Data Preparation

In its raw form, or data has some fields that are not ideal for use in machine learning models. Some have too much information, some not enough, some the wrong type. In order for us to create the best models possible we have to do several transformations to our data to get it into a format that is best suited for our purposes. Below is the raw data from kaggle (apologies for it being squished):

|  | **AnimalID** | **Name** | **DateTime** | **OutcomeType** | **OutcomeSubtype** | **AnimalType** | **SexuponOutcome** | **AgeuponOutcome** | **Breed** | **Color** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | A671945 | Hambone | 2014-02-12 18:22:00 | Return\_to\_owner | NaN | Dog | Neutered Male | 1 year | Shetland Sheepdog Mix | Brown/White |
| **1** | A656520 | Emily | 2013-10-13 12:44:00 | Euthanasia | Suffering | Cat | Spayed Female | 1 year | Domestic Shorthair Mix | Cream Tabby |
| **2** | A686464 | Pearce | 2015-01-31 12:28:00 | Adoption | Foster | Dog | Neutered Male | 2 years | Pit Bull Mix | Blue/White |
| **3** | A683430 | NaN | 2014-07-11 19:09:00 | Transfer | Partner | Cat | Intact Male | 3 weeks | Domestic Shorthair Mix | Blue Cream |
| **4** | A667013 | NaN | 2013-11-15 12:52:00 | Transfer | Partner | Dog | Neutered Male | 2 years | Lhasa Apso/Miniature Poodle | Tan |

This data set does not have many features, so we need to make sure we maximize the information in the features we have. Going through the columns, we can see certain problems listed below:

AnimalID: not relevant

OutcomeSubtype: really an extension of OutcomeType; for instance adoption can be a foster adoption or an offsite adoption

SexuponOutcome: contains animal sex as well as whether it has been spayed/neutered. So there is actually two bits of information here.

AgeuponOutcome: this is a string, should be a float or some sort of standardized number

Breed: contains info on whether the animal is a mix or not

In order to get the data into the format we want, we do a few different steps. We simply combine OutcomeSubtype with OutcomeType to create one string from two to create our final output target variable.

From SexuponOutcome, we create two new fields, one called Female and one called Intact which are both binary variables.

For AgeuponOutcome, we do something a bit different. We convert everything into weeks based on the strings it contains. We use a standard conversion of one year = 52 weeks and 1 month = 4.5 weeks. So if the original value is 3 years, it gets transformed to 156 and if the original value is 4 months it gets transformed to 18. This gives us a continuous range which will help with our models.

We add another column based on breed called MixBreed which is a binary variable if the animal is a mix or not.

Finally, we delete all the rows that have duplicate or useless information, such as animalID and SexuponOutcome.

As with most data we will come across, this one had data that was either missing or had literal values of ‘Unknown’. In general we dealt with this data by replacing it with blank strings.