**Project: Shelter Animal Outcomes**

Kaggle Competition Link: https://www.kaggle.com/competitions/shelter-animal-outcomes

**A group of dogs in a cage

Description automatically generated with low confidence**

**By**

**Venkata Saikiran**

**TABLE OF CONTENTS**

Section1 : Executive Summary................................................................................................. 03

Business Problem Statement

Data Source

Findings

Section2 : Benchmarking of Other Solutions............................................................................ 04

Comparison Table

Detailed Analysis on Solution 1

Detailed Analysis on Solution 3………………………………….……………… 06

Detailed Analysis on Solution 2………………………………………………… 05

Findings

Section3 : Data description and Initial Processing......................................................................07

Data Description

Exploratory Data Analysis

Findings from EDA……………………………………………….……………… 09

Section4 : Modelling

Relevance of independent variables.......................................................................... 10

Machine learning Models Comparison………………………………….…………11

Deep Learning Model Results……….………………………………….………....12

**Executive Summary:**

Business Problem:

If a pet cannot be given back to its owner or chosen for adoption, animal shelters may run into a number of challenging scenarios. Few circumstances, such as a lack of funding or available space for pets, can affect a shelter's decision to either move the animal to another shelter or to commit mercy killing. If an animal shelter can forecast a pet's behavior, it may aid in decision-making and assist the animal stand a higher chance of being adopted and finding a forever home.

**Data Source:**

The dataset is provided by [Austin Animal Center](http://www.austintexas.gov/department/animal-services) from time period October 1st, 2013 to March, 2016. The Austin Animal Center is the main municipal shelter center for Austin City and also for unincorporated Travis County. They always accept all kind of animals in need of shelter regardless of species ,age, health or breed. When an animal gets registered into a animal shelter center , each animal is given a unique identifier tag that is also used as a primary key in the outcomes dataset.

The dataset includes 26729 animal entries , with 9 input features. The data includes demographic information of the animals such as Name, Animal Type, Age, Sex, Breed and colour , the context of the animal’s outcome , and the outcome type while animal leaving the shelter. In this Project, models are developed to predict ‘Outcome Type’(The high-level reason for an animal leaving the shelter). This ranges from adoption / returned to owner, Died, Euthanasia and Transfer.

**Findings:**

* After the Exploratory analysis and Modelling of the overall dataset, I found that the For dogs, the most frequent outcomes were adoption, return to owner, and transfers; for cats, the most common outcomes were adoption and transfers exclusively.
* in regards to the importance of features ,random forest and XGboost provided similar results that shows that When adopting a pet, people tend to favor dogs over cats. Additionally, age was a significant factor in the dataset and is the main factor individuals take into account when looking to adopt an animal from a shelter.
* In addition to this, After the splitting of the dataset based on animal types; one being dogs and the other being cats to investigate whether there was a distribution difference between them. The findings indicate that there was a small variation between them, particularly in terms of the significance of traits. The adoption rates of cats were significantly impacted by whether or not they were given names.

**Section 2. Benchmarking of Other Solutions**

**Solution1:**

<https://www.kaggle.com/code/edwardelson/separate-classifier-for-cats-ad-dogs>

**Solution2:** [**https://www.kaggle.com/code/yy1252450987/animal-ml**](https://www.kaggle.com/code/yy1252450987/animal-ml)

**Solution3:** [**https://www.kaggle.com/code/iamtalos/machine-learning-end-to-end**](https://www.kaggle.com/code/iamtalos/machine-learning-end-to-end)

**Evaluation Metric:**

Solutions are evaluated using the metric[**Multi Class Logarithmic Loss**](https://www.kaggle.com/wiki/MultiClassLogLoss). The formula is

Text

Description automatically generated

**Comparison Table:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl.no** | **Notebook Name** | **Feature Approach** | **Model Approach** | **Logloss** |
| 1 | Loss Cat 0.5, Dog 0.9 | Split Classifier | Created new features from the existing dimensions by splitting training and test dataset into cats and dogs. | GradientBoostingClassifier  Stacking 2 Algorithms for final predictions | 0.75305 |
| 2 | animal-ML | Created Custom Functions to Preprocess the Existing Features | GradientBoostingClassifier | 0.83626 |
| 3 | Machine Learning End-to-End | Engineering New Features by combining the features | RandomForestClassifier | 0.84458 |

**Detailed Analysis on Individual Solutions:**

**Solution1:** Loss Cat 0.5, Dog 0.9 | Split Classifier

**Hypothesis** is that the Outcome Type distribution is different between cats and dogs .Based on this built different classifiers for cats' and dogs' dataset.

**Approach Followed:**

1. Spitted the training and test data in to two sets [ Cats and dogs]
2. **Feature Preprocessing** : Created custom functions to preprocess and transform the data to new features as mentioned in below table.

Used custom Coding for Converting Categorical Variables to numerical.

1. **Modelling** : Trained Gradient Boosting Classifier on two subsets [ Cats and Dogs] combined the models and made a stacked Classifier to predict the Test data

**Feature Preprocessing and Transformation:**

Original Data: Animal ID, Name, Datetime, Outcome type, Outcome subtype, Animal Type, Sexupon Outcome, Ageupon Outcome, Breed, Colour

Feature Engineering Techniques Used: Creating Bins, Grouping of data, Feature Splitting, Creating New Features

New Features Created: AgeuponOutcome, Breed, Colour, Year, Month, Day, Hour, Minute, Virginity, Sex, has\_name, hairgroup, aggressiveness, weight

**Solution 2:** animal-ML

**Approach Followed:**

1. **Feature Preprocessing** : Created custom functions to preprocess and transform the data to new features as mentioned in below table.

Used get\_dummies method to convert Categorical variables

1. **Modelling** : Trained Gradient Boosting Classifier to predict the Test data.

**Feature Preprocessing and Transformation:**

Original Data: Animal ID, Name, Datetime, Outcome type, Outcome subtype, Animal Type, Sexupon Outcome, Ageupon Outcome, Breed, Colour

Feature Engineering Techniques Used: Creating New Features, One Hot Encoding, Created Bins

New Features Created:Name, DateTime\_mmae, AnimalType, Breed, Color, HasName, DateTime\_day, Sex, IsIntact, Ispure, Ismix

**Solution 3:** Machine Learning End-to-End

1. **Feature Preprocessing** : Created custom functions to preprocess and transform the data to new features as mentioned in below table.

Used get\_dummies method to convert Categorical variables

1. **Modelling** : Trained Gradient Boosting Classifier to predict the Test data.

**Findings of 3 Solutions:**

1. The more the understanding of the features and doing feature engineering by creating customs functions results in better preparation of training data.
2. Solution 1 has created new features from existing which is meaningful and added more information to the model, so the model is able to train well resulting in better accuracy
3. Solution 3 has columns created new columns, but mostly all are the combinations of existing columns , so it is not adding much information , infact it is creating multi collinearity between columns, so the prediction accuracy is less

**Section 3 :** **Data description and Initial Processing**

**Data Description:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **DataType** | **Null Values** | **unique** |
| AnimalId | Object | 0 | 26729 |
| Name | Object | 7961 | 6374 |
| DateTime | Object | 0 | 22918 |
| Outcome type | Object | 0 | 5 |
| OutcomeSubtype | Object | 13612 | 16 |
| Animal Type | Object | 0 | 2 |
| SexuponOutcome | Object | 0 | 5 |
| AgeuponOutcome | Object | 18 | 44 |
| Breed | Object | 0 | 1380 |
| Colour | Object | 0 | 366 |

* Null values in Name can be ignored as w eare not considering the Name in the modelling
* OutcomeSubtype has almost 50% of Null values, so this can be imputed using anyone of the techniques like KNN or any other classification model
* We can see there are 2 unique animal Types and 1380 Breeds and 366 colors of pets

**Exploratory Data Analysis:**

**Analysis by Animal :**

In our data we have 26729 unique entries at the Austin Animal Center.

**Fig 1.1** : About 58% of the outcomes were associated to dogs, and 42% belongs to Cat

**Fig 1.2 :** Approximately 27.4% of dogs brought to the Austin Animal Center are returned to their owners, while 41.7% are freshly adopted. Only 4.5% of cats are given back to their owners, 38.3% are adopted for the first time, and 49.4% are moved to other shelters.

Fig 1.1 Fig 1.2

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

**Cat vs Dog :**

We can see in the below plot that Dogs are more likely to be returned to their owners and less likely to be transferred than cats.

Chart, line chart

Description automatically generated

Fortunately , if a pet was taken into an animal shelter and the shelter was able to match a name, the pet's outcomes turn out better than they could otherwise. 71% of the animals in this data set have names. Pets have a better chance of being adopted if they have a name. Similarly, if they had a name, 97% of dogs and 91% of cats were given back to their owners. However, 67% of cats were transferred if they lacked a name, for instance, because this was the case. The contrast is shown in the graph below.

Chart, waterfall chart

Description automatically generated

**Age of animals and outcomes:**

From the below plot we can infer that Animals returned to their owners tend to be slightly older. While, Cats are dying or being adopted or being transferred at a younger age.

A picture containing diagram

Description automatically generated

**Frequency of Adoptions:**

Concentrating on one of the better outcomes for pets. Figure 2.1 We can deduce that the majority of adoptions took place on the weekends by looking at adoptions by day of the week. A median of 17 adoptions take place at the Austin Animal Shelter on the weekends, compared to a typical of 9 adoptions on workdays. This pattern makes logical because adopting families may give their new pet more care on the weekends than during the workweek.

Chart, bar chart, box and whisker chart

Description automatically generated

We might also assume that there is a specific time of day when adoptions take place. The Austin Animal Center's website states that the facility is open from 11 AM to 7 PM. This explains why it makes sense that adoptions rise between 10 and 11 in the morning. The prime adoption period is from 5 to 6 PM. Additionally, we can see if hourly adoptions were affected more by weekends or by weekdays, and the trends remained consistent.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

**Findings from EDA:**

The key finding of this exploratory investigation is that dogs are adopted more frequently than cats. The chart below shows that each month, 200 to 250 dogs are adopted. However, when looking at trends in overall cat adoptions by month and year, we can detect cyclicity. We can see that cat adoptions peaked every July and fell off between February and May.

**Section 4 : Modeling**

**Relevance of independent variables:**

After the preprocessing of the initial data, the number of independent variables added to 14, where it consists of a group of categorical features as well, hence performed one hot encoding to convert in to numerical features.

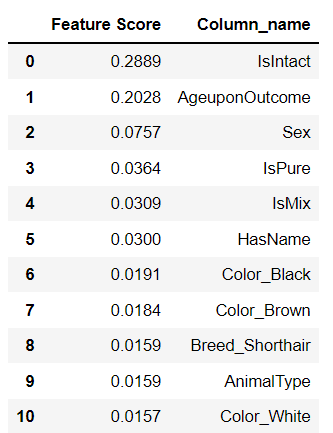
Final output shape of the data after Preprocessing is

**Training data shape**: 26729, 184

**Testing Data shape :** 11456, 184

Have implemented Random Forest Classifier Initially, A universal machine learning algorithm, which does a decent job of predicting in either regression or classification, philosophically. It builds a lots of [Decision Trees](https://en.wikipedia.org/wiki/Decision_tree) on random subsets of features and generalizes the relationships in the datasets decently to know the relationship between variables.

Top ten features as per Random Forest Classifier feature importance.



To predict the Outcome type , IsIntact , Ageuponoutcome,sex,ispure,Ismix are the Top 5 features as per the Feature Importance.

* As we know from EDA that Adoption is clearly having a much bigger percentage in the cases of Neutered Male and Spayed Female and Adoption of Intact Female is rather low.
* 'AgeuponOutcome' is the age of the pet, from this it is clear that Age of the Pets plays an

Important role in taking decision of Adoption.

* IsMix also play an Key role in influencing the decision of Adoption , If the pets are Mixed with different breeds then there chance of adoption gets low

**Model Performance comparison:**

I have tried fitting the data with all possible Classification models with the help of LazyClassifier.

lazypredict is a convenient wrapper library, that enables us to quickly fit all the models to our dataset and compare their performance.

Below is the result of LazyClaasifer:

Table

Description automatically generated

**Interpretation:**

From the comparison above, we can see that XGB classifier has better performance in terms of accuracy , but our metric is [**Multi Class Logarithmic Loss**](https://www.kaggle.com/wiki/MultiClassLogLoss)**.** So, I have taken standalone XGB classifier to see the Logloss is 84.02

References : EDA and Modelling coding part has been attached in a separate file EDA\_Modelling.ipynb