Results of 2nd Vote

1. Gentrification****
2. Is there a correlation between consumption of alcohol and the economy?
3. Spotify Song Attributes

Note: Choice 3 was of high interest but is a classification problem which can be quite challenging depending upon the dataset.

Possible Topics

Problem 1 : Crime Prediction - Predict where next crop thefts will occur by comparing crop theft with prices (worth) of crops. Use avocado or almonds as crop to train.

ML Type: Regression

Why: Growing problem along US border on imported crops. Preposition law enforcement

Data Source: Crime Data and Kaggle Avocado Prices both obtainable form Kaggle

Question Answered: Where should law enforcement be prepositioned to prevent theft crime.

Possible Topics

Problem 2 : X-Rays are used to predict the presence of Pneumonia in patients. ML can be applied to look for subtle signs in the X-Ray data to determine if present.

ML Type: Classification (Predictive)

Why: With COVID pandemic, medical staff is stretched to the limit. Using ML to define possible dianosis can help improve efficiency of medical staff.

Data Source: Kaggle - Chest X-Ray Images Pneumonia labeled dataset.

Question Answered: Preliminary Screening for diagnosis.

Possible Topics

Problem 3 : Fake Job Description Prediction. Being able to classify job listings as likely real or fake. ML Type: Classification

Why: As we come out of the COVID Pandemic, businesses will be restarting with associated job postings. However, fake job postings will also appear misleading many job searchers. Applying ML to recognize a likely Fake Job Posting from a Real Job Posting would be useful to both job searchers as well as businesses doing the hiring.

Data Source: Kaggle - Fake Job Description Prediction

Question Answered: Is this a real job posting worth my time?

Possible Topics

Problem 4 : Influential Bloggers - Getting a jump start on restarting your business after pandemic. .

ML Type: Regression

Why: After COVID, businesses will need restart sales of products and services. Influential and network marketing now tops the list for new marketing strategies in the current age. Finding influential bloggers to assist you in marketing your product can be a valuable tool to gain -- or regain recognition for your new product, restraunt or other type of business. Getting prepared now for the biggest comeback in history will set the stage for winners and losers.

Data Source: Kaggle - Identifying Influential Bloggers using standard blog measures.

Question Answered: Who are the most influential bloggers which might assist you in restarting your business.

Topic #5 ***

Hypothesis: Is there a correlation between consumption of alcohol and the economy?

ML Type: Using Linear Regression and past economic upswings and downswings, demographic data is disaggregated and plotted against alcohol sales to determine if a true correlation exists and along what parameters.

Why: Economic imbalance and disparity in the US has been growing since 2000. Could easily obtainable alcohol sales be an indicator of citizen satisfaction or dissatisfaction with the state of affairs?

Data Sources: Data.gov census data on demographics and economic. Alcohol sales from Kaggle.

Question Answered: Can alcohol sales be used to sense citizen sentiment or stress levels within society.

Alena ***

Gentrification

G. is a growing concern among numerous neighborhoods in different cities and a hot topic for developers looking to invest money.

With this research we are hoping to find out the combination of factors that play the key role in an area potentially becoming gentrified. And define the opportunity zones through clustering algorithms.

(Opportunity zone – is basically where you invest money in local businesses for certain timeframe and get tax breaks on the money)

Data sources I've got so far:
https://opendata.imspdx.org/dataset/gentrification-and-displacement-risk-typology/resource/5dd02ef6-cd29-42cd-bd3f-f568a2ed2f65

http://gis-pdx.opendata.arcgis.com/datasets/b4f89af7f4964d1db0a76faac2cbb811_245

https://datasetsearch.research.google.com/search?query=Gentrification&docid=ZNw%2Ff8ouOj1nSTtTAAAAAA%3D%3D

University Rankings *

ML type: neural net

Looking at the different university grading policies to accurately determine the ranking of each universities. What university is the best? Do influence, employees, publications, or citations make a difference?

Kaggle https://www.kaggle.com/mylesoneill/world-university-rankings

What is the top ranking universities in the world?

NBA stats prediction

MI type: neural net

Predicting the stats of the 2019-2020 season to see who could be the potential mvp and other awards. Which players should be selected to create the perfect fantasy team?

Kaggle https://www.kaggle.com/calvingee/nba-stat-projections-2019

Looking at previous stats to potentially determine which players could be chosen for a fantasy league?

Spotify Song Attributes ***

MI type: unsupervised learning

Analyzing the different features of a song to then categorize them into different groups using unsupervised learning. Does time signature, tempo, loudness, etc all make a difference and are there more correlation between different genre music over similar genre music?

Kaggle https://www.kaggle.com/geomack/spotifyclassification

What determines the genre of music and can ml properly categorize music?

Horse Racing *

MI type: neural net

Analyzing horse racing data to compare and create an accurate prediction on the next winning horse for the next race.

Kaggle https://www.kaggle.com/gdaley/hkracing#runs.csv

Which horse can be the next winner in the hk race to potentially earn the winning prize?

Wildfire risk in California *

ML Type: Regression/predictive

Why: Wildfires have been affecting California greatly over the past few years. Understanding the most at-risk areas would benefit residents, landowners, and business owners.

Data Source:

ncdc.noaa/gov/stormevents/

Bulk data downloads (csv) from database

-search for datasets of parameters that may impact risk potential

Questions to answer:

What are the chances of a wildfire in specific areas? (search by zip code or GPS coordinates)

Are wildfires affecting specific zones more regularly than others?

Are the number of days with wildfire events per year increasing? If yes, what can we project for 2025-2030?

Drought in California

ML Type: Regression?

Why: Increasing concern throughout the world, but specifically in California.

Data Source:

Kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data

Drought dataset from USGS

ArcMap

Questions to answer:

What areas are being impacted by drought the most currently? What areas will be impacted the most in the future?

Homelessness in California

ML Type: ??

Why: Concern in San Francisco

Data Sources: ??

Questions to answer: How will crime rates be affected as homelessness decreases?