Data Science

(Spring semester)

Term Project Additional Specification

Team 8

202135810 이규석

202235019 김선영

202239868 김현우

202334452 남경민

"

1. Dataset

- Must contain some categorical features and some numerical features

- Must contain a reasonable amount of dirty data (missing data, wrong data)

• **Sample(data) count:** Total 75.

5 countries for each developed, middle income, underdeveloped countries. Each country have 5 records from different years. (ex. 2010, 2020..)

• **Feature(column) Type**

Feature1: population (numerical)

Feature2: minimum wage (numerical)

Feature3: GDP (numerical)

Feature4: GDP per capita (numerical)

Feature5: unemployment rate (numerical)

Feature6: income distribution(gini coefficient) (numerical)

Feature7: national debt (numerical)

Feature8: government debt (numerical)

Feature9: poverty rate (numerical)

Feature10: exchange rate (numerical)

Feature11: economic growth rate (numerical)

Feature12: interest rate (numerical)

Feature13: inflation rate(CPI) (numerical)

Feature14: national credit rating (categorical)

Target feature for regression algorithm: inflation rate(CPI) = numerical feature

Target feature for classification algorithm: national credit rating (ex. AAA, AA+) = categorical feature

• **Revalue data**

Dirty data: In the case of developing countries, it is expected that there is no numerical data or a high possibility of errors in the data itself. Such data will be treated as dirty data.

Wrong data: Since the unemployment rate data is measured differently in each country, it can be judged as wrong data that is not suitable for model predict(learning). It will be standardized as treat by giving appropriate weights.

• **Scaling**

In the case of 'minimum wage' feature data, the units are expected to be different for each country. (each country has its own unique currency unit.) Therefore, the data will undergo a scaling process to unify the units into US dollars based on exchange rates. Since the data is large in size, we plan to use a standardization scaling technique.

• **Data resource**

We will get dataset from World Bank Open data(https://data.worldbank.org/, use this site mainly), and will use https://tradingeconomics.com/country-list/rating site to get national credit rating feature data, and https://www.ilo.org/topics/wages/minimum-wages site to get minimum wage feature data.

2. Algorithms

- Must use data scaling (standardization/normalization)

- Must use 2 of the following 3 types of algorithms Regression, classification, clustering Must NOT use algorithms that were not taught in this course.

• **Regression algorithm**

We will predict 'inflation rate(CPI)' by other independent features(variables). In order to derive CPI prediction data which consist with numerical values, we decide it is appropriate to use a regression algorithm that is effective in finding linear relationships between datas. Therefore, we plan to proceed with learning and prediction based on a linear regression algorithm.

• **Classification algorithm**

We will predict 'national credit rating' by other independent features. In the case of national credit ratings, they are expressed by categorizing them into classes (such as AAA and AA+ etc). Therefore, we decided that a classification algorithm would be appropriate to predict data values with these classes, and we planned to use it to predict national credit ratings for the second time. Visualize such categorical data in a graph, it is appropriate to convert it into numerical data, so we expect to additionally perform feature engineering using the OrdinalEncoder technique.

3. Evaluation

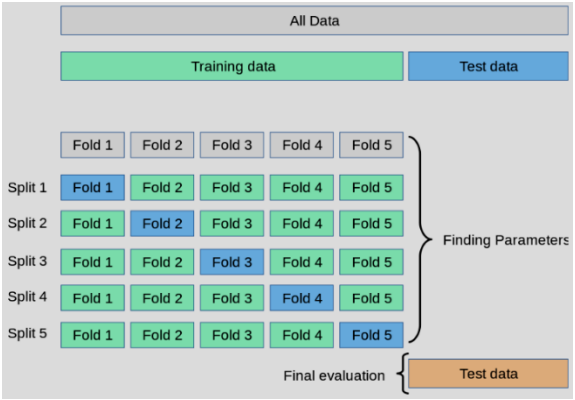
- Must use k-fold cross validation for testing classification models

• **Regression algorithm: based on MSE**

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Since regression problems are essentially about numerical differences, numerical indicators MSE, which sensitively reflect the error size, are natural and appropriate. MSE penalizes larger errors by squaring them, and it penalizes positive or negative errors (regardless of direction). Therefore, MSE can be used to precisely evaluate how close the model's predicted value is to the actual value.

• **Classification algorithm: use k-fold cross validation**



Classification is important not only for predicting each one and seeing if it is correct or not, but also for evaluating how well the model can generalize to the entire data, so the validation process is even more essential. Out of k-fold (k data subsets), (k-1) are used for training and remaining 1 is used for validation. This process is repeated k times and the results are averaged. Since all data is used for both training and validation, it can prevent good results for a specific data set from accidentally occurring (making it possible to determine actual generalization performance).