Shelter Animals Outcomes

Abril / 2016

Sola Comino, Alberto
albsolac@gmail.com

El problema

- Predecir el futuro de los animales en el refugio Austin Animal Center.
- Clasificación multiclase.
- Conjunto de datos:
  - Alimento
  - Color
  - Edad
  - Etc.

Herramientas utilizadas

- Lenguaje: R.
- Entorno: RStudio.
- Paquetes:
  - Entrada / Salida (csv): gdata.
  - Fecha: lubridate.
  - Visualización: ggplot2.
  - Clasificación: xgboost, gbm y randomForest.

Medida de error. *Logarithmic Loss*

- Es el logaritmo de la función de probabilidad de la distribución aleatoria de Bernoulli.
- Esta medida de error es utilizada cuando hay que predecir un suceso con una determinada probabilidad.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

- Penaliza a los clasificadores que estén muy seguros en su decisión.
- El mismo concepto se utiliza para clasificación multiclase.

Visualización

Visualización

	AnimalID	Name	DateTime	OutcomeType	OutcomeSubtype	AnimalType	SexuponOutcome	AgeuponOutcome	Breed	Color
1	A671945	Hambone	2014-02-12 18:22:00	Return_to_owner		Dog	Neutered Male	1 year	Shetland Sheepdog Mix	Brown/White
2	A656520	Emily	2013-10-13 12:44:00	Euthanasia	Suffering	Cat	Spayed Female	1 year	Domestic Shorthair Mix	Cream Tabby
3	A686464	Pearce	2015-01-31 12:28:00	Adoption	Foster	Dog	Neutered Male	2 years	Pit Bull Mix	Blue/White
4	A683430		2014-07-11 19:09:00	Transfer	Partner	Cat	Intact Male	3 weeks	Domestic Shorthair Mix	Blue Cream
5	A667013		2013-11-15 12:52:00	Transfer	Partner	Dog	Neutered Male	2 years	Lhasa Apso/Miniature Poodle	Tan
6	A677334	Elsa	2014-04-25 13:04:00	Transfer	Partner	Dog	Intact Female	1 month	Cairn Terrier/Chihuahua Shorthair	Black/Tan
7	A699218	Jimmy	2015-03-28 13:11:00	Transfer	Partner	Cat	Intact Male	3 weeks	Domestic Shorthair Mix	Blue Tabby
8	A701489		2015-04-30 17:02:00	Transfer	Partner	Cat	Unknown	3 weeks	Domestic Shorthair Mix	Brown Tabby
9	A671784	Lucy	2014-02-04 17:17:00	Adoption		Dog	Spayed Female	5 months	American Pit Bull Terrier Mix	Red/White
10	A677747		2014-05-03 07:48:00	Adoption	Offsite	Dog	Spayed Female	1 year	Cairn Terrier	White

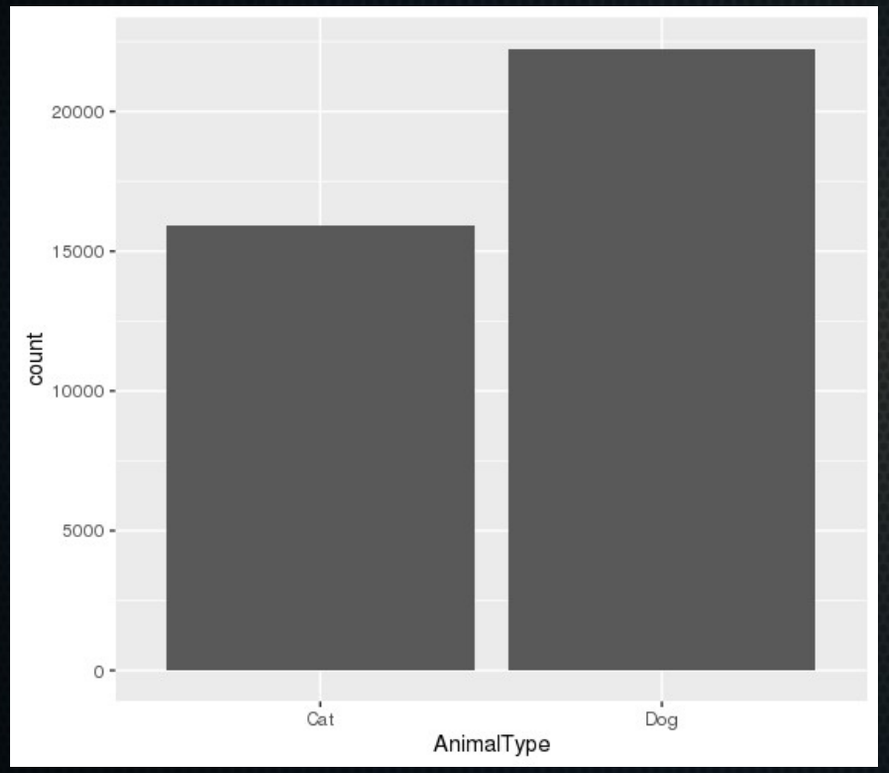
train.csv

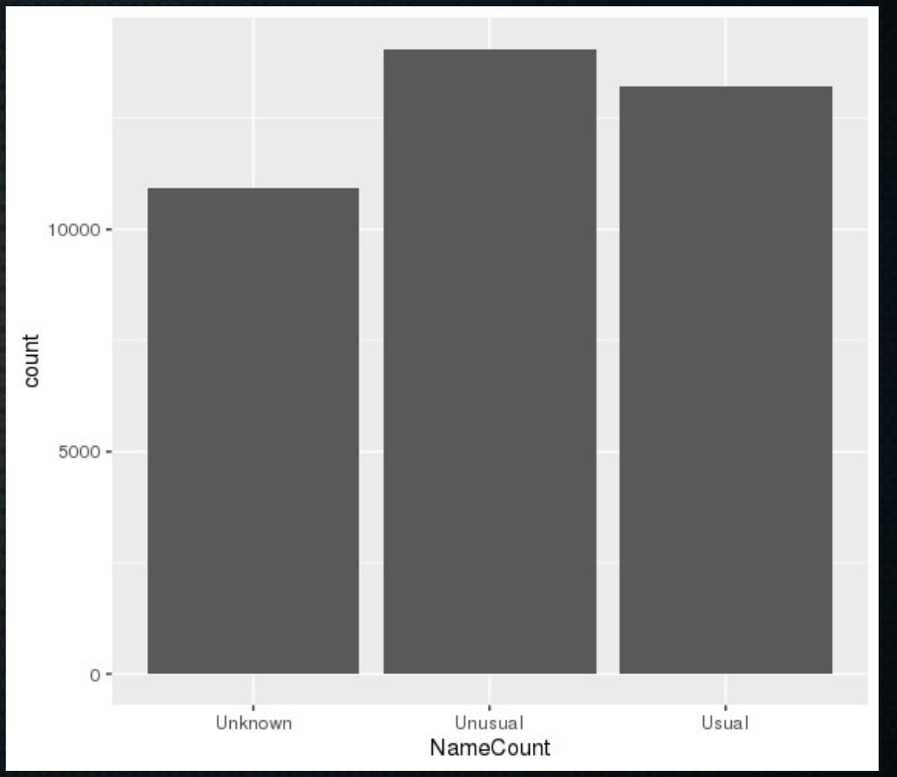
Visualización

ID	Name	DateTime	AnimalType	SexuponOutcome	AgeuponOutcome	Breed	Color
1	Summer	2015-10-12 12:15:00	Dog	Intact Female	10 months	Labrador Retriever Mix	Red/White
2	Cheyenne	2014-07-26 17:59:00	Dog	Spayed Female	2 years	German Shepherd/Siberian Husky	Black/Tan
3	Gus	2016-01-13 12:20:00	Cat	Neutered Male	1 year	Domestic Shorthair Mix	Brown Tabby
4	Pongo	2013-12-28 18:12:00	Dog	Intact Male	4 months	Collie Smooth Mix	Tricolor
5	Skooter	2015-09-24 17:59:00	Dog	Neutered Male	2 years	Miniature Poodle Mix	White
6	Beau	2015-06-23 11:17:00	Dog	Neutered Male	3 years	Beagle Mix	Brown/White
7	Bobo	2014-03-12 09:45:00	Cat	Neutered Male	13 years	Domestic Medium Hair Mix	Brown Tabby/White
8	Abby	2014-06-25 08:27:00	Cat	Spayed Female	6 months	Domestic Shorthair Mix	Brown Tabby
9	Ruby Grace	2014-11-12 18:05:00	Dog	Spayed Female	3 months	Cairn Terrier	Black/Cream
10	Ruby	2014-04-07 17:41:00	Dog	Spayed Female	1 year	Pit Bull Mix	Brown/White

test.csv

Visualización





Primer intento

- Poco familiarizado con R.
- No sabía ni qué clasificador ni cómo utilizarlo.
- Clasificador elegido: XGBoost.
- Datos sin preprocesar.

Puntuación: 10.06698
Posición: 200 / 230

Algo de preprocesamiento

- Se combinan 'train.csv' y 'test.csv'.
- Preprocesado: se añaden nuevos atributos.
  - TieneNombre, TipoComida
  - Año, Mes
  - Sexo, Estado
  - Edad, Color

Puntuación: 5.83736
Posición: +2

Cross Validation

- Mismo clasificador (XGBoost).
- Mismo preprocesado.
- *Folds*: 5/10
- Permite obtener el número de iteraciones con menor error.

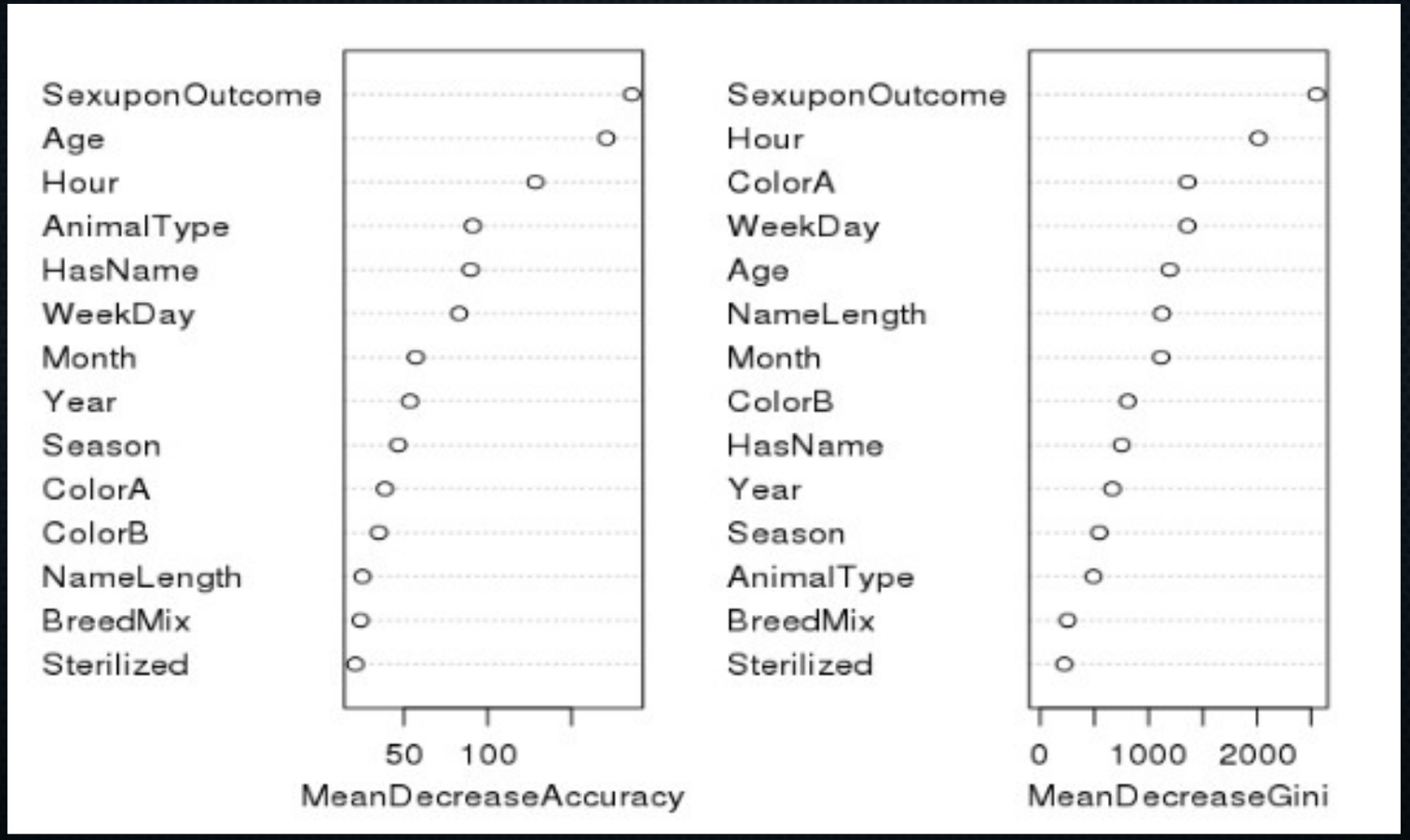
Puntuación: 1.48825
Posición: 140

Más preprocesamiento

- Valores perdidos de edad completados con la media.
- Otros valores perdidos como 'Desconocido'.
- Hora y día de la semana.
- Longitud del nombre.
- Alimento separado en dos columnas.
- Tamaño del animal (en función del pienso).

Importancia de características

(Random Forest)



Gradient Boosting

- Paquete 'gbm'.
- Mejores resultados con un sólo clasificador.
- Parece que se adapta mejor que *XGBoost* o *Random Forest*.

Puntuación: 0.79814
Posición: 80

Ensemble

- Clasificador utilizado: XGBoost.
- Entrenar varios clasificadores.
- *Cross Validation* para obtener el número de iteraciones.
- Unificar resultados.

Puntuación: 0.74968
Posición: 54

Posibles mejoras

- Mejor preprocesado de datos.
- Seleccionar características más importantes.
- Realizar imputación de valores perdidos.
- Realizar *ensemble* con diferentes tipos de clasificadores.

¿ Preguntas ?