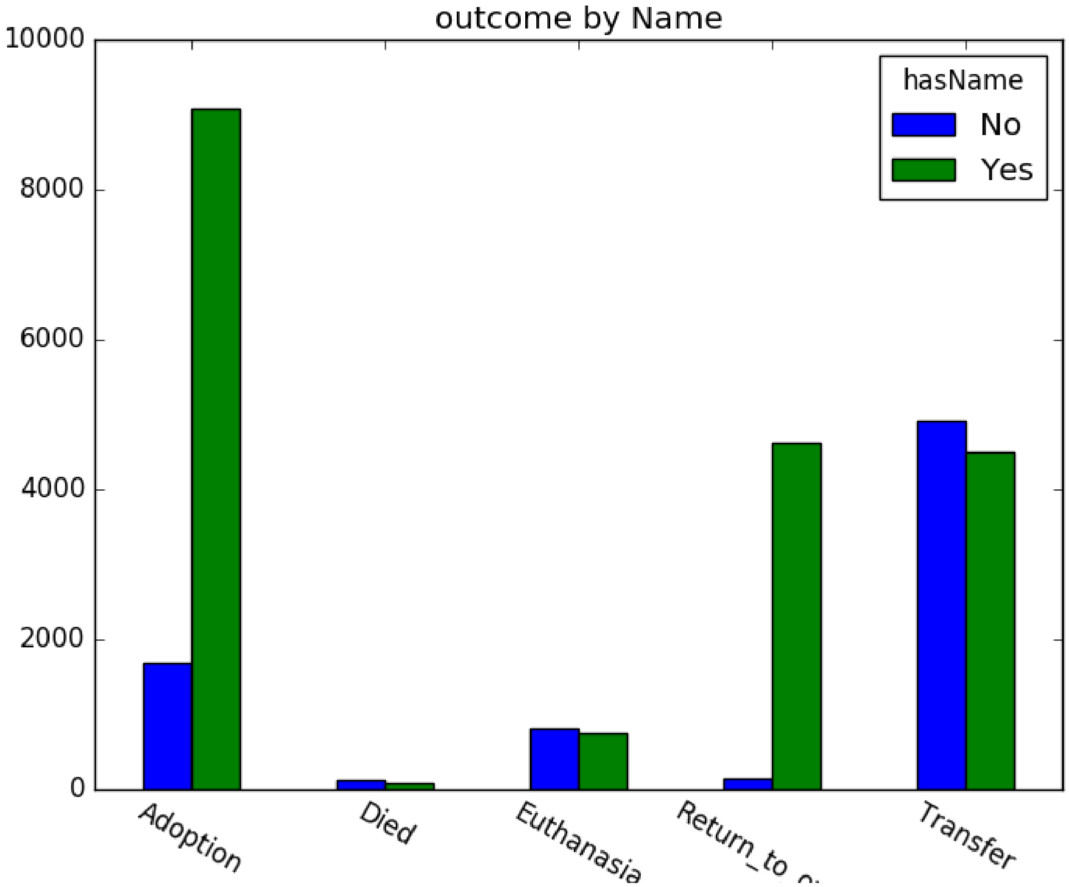


Fig. has Name vs outcome type

This graph shows that having a name to a pet animal is significant factor for animal getting adopted or returning to owner. While having name doesn't matter much for rest of the outcomes.

3.7 MIX BREED VS OUTCOME TYPE

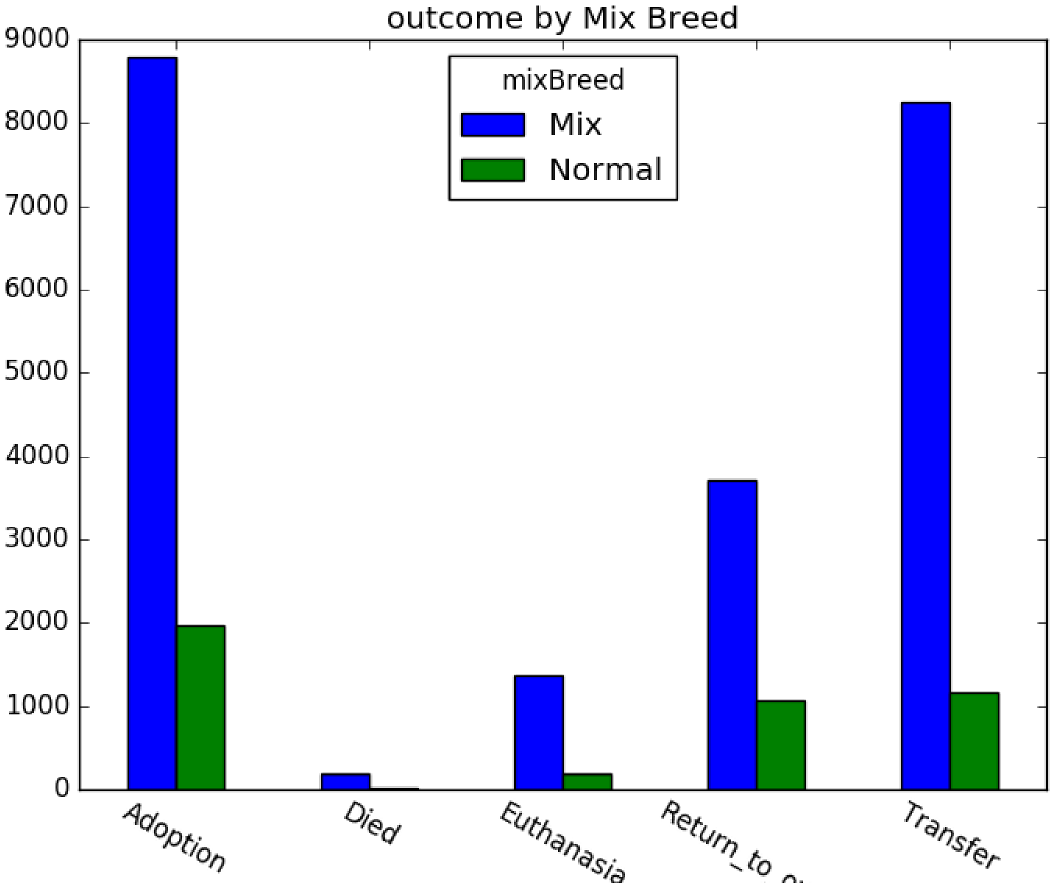


Fig. Mix Breed vs Outcome Type

Mix breed is seen more in all the outcomes. This shows that the pet animals with mix color are more in the dataset, so this also doesn’t provide much information for predicting the outcome.

4 RESULTS OBTAINED FROM TEST AND TRAINING DATA

- **Naïve Bayes**

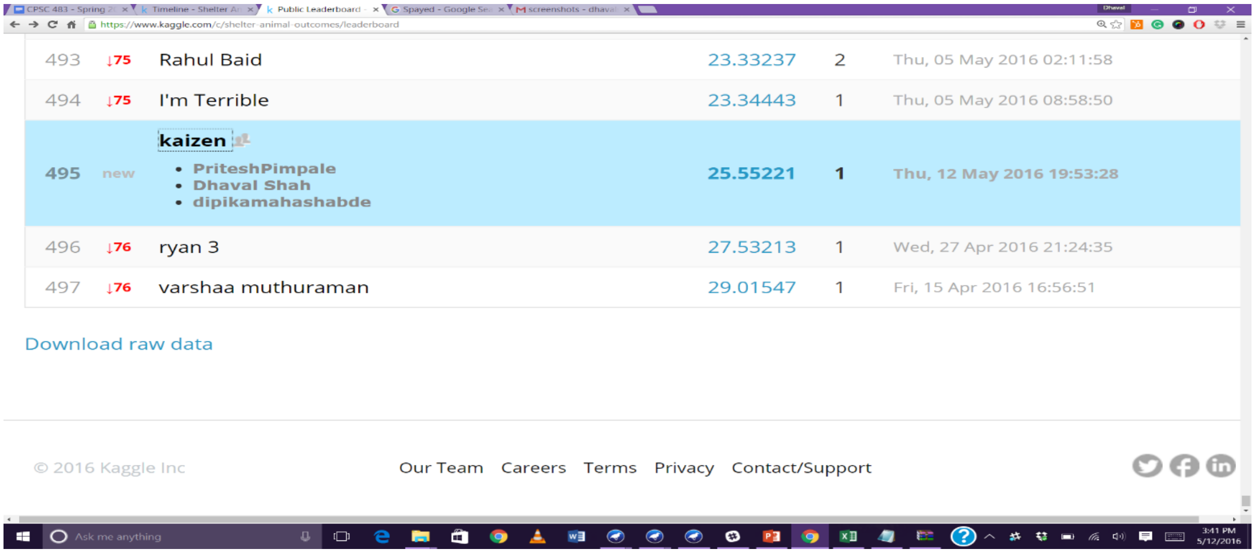
  Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness and diameter features.
  We have implemented naïve Bayes algorithm. We used Naive Bayes from SK-learn (scikit learn, n.d.) Library.
  Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features.
  Gaussian NB implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma^2_y}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma^2_y}\right)$$

The parameters σ_y and μ_y are estimated using maximum likelihood.
We have train the data on the outcome type and then predicted whether the outcome type is Adoption, Transferred, Return to owner, Died, Adopted, and Euthanized.

Our initial approach was to use this classifier for the prediction. The initial results were not that good for the prediction as there were no extra parameters considered for the prediction.

Kaggle score is as follow.



493	↓75	Rahul Baid	23.33237	2	Thu, 05 May 2016 02:11:58
494	↓75	I'm Terrible	23.34443	1	Thu, 05 May 2016 08:58:50
495	new	**kaizen** • PriteshPimpale • Dhaval Shah • dipikamahashabde	25.55221	1	Thu, 12 May 2016 19:53:28
496	↓76	ryan 3	27.53213	1	Wed, 27 Apr 2016 21:24:35
497	↓76	varshaa muthuraman	29.01547	1	Fri, 15 Apr 2016 16:56:51

Download raw data

- **Random Forest**

  Many Classification trees are grown by using random forest (scikit-learn, n.d.). An input vector is given to predict its class. This input vector passes through all tress which are produced. Each tree gives a result that is, which class it belongs to. Votes are considered from all the trees to decide the final outcome of the classifier. The class with most vote is assign to the input given.

  Growing of the trees

  There are ‘N’ number of observation in the training set. Every tree selects observations at random by replacing the actual training data. These new data generated are used for the growing of the tree. There are ‘M’ number of attributes, they are also selected at random, and the best split on these attributes (selected) is used to split the node. The value of the M can be given while running the algorithm. While growing the tree the selected attribute won’t change till the tree is grown to the output leaves. (random forest, n.d.)

  Our next approach was to use random forest classifier for the prediction.

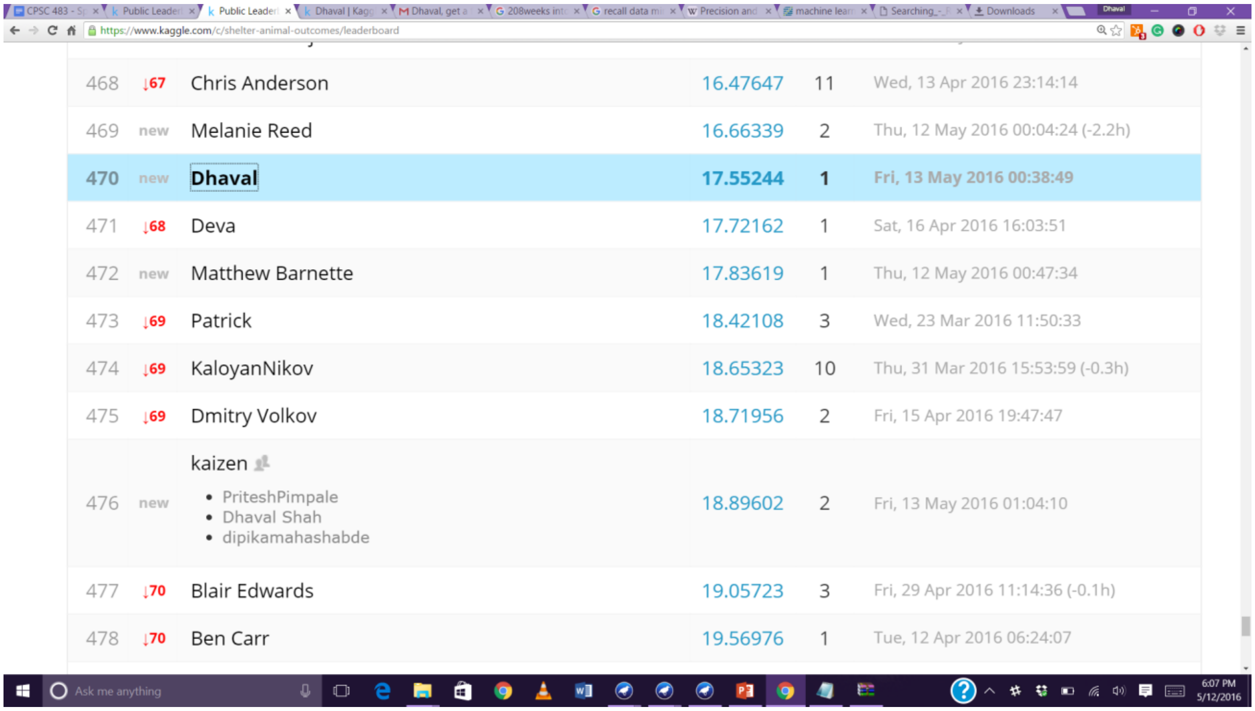
  We tried with the original parameters and following configurations for the random forest classifier.

  ○ No. of trees = 10

  ○ max features = 5

  ○ max depth = 10

  Initial result with this parameters is as below



468	↓67	Chris Anderson	16.47647	11	Wed, 13 Apr 2016 23:14:14
469	new	Melanie Reed	16.66339	2	Thu, 12 May 2016 00:04:24 (-2.2h)
470	new	Dhaval	17.55244	1	Fri, 13 May 2016 00:38:49
471	↓68	Deva	17.72162	1	Sat, 16 Apr 2016 16:03:51
472	new	Matthew Barnette	17.83619	1	Thu, 12 May 2016 00:47:34
473	↓69	Patrick	18.42108	3	Wed, 23 Mar 2016 11:50:33
474	↓69	KaloyanNikov	18.65323	10	Thu, 31 Mar 2016 15:53:59 (-0.3h)
475	↓69	Dmitry Volkov	18.71956	2	Fri, 15 Apr 2016 19:47:47
476	new	kaizen • PriteshPimpale • Dhaval Shah • dipikamahashabde	18.89602	2	Fri, 13 May 2016 01:04:10
477	↓70	Blair Edwards	19.05723	3	Fri, 29 Apr 2016 11:14:36 (-0.1h)
478	↓70	Ben Carr	19.56976	1	Tue, 12 Apr 2016 06:24:07

Random forest algorithm can give better predictions if number of trees are set to a larger values.

Also the new derived parameters are used while training the random forest classifier.

Parameters used for the final submissions are

1. Name
2. NameFreq
3. hasName
4. AnimalType
5. SexuponOutcome
6. AgeuponOutcome
7. AgeGroups
8. Breed
9. mixBreed
10. Color
11. MixColor
12. firstColor
13. Year
14. Month
15. Day
16. Hour
17. Minute

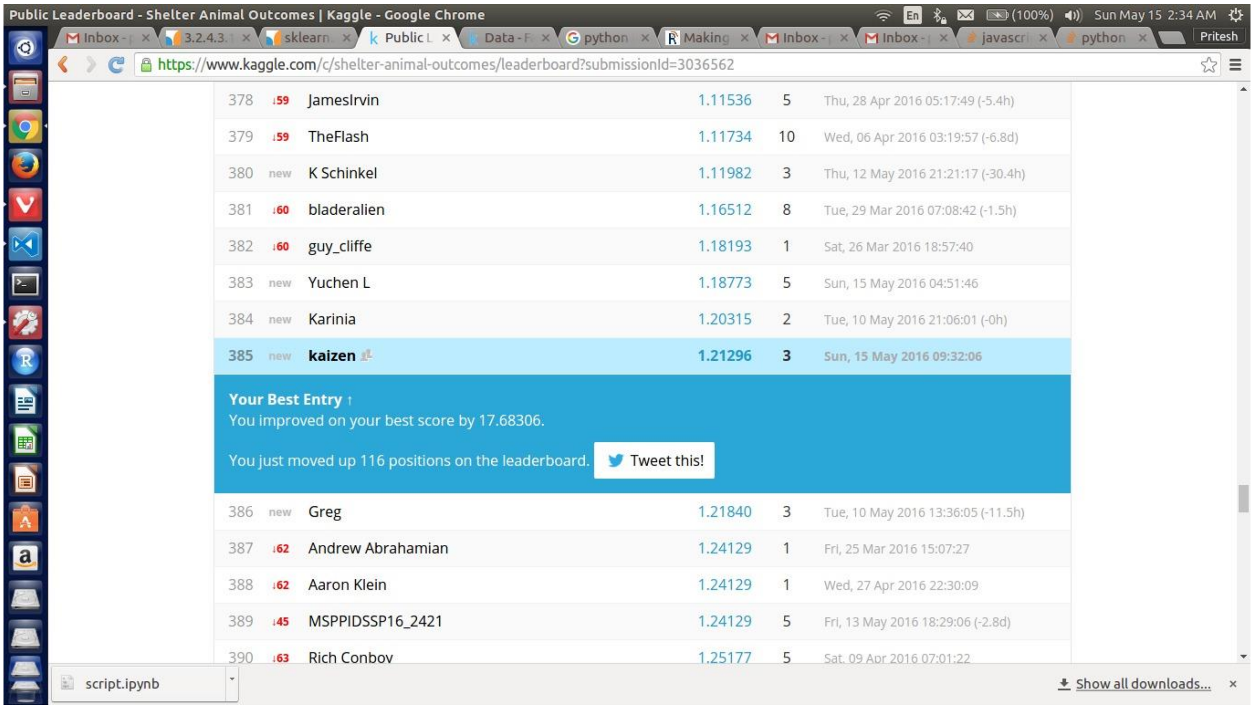
For random forest classifier there is no need to do the feature selection as it is best trained with all the features provided. Each tree generated uses only ‘n’ number of features specified while running it from all features. It also trains the generated trees on the result of the earlier trees. This increases the prediction accuracy of the trees in the classifier.

We tried with multiple values for the number of trees and number of features. The final submission was the one with minimum logloss obtained from the training data.

Final parameters used for the prediction.

- No. of trees = 1000
- max features = 10
- max depth = 10

Final result from Kaggle


Public Leaderboard - Shelter Animal Outcomes | Kaggle - Google Chrome
En
(100%)
Sun May 15 2:34 AM
Inbox -
3.2.4.3.
sklearn
Public L
Data - F
python
Making
Inbox -
Inbox -
javascri
python
Pritesh
https://www.kaggle.com/c/shelter-animal-outcomes/leaderboard?submissionId=3036562
script.ipynb
Show all downloads...

378	↓59	JamesIrvin	1.11536	5	Thu, 28 Apr 2016 05:17:49 (-5.4h)
379	↓59	TheFlash	1.11734	10	Wed, 06 Apr 2016 03:19:57 (-6.8d)
380	new	K Schinkel	1.11982	3	Thu, 12 May 2016 21:21:17 (-30.4h)
381	↓60	bladeralien	1.16512	8	Tue, 29 Mar 2016 07:08:42 (-1.5h)
382	↓60	guy_cliffe	1.18193	1	Sat, 26 Mar 2016 18:57:40
383	new	Yuchen L	1.18773	5	Sun, 15 May 2016 04:51:46
384	new	Karinia	1.20315	2	Tue, 10 May 2016 21:06:01 (-0h)
385	new	**kaizen**	**1.21296**	**3**	**Sun, 15 May 2016 09:32:06**

Your Best Entry ↑
You improved on your best score by 17.68306.

You just moved up 116 positions on the leaderboard. Tweet this!

386	new	Greg	1.21840	3	Tue, 10 May 2016 13:36:05 (-11.5h)
387	↓62	Andrew Abrahamian	1.24129	1	Fri, 25 Mar 2016 15:07:27
388	↓62	Aaron Klein	1.24129	1	Wed, 27 Apr 2016 22:30:09
389	↓45	MSPPIDSSP16_2421	1.24129	5	Fri, 13 May 2016 18:29:06 (-2.8d)
390	↓63	Rich Conboy	1.25177	5	Sat, 09 Apr 2016 07:01:22

5 ENVIRONMENT

Environment: - Python (install the new version)

Libraries: -

1. Pandas: - It offers data structures and operations for manipulating numerical tables and time series
2. Scipy: - It helps in scientific and technical computing.
3. Sklearn: - They have all algorithm with their functions.
4. Numpy: - It helps in mathematical computation.
5. Matplotlib: - This library is used to plot graphs in python, it has inbuilt functions for it.

Tool: - Visual Studio Code.

1. **Installation of Python and it libraries**

   Step 1.

   Download python: - **https://www.python.org/downloads/**

   Step 2.

   Install all the libraries needed

   1. Install pip(This will help us install other libraries)
   2. Install scipy (pip install scipy)
   3. Install numpy (pip install numpy)
   4. Install sklearn (pip install sklearn)
   5. Install pandas (pip install pandas)
   6. Install matplotlib (pip install matplotlib)

2. **Download the data from Kaggle**

   Data: - https://www.kaggle.com/c/shelter-animal-outcomes/data (Kaggle , n.d.)

   - Download all the files on this link.
   - Train.csv: - this data will help us to model the data using the algorithm.
   - Test.csv: - the modeled algorithm will help us to predict on the test data.
   - Sample_Submission.csv: - the predicted values need to be stored in the format shown in this file.

6 REFERENCES

Kaggle . (n.d.). Retrieved from Kaggle : https://www.kaggle.com/c/shelter-animal-outcomes/data

random forest. (n.d.). Retrieved from random forest: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#intro

scikit learn. (n.d.). Retrieved from scikit learn: http://scikit-learn.org/stable/modules/naive_bayes.html

scikit-learn. (n.d.). Retrieved from scikit-learn: http://scikit-learn.org/stable/modules/ensemble.html#forest