2-Day Project

You are given 3 questions in this area. Work on 1 question of your choice.

1. Clock Image Recognition System

Problem Description. Analog clocks have a clock face which indicates time using rotating pointers called "hands" on numbered dials. The standard clock face has (1) a short "hour hand" which indicates the hour on a circular dial of 12 hours, making two revolutions per day, and (2) a longer "minute hand" which indicates the minutes in the current hour on the same dial, which is also divided into 60 minutes. In this project, our goal is to build a model that can recognize images of clocks and predict the time. We have a dataset consists of synthetically generated images of clocks with the clock hands placed on one of many different clock faces. Each clock image indicates certain time by two hands explained above and is labelled in the form hour-min, so for a time like 1:30 the associate label would be 1-30, and we use it as folder names to properly manage clock images in the dataset. The dataset includes a single train subset and two different test subsets. Some images in the dataset are also randomly rotated between 0,-90, +90 and 180 degrees.

- Please refer to the **train subset** in the link below:

  `https://tinyurl.com/h5sjkj5c`

  The train subset includes images, and labels of them are indicated by folder names. In this project, you are **not allowed to use additional data** to improve the learning.

- Please refer to **two different test subsets** in the link below:

  `https://tinyurl.com/3wt3afzb` and `https://tinyurl.com/mpvva9xz`

  These two test subsets are structured in the same way as the train one. In this project, our goal is to have a model that similarly **performs well on both test subsets** (at least on the first one).

(a) **Data Preprocess.** Describe how you preprocess the data into a format to build your model and how you split the dataset to properly build your model.

(b) **Exploratory Data Analysis.** Explore the dataset with descriptive statistics and visualization to understand the problem.

(c) **Learning Algorithm.** Build a model to achieve the goal explained above using available data. Describe your learning algorithm (or statistical method) to build your model, and explain how it is implemented.

(d) **Evaluation Metric.** Define your evaluation metric, and explain why it is appropriate for the problem.

(e) **Result and Discussion.** Present your work at the oral exam, including (but not limited to) problem setting, hypotheses, how you preprocess data, description and justification of your learning algorithm, prediction results, interesting observations or insights learned, and potential future work.

2. Pokemon Character Classification and Prediction

Problem Description. Pokemon is one of the most successful media franchises. The franchise is centered on fictional characters called "Pokemon." In Pokemon, humans are known as trainers. Trainers capture and train Pokemon to battle other trainer's Pokemon for sport.

Suppose you are a new manager (in the real world not in the Pokemon world) in charge of Pokemon creature development. Assume that you do not have prior knowledge on Pokemon creatures.

1) As the first task, you want to understand the relationship among various Pokemon attributes. For example,

1. How are one's height and weight related with one's HP or attack/defense?
2. How do these relationships evolve over generations of Pokemon? (Pokemon has released 8 generations so far.)
3. Which attributes significantly explain why certain Pokemon characters are classified as legendary or mythical?

Develop some of these questions into the relationship among variables and answer with exploratory data analysis and/or regression.

2) As you gain some understanding into Pokemon creatures, you plan to develop a new Pokemon character. Since each Pokemon has one or more type (e.g., water type, fire type, etc.), your newly developed character must have one or more types. Develop a model to appropriately suggest one or more types for the new character based on other attributes such as HP, attack, defense, etc. To be more concrete, first develop a Pokemon type prediction model of based on generations 1 thru 7 to make predictions for generation 8 Pokemons.

3) Some Pokemon creatures are classified as legendary. You want to check whether your newly developed character should be marked legendary. Develop a model to predict whether a given Pokemon character should be classified as legendary.

4) Evaluate and justify the performance of your models.

- Please refer to the datasets in the link below:
  `https://www.kaggle.com/datasets/joshuabetetta/complete-pokedex-v100?select=Complete+Pokedex+V1.1.csv`[1]

  Please feel free to use these additional datasets to build your model, although it is not required for this problem.

(a) **Exploratory Data Analysis.** Explore the data with descriptive statistics and visualization. (ex. finding important features or factors having a high positive or negative impact on the target variable)

(b) **Learning Algorithm.** Build a model to achieve the goal explained above using available data and information or additionally collected data. Describe your learning algorithm (or statistical method) to build your model, and explain how it is implemented.

(c) **Evaluation Metric.** Define your evaluation metric, and explain why it is appropriate for the problem.

(d) **Result and Discussion.** Present your work at the oral exam, including (but not limited to) problem setting, hypotheses, how you collect and preprocess data, description and justification of your learning algorithm, prediction results, interesting observations or insights learned, and potential future work.

[1] The file is also available below. (`https://www.dropbox.com/s/hcuqctfrmz6xsmz/Complete%20Pokedex%20V1.1.csv?dl=0`).

3. Stock price predictions and anomaly detection.

Problem Description. Stock prices are public time series data that pertains a long history. Even right at this moment, many trading companies are analyzing historical stock prices in order to beat the market with the automated-trading system. The stock market dataset is interesting as it contains hundreds of time series (one per company) that are dependent to each other. For example, there could be a sector information that groups multiple timeseires together or there could be external factors such as war and pandemics (e.g. COVID) which could be regarded as an outlier.

We will mainly work with historical stock prices of S&P 500 companies[2]. The following data has 'Date', 'Volume', 'High', 'Low', and 'Closing Price' that are updated daily.

- Please refer to the dataset in the link below (stock_market_datasp500):
  `https://www.kaggle.com/datasets/paultimothymooney/stock-market-data`

  The dataset also has 'forbes2000', 'nasdaq', 'nyse' folders in addition to 'S&P500'. Feel free to use these datasets as well if you wish to.

- You may also crawl your own S&P dataset, following this blog (very easy).

In this project, the goal is to build models that can spot outliers, an erratic market behaviors just by monitoring market trends. Outlier detection may be very important in the automated trading; upon the emergence of such outlier events, stock prediction model is likely fail as the model is not accustomed to such outside events. How would you define outliers? Build or split your own test set based on your definition. You may use future prediction models or any other machine learning (statistical or neural) models that you can think of, however, be careful to not look into the future ($t+1,\ t+2,\ \dots$) in order to report a time point to be an outlier.

(a) **Additional Data Collection.** In addition to the data listed above, you may also use additional data to improve the model fit, and please specify the datasets and explain why and how the additional datasets are needed.

(b) **Data Cleansing & Data Loading.** Describe how you preprocess the data into a format (ex. how to resolve issues of missing data / use of dimension reduction technique / normalization of values) to build your model and how you split your training and testing sets.

(c) **Exploratory Data Analysis.** Explore the data with descriptive statistics and visualization. (ex. finding important features, statistics, or factors having a high positive or negative impact on the target variable)

(d) **Learning Algorithm.** Build a model to achieve the goal explained above using available data and information or additionally collected data. Describe your learning algorithm (or statistical method) to build your model, and explain how it is implemented.

(e) **Evaluation Metric.** Define your evaluation metric, and explain why it is appropriate for the problem.

(f) **Result and Discussion.** Present your work at the oral exam. Your presentation should include (but not limited to) a description of problem setting, hypotheses, how you collect and preprocess data, description and justification of your learning algorithm, prediction results, interesting observations or insights learned, and potential future work.

[2] Standard and Poor's 500 companies that tracks 500 large companies listed on exchanges in the United States.