2-Day Project

You are given 2 questions in this area. Answer 1 question of your choice.

1. Crime Rate Analysis

Problem Description. Crime rate analysis is a crucial yet challenging problem, which requires a lot of knowledge across domains. Furthermore, it's important to foresee the trend in crime to prevent particular crimes from happening rather than deal with them after they occur. Besides, finding the root causes of crime rates helps the government come up with appropriate policies to improve the situation. The related dataset, Crime Rate Dataset[2], was derived from the Summary Reporting System (SRS), which contains estimated crime data of 50 states from 1979 to 2019. These data reflect the estimates that the FBI has traditionally included in its annual publications. The crime rate data set has been downloaded and stored on Google drive via the link below for your convenience.

`https://drive.google.com/drive/folders/1oFGX6vCWlGcTZtjjt2WLoGfqEZXk-QkC`

The crime data (`estimated_crimes.csv`) includes such as crime type and number of occurrences. We also include 3 additional data sources on Google drive, which contains relevant data sets to crime rates, including:

- `demography.csv`: Demographic data, including population, race, and age distribution, for each state, etc.
- `economics.csv`: Economic data, including workforce and employment statistics, natural resources, industry distribution, etc.
- `social_characteristics.csv`: Social characteristics, including average education level, ethnic background, percentage of residences were foreign-born, percentage of residences were illegal immigrants, etc.

In this project, our goal is to use these datasets to build models to predict (1) the crime rate[3] of the whole USA (including 50 states and territories) and any particular 5 states (having different profiles in demography, economy, and social characteristics). We further want to predict (2) the crime rates of any particular 2 crime types[4] in the USA.

(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) to build your model and how it is split into training and testing sets.

(c) **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)

(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, 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.

[2]Original source: `https://www.kaggle.com/tunguz/us-estimated-crimes`

[3]Crime rate = total number of crime / population

[4]Please pick from violent crime, homicide, robbery, aggravated assault, property crime, burglary, larceny, and motor vehicle theft.

2. Seoul Bike Sharing System

Problem Description. Seoul bike sharing system (`https://www.bikeseoul.com/`) is widely used as an eco-friendly public transportation in Seoul, offering 24/7 convenient access to nearly 40,000 bikes and 2,600 stations across the city. As the number of active users is consistently increasing thanks to its convenience, Seoul metropolitan government is encouraging data scientists to explore and build prediction models that can further improve the bike share system. For this reason, the government has been publicly sharing related datasets `https://data.seoul.go.kr/`, and our goal in this project is to use them to build and test prediction models. More specifically, we expect to have a model that can be useful to properly allocate bike resources in the stations, and this will let all users to conveniently and smoothly pick up bikes without any wait. To do that, the simplest way is to build a model that can predict the destination station given the departure station with some extra information such as date and time. We can further consider more sophisticated models such as a model that can temporally and jointly predict the number of station-wise incoming (and outgoing) bikes.

- Please refer to the datasets in the link below:
  `http://data.seoul.go.kr/dataList/OA-15182/F/1/datasetView.do`[5]

  The datasets include multiple rows (data points), where each corresponds to a single bike ride route, and columns such as departure / destination stations and time information. For the simplest setting, the departure station (and time information) can be used as a input variable, and the destination station (and time information) as a target variable.

- There are more datasets related to this project (bike sharing) in the link below:
  `https://data.seoul.go.kr/dataList/5/literacyView.do`

  Please feel free to use these additional datasets to build your model. It is highly recommended to further collect and utilize more data sources if you can do it.

(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) to build your model and how it is split into training and testing sets.

(c) **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)

(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, 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.

[5]Some of them are available here (`https://drive.google.com/drive/folders/17Z-DpQGbsRpiHNDzWMXdCVnGcqh5wx9B`).