Capstone 3 Project Idea

Springboard Data Science Career Track

Zizhong Liu

1. Applying classification models for predicting restaurant inspections in Las Vegas

Datasets:

- City of Lags Vegas Portal (https://opendataportal-lasvegas.opendata.arcgis.com/datasets/restaurant-inspections-open-data/explore)
- Food Establishment Inspection Report (https://www.southernnevadahealthdistrict.org/download/eh/fe-inspection-report.pdf)

Objectives:

- Features: restaurant serial number, restaurant permit number, restaurant name, restaurant location, restaurant category, address, city, state, zip, current demerits, current grade, employee count, median employee age, median employee tenure, inspection time, inspection type, inspection demerits, violations raw, record updated, lat long raw, first violation, second violation, third violation, first violation type, second violation type, third violation type, number of violations, next inspection grade c or below, lat, long
- Target variable: Next Inspection Grade C or Under C
- Tasks:
    1) Conduct exploratory data analysis of the dataset.
    2) Build classification models that predict the outcome of a restaurant's next inspection, using NEXT_INSPECTION_GRADE_C_OR_BELOW as the response.
    3) Provide recommendations for how the dataset can enhanced to improve the predictive power of the model.

2. Applying neural network model for predicting CO2 gas emissions

Datasets:

On the basis of the dataset in Capstone 2, more data set are added:

- World Life expectancy by country from 1949 to 2019 annually (https://github.com/owid/co2-data)
- World Electricity net production by county from 1990 to 2019 annually (http://data.un.org/Data.aspx?d=EDATA&f=cmID%3aEL%3btrID%3a019)
- World total fuel oil consumption by county from 1990 to 2019 annually (http://data.un.org/Data.aspx?d=EDATA&f=cmID%3aRF)
- World precipitation by country from 1901 to 2020 monthly (http://data.un.org/Data.aspx?q=precipitation&d=ENV&f=variableID%3a6) (https://climateknowledgeportal.worldbank.org/download-data)
- Global temperature by country from 1743 to 2020 monthly (https://www.kaggle.com/akshaychavan/average-temperature-per-country-per-year/data?select=GlobalLandTemperaturesByCountry.csv), (https://climateknowledgeportal.worldbank.org/download-data)
- World religion distribution of each country
- World capital cities and their location
- World official languages of each country

The following are datasets used in Capstone 2:

- Global CO2 Emission by country from 1949 to 2019 annually (https://github.com/owid/co2-data)
- Agricultural land use from 1961 to 2018 annually (http://data.un.org/Data.aspx?d=FAO&f=itemCode%3a6610), (http://data.un.org/Explorer.aspx)
- Beef production use from 1961 to 2018 annually (http://data.un.org/Data.aspx?d=FAO&f=itemCode%3a6610), (http://data.un.org/Explorer.aspx)
- World population by country from 1949 to 2019 annually (https://github.com/owid/co2-data)
- World GDP change by country from 1949 to 2019 annually (https://github.com/owid/co2-data)

Objectives:

- Predict CO2 emission based on the above features.
- Features: Date (date-type data); Countries, capital cities, religions, languages (categorical data); Agricultural land use, temperature, precipitation, population, GDP, Life expectancy,

Electricity net production, total fuel oil consumption (numerical data); capital city locations (geographical data)

- Target variable: CO_2 emission (numerical data)
- Examples: global CO_2 Emission by country from 1949 to 2019 annually. (examples >> features)
- Suggestions: 1) apply NLP to categorical data to perform feature engineering; 2) combine different features with ANN method to check which feature combination give the most exact prediction.

3. Stock price predicting: using time-series forecasting method

Datasets:

- Stock price: Date, Volume, High, Low, and Closing Price for all NASDAQ, S&P500, and NYSE listed companies since their listing up till now. (https://www.kaggle.com/paultimothymooney/stock-market-data)
- Gold daily price in different currencies since 1978 (https://www.gold.org/goldhub/data/gold-prices)
- WTI crude oil price daily price in USD since 1986 (https://fred.stlouisfed.org/series/DCOILWTICO)
- US Consumer Price Index in U.S. City Average since 1950 in daily manner (https://fred.stlouisfed.org/series/CPIAUCSL)
- US Inflation Rate monthly since 1913 (https://inflationdata.com/Inflation/Inflation_Rate/HistoricalInflation.aspx)
- Effective Federal Funds Rate daily since 1955 (https://fred.stlouisfed.org/series/DFF)

Objectives:

Subproject 1 (entry-level warm-up):

- Choose one stock and predict its price using time-series forecasting method.
- features: Date (date-type data)
- target variable: closing price (numerical data)
- examples: daily stock prices since its listing. (examples >> features)

Subproject 2:

- Predicting QQQ and VOO based on gold price, crude oil price, US Consumer Price Index, US Inflation Rate, and Effective Federal Funds combining time-series forecasting method.[it is temporal data; shift target one unit, ARIMA]+[NLP news related to predict human behavior (JP Morgan vs. Deloitte)]
- features: Date (date-type data); gold price, crude oil price, US Consumer Price Index, US Inflation Rate, and Effective Federal Funds (numerical data)
- target variable: closing price (numerical data)
- examples: daily stock prices since its listing. (examples >> features)
- suggestions: 1) built a long term model for both co2 and stocks; 2) combining co2 and voo, create annual income joined dataset (or month); 3) work on the dimension reduction technique

4. Climate Change Impacts on the Global Food Supply

Agriculture is an important sector of the U.S. economy. The crops, livestock, and seafood produced in the United States contribute more than $300 billion to the economy each year. When food-service and other agriculture-related industries are included, the agricultural and food sectors contribute more than $750 billion to the gross domestic product. Agriculture and fisheries are highly dependent on the climate. Increases in temperature and carbon dioxide (CO_2) can increase some crop yields in some places. But to realize these benefits, nutrient levels, soil moisture, water availability, and other conditions must also be met. Changes in the frequency and severity of droughts and floods could pose challenges for farmers and ranchers and threaten food safety. Meanwhile, warmer water temperatures are likely to cause the habitat ranges of many fish and shellfish species to shift, which could disrupt ecosystems. Overall, climate change could make it more difficult to grow crops, raise animals, and catch fish in the same ways and same places as we have done in the past. The effects of climate change also need to be considered along with other evolving factors that affect agricultural production, such as changes in farming practices and technology. (https://climatechange.chicago.gov/climate-impacts/climate-impacts-agriculture-and-food-supply)

Datasets:
- Global CO2 Emission by country from 1949 to 2019 annually (https://github.com/owid/co2-data)
- Global temperature by country from 1743 to 2020 monthly (https://www.kaggle.com/akshaychavan/average-temperature-per-country-per-year/data?select=GlobalLandTemperaturesByCountry.csv), (https://climateknowledgeportal.worldbank.org/download-data), (https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND)
- World precipitation by country from 1901 to 2020 monthly (https://climateknowledgeportal.worldbank.org/download-data)
- World Food Supply - Crops Primary Equivalent by country from 1969 to 2018 yearly (http://data.un.org/Data.aspx?d=FAO&f=itemCode%3a2516)
- World Food Supply - Livestock Primary by country from 1969 to 2018 yearly (http://data.un.org/Data.aspx?d=FAO&f=itemCode%3a1806)

Objectives:
- Predict crop and livestock production based on the above features.
- Features: Date (date-type data); Country (categorical data); Global CO2 Emission, temperature, precipitation, (numerical data).
- Target variable: crop and livestock production (numerical data)

- Examples: crop and livestock production by country from 1969 to 2018 annually. (examples >> features)