Domain background:

Open Data Initiatives around the world are allowing government and individuals to innovate and collaborate in new ways. Open Data provides the opportunity for transparency in communities, public service improvement, innovation, and economic value. For more information about Open Data Initiatives and resources, please visit: http://opendatatoolkit.worldbank.org/en/starting.html.

The Open Data domain of study chosen is: Animal Welfare.

"There is no simple solution to the complex problem of animal overpopulation and shelters remain an interim solution for the foreseeable future. Stray animals, especially cats, are often reviled and may enter shelters with a visible history of neglect and abuse. They are some of the most vulnerable animals that exist in our communities but despite their significant numbers, they remain largely unseen and unwanted by society. Veterinarians should strive to ensure that this population of disadvantaged animals receives compassionate, humane, and high-quality medical care in life and dignity in death. An awareness of the complex issues accompanying animal shelter management and the critical need for appropriate education and training of all personnel in contact with animals is essential to ensure good welfare of shelter animals." [Animal shelters and animal welfare: Raising the Bar, 2012]

Problem Statement:
There are plethora of animals in shelters waiting for adoption and rescue from people in communities. To better understand the trends and gaps in data about these animals, a dataset will be studied, and a model trained to predict the probability of animals being adopted within the allowable timeframe of retention. Further, study of impacts of the pandemic on the seasonality of animal rescue is limited, so this is a project to better understand the impacts of the pandemic on animal shelters and their residents.

For the purposes of the course this will be submitted to, a single city will be focused on, and that city will be Austin, Texas.

Datasets and inputs:

Information about the dataset can be found at the following links:

- https://www.kaggle.com/datasets/aaronschlegel/austin-animal-center-shelter-intakes-and-outcomes
- https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Intakes/wter-evkm

The dataset is public record of animal rescue profiles for intake and outcomes for animals rescued in animal shelters in Austin, Texas from the Austin Animal Center as part of the city of Austin's Open Data Initiative.

There are a total of three tabular datasets each with the primary key: animal id. These are the datasets:

- 'acc_intakes.csv': includes columns and features for age upon intake, animal type, breed, color, datetime of intake, location found, intake condition, and intake category
- 'acc_intakes_outcomes.csv': includes columns and features for age upon outcome, date of birth, outcome subtype, outcome type, sex upon outcome, age upon outcome in days and years, and outcome datetime

- 'acc_outcomes.csv': includes columns and features for age upon outcome, animal type, breed, color, date of birth, datetime, month and year of outcome, name of animal, and outcome subtype

Notice some of the columns and features are redundant in datasets 'acc_intakes_outcomes.csv' and 'acc_outcomes.csv'. A portion of time for this project will be dedicated to exploratory data analysis (EDA) to become familiar with these redundancies and feature correlations. Then the data will be split and used to train, test, and validate a model that will aim to predict the probability of an animal adoption.

These datasets were obtained from the Kaggle platform and the data collected daily from the Socrata Open Data Access API listed in the Kaggle profile. The datasets are also listed on the Austin, Texas government site 'data.austintexas.gov' and can be verified/downloaded directly from this site.

Solution Statement:

A solution machine learning model that will predict probability of an animal being adopted will be developed and trained with the dataset in AWS Sagemaker notebooks and then deployed and invoked on AWS services (AWS Sagemaker and Lambdas) to provide a proof-of-concept deployment method that may be integrated into organization operations.

With a better understanding of animal adoption statistics and outcomes, this citizen aims to uncover insights about the impact of recent global events on adoption rates and outcomes of sheltered animals. This initial project for an open data initiative may also contribute to discussions on what data may be valuable in addition to what is already being documented and contribute to the open data initiative purpose of informing the government of gaps there may be in datasets for further research.

Benchmark Model:

I will be benchmarking with a notebook and model completed by a Kaggle user of similar background (beginner-level): https://www.kaggle.com/code/wenlie/exploring-and-predicting-the-animal-s-outcomes

The model is a random forest classifier that predicts animal outcomes.

Evaluation Metrics:

This project can be measured by the accuracy of the model on validation dataset and number of uncovered insights proposed for improving open data initiative dataset.

The insights themselves can be measured by the number of actionable takeaways they create for improving adoption rates and intake and outcome rates of animals in these shelters.

Project Design:

PROCESS/WORKFLOW

- Exploratory Data Analysis (EDA): that dataset will be explored using tools demonstrated in this course, including Pandas python package, Auto-visualization tools
- Data Cleaning: Following exploratory data analysis, the data will be cleaned if there are missing values, inconsistencies in data, or imbalances. Further, feature engineering may be required for categories of outcomes and animals.

- Model Selection: Since the data is tabular and has a mix of data types, the most likely models challenged against each other will be XGBoost-related and ensemble learning techniques. An AutoML tool may be used to expedite comparison of model options.
- Model Hyperparameter Optimization: Once models are compared and a champion is selected, the model hyperparameters will be optimized using AWS Sagemaker tools.
- Model Training: From hyperparameter optimization, the best hyperparameters will be extracted and used for final training of model predicting outcomes of animals based on features.
- Model Deployment: Finally, the model will be deployed to an endpoint and tested to ensure the response is understandable. Further considerations about the analysis and proposed takeaways from model results will be documented.

DELIVERABLES

- AWS Sagemaker Notebook Outlining Process Findings
    - EDA Visualizations and Summary
    - Machine Learning Model Testing and Training
    - Sample Code for Deployment
- Blog-ready Documentation about Model and Analysis Key Takeaways
    - Benchmark Model Comparative Analysis
    - List of possible actionable takeaways for operation and stakeholders
    - Stretch Goal: Cost/Value analysis of Model Operationalization