UNIVERSIDADE FEDERAL DE SANTA CATARINA - UFSC
SISTEMAS DE INFORMAÇÃO - INE - CTC
DATA MINING - INE 5644

Trabalho prático de Data Mining

Bruno Eduardo D'Angelo de Oliveira
Eduardo Demeneck
Carlos Henrique Costa Jr

FLORIANÓPOLIS, 2016

1. Definição do problema

A plataforma online Kaggle é destinada ao aprendizado de técnicas de *Data Science*, e para tal mesma proporção um ambiente para a criação de uma comunidade de entusiastas e profissionais, o compartilhamento de *datasets*, tutoriais básicos e introdutórios aos conceitos de exploração de dados, e a promoção de competições utilizando dados reais.

Uma das competições promovidas é relacionada a dados de um abrigo de animais da cidade de Austin, Texas nos Estados Unidos. O texto publicado elabora sobre como 7.6 milhões de animais acabam em abrigos nos Estados Unidos todos os anos, e como muitos destes animais acabam por serem eutanasiados. Após contextualizar o problema os organizadores da competição disponibilizaram arquivos com dados do abrigo de Austin, sugerindo que sejam encontrados neles os padrões que influenciam no destino de um animal, para que o abrigo possa vir a tomar decisões mais informadas sobre como lidar com cada animal que chega aos seus cuidados.

2. Descrição do entendimento dos atributos

Os arquivos disponibilizados estão no formato *.csv* e são os seguintes: um arquivo de teste, um arquivo de treino e um exemplo de submissão de dados à competição (o qual será ignorado para este trabalho). Os dados fornecidos são compostos pelas colunas: *AnimalID, Name, DateTime, OutcomeType, OutcomeSubtype, AnimalType, SexuponOutcome, AgeuponOutcome, Breed e Color*.

Antes de iniciarmos qualquer tipo de análise é necessário avaliar cada atributo, entender quais são os valores possíveis e seus significados e, por fim, decidir quais tem a maior chance de serem relevantes e quais devem ser descartados. Tal estudo foi realizado e os valores de cada atributo (e uma breve explicação de cada), e as decisões relacionadas a cada um deles, estão nas seções seguintes.

i. Atributos e seus valores

1. ***AnimalID*** é composto pela letra "**A**", seguida por um número de identificação de 6 dígitos, que representa o animal dentro do sistema de abrigos dos Estados Unidos.

2. ***Name*** é a coluna dedicada ao nome, caso se saiba, do animal.

3. ***DateTime*** é a data relacionada com o atributo ***OutcomeType***, ou seja a data em que a situação do animal teve um desfecho.

4. ***OutcomeType*** sinaliza o que ocorreu com o animal em questão, este atributo assume os valores descritos na tabela abaixo.

Tipo	**Descrição**
Adoção	Acolhido por família ou outra entidade
Morte	Óbito durante estada com o abrigo
Eutanásia	Eutanasiado / Sacrificado
Devolvido ao dono	Animal reunido ao dono original
Transferido	Transferido para outras instituições de abrigo e cuidados

Tabela 1: descrição dos dados da coluna OutcomeType.

5. ***OutcomeSubtype*** fornece informações adicionais em relação ao desfecho descrito na coluna acima. Este atributo assume os valores descritos na tabela que vem a seguir. É necessário notar que muitos destes valores só podem ser atribuídos a certas situações, pois não fariam sentido em outros contextos.

Tipo	**Descrição**
Agressivo	Animal sacrificado por comportamento agressivo
No veterinário	Faleceu no veterinário
Barn	Transferido para o *Barn Cat Program*
Comportamento	Sacrificado por conta do comportamento
Corte/Investigação	Sacrificado por ordem judicial
Em rota	Faleceu; está a caminho.
Adotado	Adotado por família
No lar adotivo	Faleceu no lar adotivo
No canil	Faleceu no canil
Em cirurgia	Faleceu em cirurgia
Médico	Sacrificado por motivos médicos

Fora do abrigo	Adotado por eventos Offsite(fora da instituição)
Parceiro	Transferido para uma entidade parceira do abrigo
Risco de Raiva	Sacrificado por risco de raiva
SCRP	Transferido para *Stray Cat Return Program*
Sofrimento	Sacrificado por estar em sofrimento

Tabela 2: descrição dos dados da coluna OutcomeSubtype.

6. ***AnimalType*** identifica se o animal em questão se trata de um cão ou gato.
7. ***SexuponOutcome*** identifica a situação do sexo do animal na data do desfecho, o atributo pode assumir os valores descritos na tabela 3.

Tipo	**Descrição**
Fêmea intacta	Fêmea que não foi esterilizada
Macho intacto	Macho que não foi esterilizado
Macho castrado	Macho foi castrado
Fêmea castrada	Fêmea foi castrada
Desconhecido	Sexo desconhecido

Tabela 3: descrição dos dados da coluna SexuponOutcome.

8. ***AgeuponOutcome*** descreve a idade do animal ao final da sua estada com o abrigo. Este atributo pode estar em: anos, meses ou semanas, dependendo de quão jovem é o animal.
9. ***Breed*** se refere a raça do animal e se ele é de raça pura ou não.
10. ***Color*** se refere a coloração da pelagem do animal e pode vir a ser composta por uma mistura de mais de uma cor.

ii. Decisões preliminares de uso de atributos

Após estudarmos os atributos da tabela fornecida e seus valores possíveis, pudemos notar que alguns deles não nos dão introspecção suficiente sobre o possível futuro do animal em questão. Enquanto outros atributos parecem influenciar pesadamente as chances de

adoção, contudo essas colunas necessitam ser trabalhadas de maneira apropriada para que seja possível utilizá-las. Nesta seção serão expostas as decisões de utilizar ou não certas colunas e o motivo de sua inclusão ou não no estudo.

Colunas selecionadas para análise

- ***Name*:** A existência ou não do nome do animal pode ser considerada como um dos fatores relevantes em alguns dos resultados; como adoção.
    - ➢ Esta coluna será transformada em um atributo booleano indicando a presença de nome ou não.
- ***DateTime***: A data da conclusão da estadia do animal na instituição foi escolhida pois é possível que datas significativas, como feriados, fim de semana e estações do ano possam influenciar no resultado de adoção. Também é possível que exista uma certa sazonalidade na frequência com que as pessoas adotam animais de estimação.
- ***AnimalType***: O fato do animal ser um gato ou cão pode influenciar fortemente o resultado, como chance de adoção ou transferência. Também é interessante separar o estudo para cães e gatos, desenvolvendo modelos diferentes para ambos os tipos de animais.
- ***SexuponOutcome***: Esterilização, ou a falta dela, pode ser um atrativo para certas pessoas quando procuram adotar animais; assim como o próprio sexo do animal.
- ***AgeuponOutcome***: A idade do indivíduo pode influenciar alguns resultados; como as pessoas escolherem o mesmo para adoção, pois é provável que filhotes tenham mais chances de serem adotados do que um indivíduo de maior idade e que possa apresentar problemas de saúde num futuro próximo.
    - ➢ Para este atributo é necessário que os valores sejam normalizados, pois estão em unidades de medida distintas (anos, meses e semanas).
- ***Breed***: A raça do animal influencia fortemente na escolha das pessoas no processo de adoção, visto que raças diferentes tem necessidades e características diferenciadas.
    - ➢ Para gatos os valores desse atributo serão utilizados para gerar uma nova coluna que classificará os gatos dependendo do comprimento de sua pelagem. Isso será feito em vista do fato de que a maioria dos valores de raça para gato são valores genéricos que descrevem comprimentos de pêlo, sem dar maiores informações sobre qual a sua raça. Para isso um documento oficial que detalhe os comprimentos da pelagem das raças de gatos domésticos será utilizado.
    - ➢ Para cachorros é necessário realizar uma generalização das raças baseadas em uma lista oficial das raças de cães existentes. Para esse objetivo ser alcançado,

será criada uma nova coluna, contendo o grupo de características raciais do indivíduo.

- ***Color***: A coloração da pelagem dos animais pode vir a influenciar na decisão de se adotar, ou não, o animal.

    ➢ Serão criados grupos de cores, baseados em listas oficiais de coloração para gatos e cães, para agrupar as características de coloração dos indivíduos.

Colunas descartadas

- ***AnimalID***: Este atributo representa apenas a identificação dada ao animal referente ao abrigo de animais, portanto não influencia em seu destino final.
- ***OutcomeType***: Este atributo representa a classe tentando ser prevista, assim não entrando no estudo.
- ***OutcomeSubtype***: Esta coluna age como "informações adicionais" para o desfecho da situação do animal, sendo assim não utilizaremos estas informações para o modelo que será criado.

3. Objetivo da mineração

O objetivo da mineração de dados é encontrar significado em informações desconexas, assim como correlações de padrões e tendências relevantes para tentar predizer o que acontecerá com os animais do *Austin Animal Center*. Avezado deste conhecimento, o abrigo de animais poderá melhor alocar seus recursos, para que os animais que precisam de um esforço maior para encontrar um lar recebam toda a atenção necessária.