Comprensión del negocio:

Se desea analizar los datos de ingresos y salidas del Refugio de Animales de la ciudad de Austin, para comprender las tendencias de adopción de animales, incluyendo qué atributos de estos animales resultan en una probabilidad de adopción mayor. El objetivo final es predecir si un cierto animal será adoptado o no, basado en las características que el Refugio de Animales de Austin puede identificar en el momento del ingreso. Este problema es importante porque aproximadamente 6.5 millones de animales ingresan a los refugios anualmente. Más aún, cada año 1.5 millones de animales de los refugios son sacrificados.

Comprensión de los datos:

Tenemos dos datasets uno de datos de los animales cuando entran y otro cuando salen. Los datos son los siguientes:

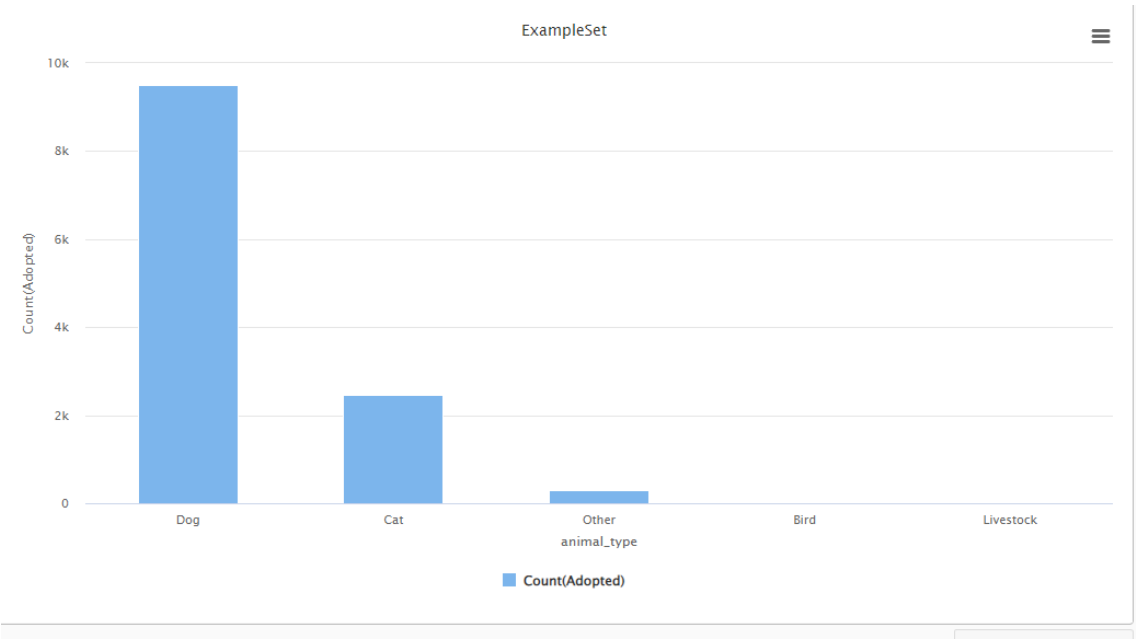
Dataset intakes:

Atributo	**Tipo de datos original de la base de datos**
animal_id	Texto
name	Texto
datetime	datetime
monthyear	datetime
found_location	Texto
intake_type	Texto
intake_condition	Texto
animal_type	Texto
sex_upon_intake	Texto
age_upon_intake	Texto
breed	Texto
color	Texto

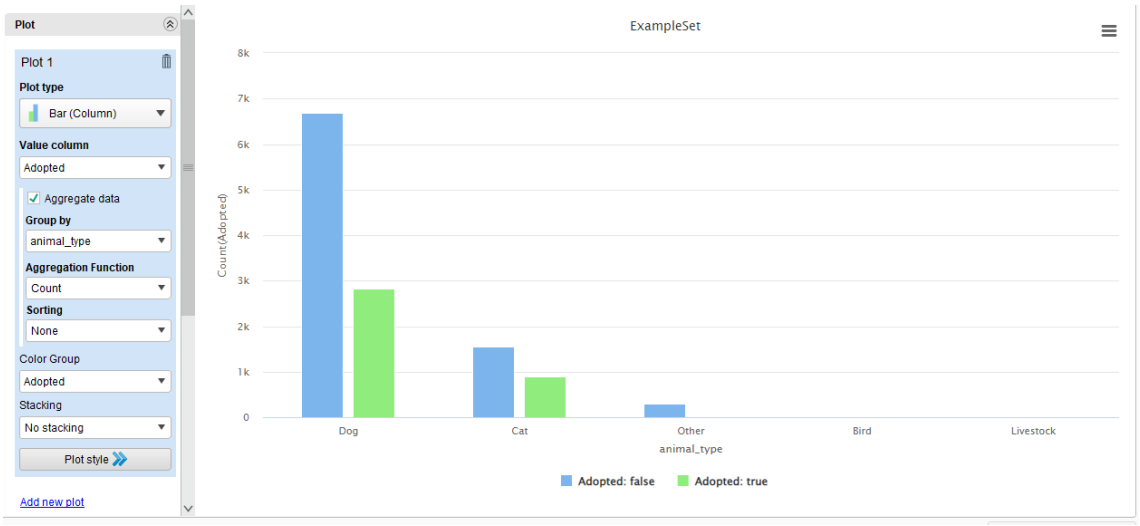
Dataset outake:

Atributo	**Tipo de datos original de la base de datos**
animal_id	Texto
name	Texto

Atributo	Tipo de datos original de la base de datos
datetime	datetime
monthyear	datetime
date_of_birth	Texto
outcome_type	Texto
outcome_subtype	Texto
animal_type	Texto
sex_upon_outcome	Texto
age_upon_outcome	Texto
breed	Texto
color	Texto



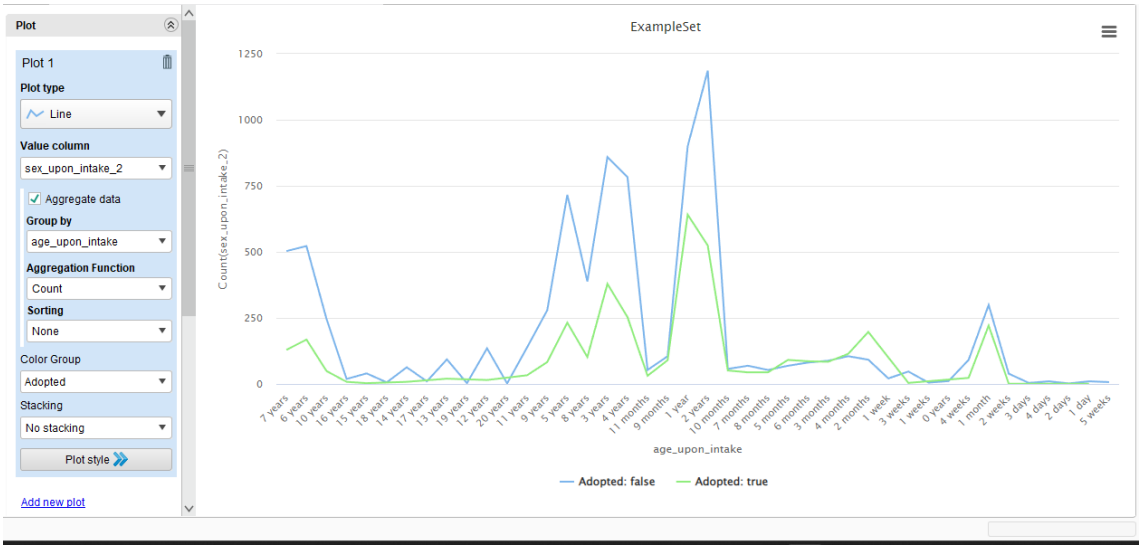
Se adoptan principalmente perros. Lo cual es normal puesto que el refugio recibe principalmente perros.



Total, Perros Adoptados 2823/9509

Total, Gatos Adoptados 904/2461

Lo cual nos da un animal de preferencia son los gatos.



Podemos observar también que la principal edad en la que se adopta es al año con 641 adopciones.

Preparación de los Datos:

Las decisiones que se tomaran para trabajar sobre los mismos son:

Trabajar sobre los dataset:

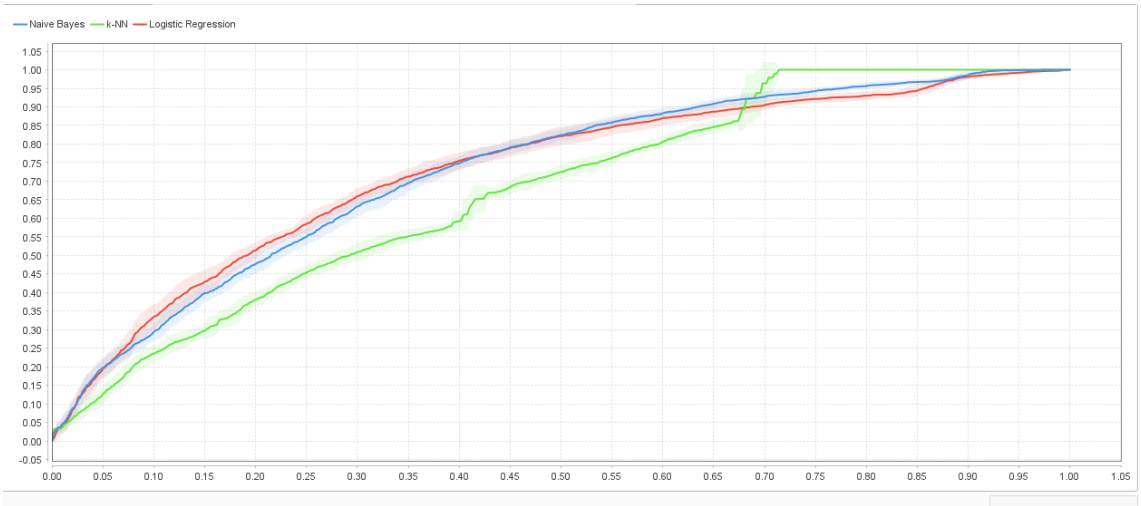
- Converitr sex upon intaken/outcome a sex y castrated por separado
- Definir atributo adoptado como label
- Sacar monthyear que es la mismo que date
- Eliminar atributos inecesarios
- Asignar role id a animal_id
- Eliminar duplicados
- Generar clase Adopted donde [outcome_type]=="Adoption"&&[outcome_subtype]!="Foster"

Todo esto se hace con turboprep sin ningún operador en específico.

Las columnas se remueven en transform remove. Las seleccionadas fueron principalmente las que no aportaban al modelo

Modelo:

Comparación de ROCs:

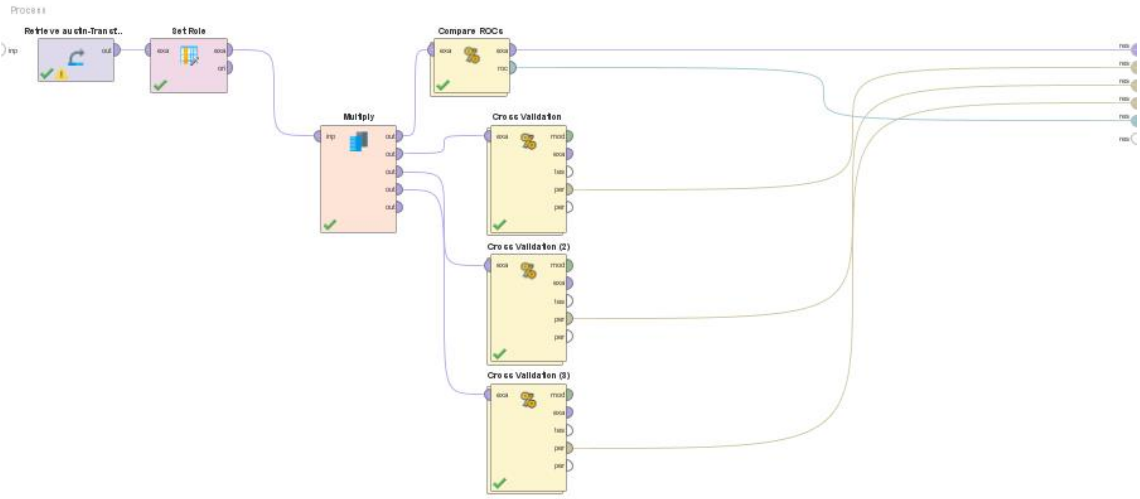


Se utilizarán tres modelos de clasificación:

- RL
- KNN
- NB

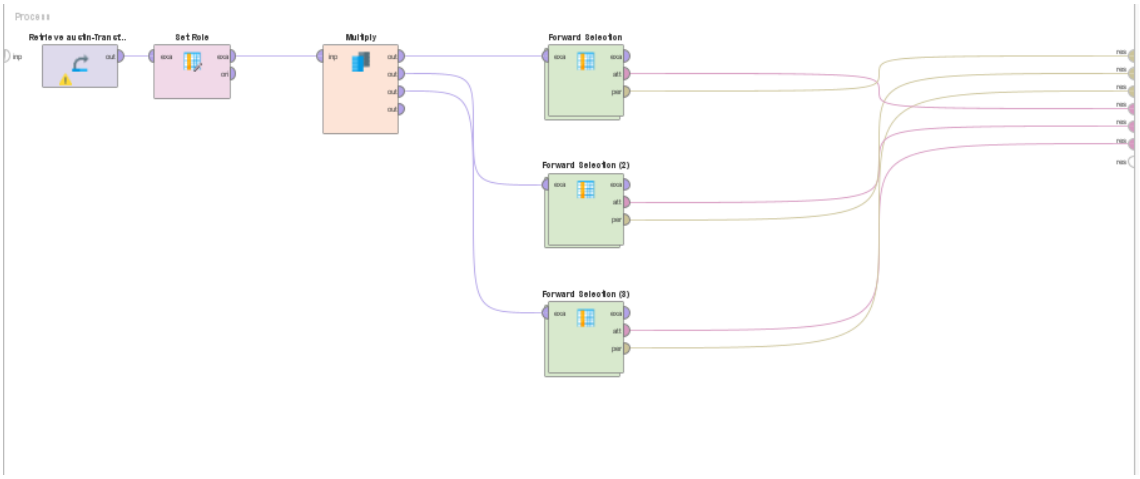
LDA se lo va a tratar como separado y se evaluara de manera particular puesto que debido a la característica de este hay que realizar unos cambios puesto que no soporta valores categóricos. (Se hará una conversión dummy)

El modelo para utilizar es el siguiente:



Cada cross validation posee un modelo.

También se realizará feature selection para cada modelo.



Evaluación:

RL

accuracy: 72.51% +/- 1.03% (micro average: 72.51%)

	true false	true true	class precision
pred. false	7551	2383	76.01%
pred. true	991	1347	57.61%
class recall	88.40%	36.11%	

- Foward selection:

accuracy: 72.75% +/- 0.82% (micro average: 72.75%)

	true false	true true	class precision
pred. false	7576	2378	76.11%
pred. true	966	1352	58.33%
class recall	88.69%	36.25%	

NB

accuracy: 69.90% +/- 0.57% (micro average: 69.90%)

	true false	true true	class precision
pred. false	6727	1879	78.17%
pred. true	1815	1851	50.49%
class recall	78.75%	49.62%	

- Foward selection:

accuracy: 72.20% +/- 1.36% (micro average: 72.21%)

	true false	true true	class precision
pred. false	7528	2397	75.85%
pred. true	1014	1333	56.80%
class recall	88.13%	35.74%	

KNN

accuracy: 67.57% +/- 1.31% (micro average: 67.57%)

	true false	true true	class precision
pred. false	7092	2530	73.71%
pred. true	1450	1200	45.28%
class recall	83.03%	32.17%	

- Foward selection:

accuracy: 72.12% +/- 1.16% (micro average: 72.12%)

	true false	true true	class precision
pred. false	7746	2626	74.68%
pred. true	796	1104	58.11%
class recall	90.68%	29.60%	

Algo a destacar es lo siguiente:

attribute	wei...
intake_type	1
intake_condition	1
animal_type	1
age_upon_intake	1
breed	1
color	1
datetime	0
castrated	0
sex_upon_intake_2	0

attribute	wei...
intake_type	1
intake_condition	1
animal_type	1
sex_upon_intak...	1
age_upon_intake	1
breed	1
color	1
datetime	0
castrated	0

attribute	weight
sex_upon_intake_2	1
intake_type	1
intake_condition	1
datetime	0
color	0
castrated	0
breed	0
animal_type	0
age_upon_intake	1

Se puede ver que a pesar de realizar el mismo foward selection, cada vez tomo atributos diferentes, aunque con algunas coincidencias. Esto se debe a que el algoritmo es avido, es decir que todo depende de donde tenga el punto de origen.