Nicholas Brower
Springboard DSCT – May 2022
Unit 7.1.1 - Capstone Two - Project Ideas

1. Credit Default Prediction

Use anonymized account activity data to predict credit default among American Express cardholders. This dataset includes purchase, payment, and other account activity information for users within a specific time window.

Data:

https://www.kaggle.com/competitions/amex-default-prediction/data

2. Book Review Prediction

Predict a book review rating on a scale of 1 to 5 for Goodreads users. This dataset provides information on a user's account activity on Goodreads, including user ID, review dates, review ratings, book identification, and the text of each review. Build a model to predict a user's score given their history, the text of a review, and the ID of a book.

Data:

https://www.kaggle.com/competitions/goodreads-books-reviews-290312/data

3. Pennsylvania Roadway Maintenance

Given roadway condition, traffic volume, construction, weather, and maintenance data, predict when segments of roadways will require maintenance. Either focus on Pennsylvania's most traveled road segments, or limit data to road segments in a single county.

There are many datasets relevant to the creation of a predictive model. Pennsylvania's Roadway Mangement System and Department of Transportation websites offer useful introductions to much of the required roadway data. Weather data is easily obtained. Underlying soil compositions may be relevant for some roads; the US Department of Agriculture offers this data in many useful formats.

Info:

PA Roadway Mangement System
https://www.penndot.pa.gov/ProjectAndPrograms/ResearchandTesting/RoadwayManagementandTesting/Pages/Road-Management-System.aspx

PennDOT Open Data
https://data-pennshare.opendata.arcgis.com/

Data:

Traffic Volumes
https://data-pennshare.opendata.arcgis.com/datasets/PennShare::rmstraffic-traffic-volumes/about

Transportation Improvement Projects – Points
https://data-pennshare.opendata.arcgis.com/datasets/PennShare::transportation-improvement-projects-points/about

Transportaiton Improvement Projects - Lines
https://data-pennshare.opendata.arcgis.com/datasets/PennShare::transportation-improvement-projects-lines/about

Weather:
http://www.climate.psu.edu/data/city_information/index.php?city=phl&page=dwa&type=big7

USDA Soil Survey
https://www.nrcs.usda.gov/wps/portal/nrcs/surveylist/soils/survey/state/?stateId=PA

4. Video Game Recommendations (Steam)

Create a game recommendation system based on a Steam user's habits and preferences. Tailor recommendations to account for how many titles a user plays per year, the amount of time spent playing per week, and the amount of time spent playing specific titles. Use game critic reviews, aggregated external user review data, and Steam popularity charts to inform recommendations.

Steam review data is publically available. Various subsets of Steam user and application data are available online, but I would scrape my own data for this project. Game critic and aggregate user reviews would also be scraped from Metacritic and Amazon. For reference, some publically available subsets are listed below.

Data:
Jianmo Ni's 2018 Amazon Review Data (Video Games)
https://nijianmo.github.io/amazon/index.html

Steam Users (time-series and purchase data)
https://www.kaggle.com/datasets/tamber/steam-video-games

Top Video Games 1995-2021 Metacritic (Kaggle)
https://www.kaggle.com/datasets/deepcontractor/top-video-games-19952021-metacritic

5. Better Bird Call Identifier

Predict a bird's species given an audio recording of its call. The Cornell Lab of Ornithology hosted several Kaggle competitions with this aim in mind. Implement robust filtering, preprocessing, feature engineering, and data synthesis stages in attempt to provide better accuracy than previous models. Engineer song classification features based on the Smithsonian's bird song classification key and the “PDHF” principles outlined in the research paper linked below. Before training a model, generate synthetic data for each target species using gated, filtered, dynamically modulated portions of existing calls over environmental noise loops from unrelated or external samples. Use equalization to simulate a range of distances and the off-axis frequency reponse of various microphones. Specify modulation parameters using the provided training data, estimated population variance, and bioacoustic morphology of each species. Generate the synthetic data by drawing random samples (without replacement) from filtering, noise, and modulation parameter distributions.

Data:
Birdclef Data
https://www.kaggle.com/competitions/birdclef-2022/data
https://www.kaggle.com/competitions/birdsong-recognition/data

Smithsonian's National Zoo & Conservation Biology
https://nationalzoo.si.edu/scbi/migratorybirds/education/nasongkey.pl?glance=

Info:
Automatic Classification of Monosyllabic and Multisyllabic Birds Using PDHF (Perceptual, Descriptive, and Harmonic Features)
https://www.mdpi.com/2079-9292/10/5/624

The Journal of the Acoustical Society of America: Semi-automatic classification of bird vocalizations using spectral peak tracks
https://asa.scitation.org/doi/abs/10.1121/1.2345831