Data and Model Report

Fabian Juarez, Surya Sama

May, 2022

Introduction

As described in the project proposal, we are trying to test our hypothesis regarding the convergence of the performance metrics on various Machine learning algorithms. So we categorized the algorithms into 3 different categories:

- ***Classification Algorithms:*** The algorithms where the target variable is categorical. And the algorithms we would like to test our hypothesis on are Logistic Regression, Support Vector Classifier, KNN Classifier, Random Forests Classifier, Decision tree classifier, and the Naive Bayes Classifier. And the performance metrics we would like to analyze are Accuracy score, Precision, Recall, F1 score, logarithmic loss and ROC AUC.
- ***Continuous and Regression based algorithms***: The algorithms where the target variable is continuous and numerical. We want to test if the algorithms like Ridge, Lasso, Elastic net , Decision tree regressor, KNN regressor and the Random Forests regressors follow our hypothesis pertaining to the performance metrics like the root mean squared error (RMSE), Mean squared error (MSE), mean absolute error (MAE), etc.
- ***Deep learning algorithms:*** We are planning to apply different deep learning classifiers and regressors to the datasets used for the classification algorithms and the regression based algorithms described above. We would like to use multi layer perceptrons, Neural networks and Convolutional Neural Networks for the estimation and classification to derive the performance metrics from each of these models. And then test our hypothesis.

Description and Data

We would like to use different datasets for each of the categories described above.

I. Macroeconomic Country Data

Source: The Conference Board Total Economy Database™ - Output, Labor and Labor Productivity, 1950-2021.

The Conference Board. (August 2021). *The Conference Board Total Economy Database™*. Downloaded from:

https://www.conference-board.org/data/economydatabase/total-economy-database-productivity-growthaccounting

Dataset Description: This dataset features 16 macroeconomic variables of 130 countries from 1950 to 2021 plus categorical variables that specify the region and country. The target was chosen to be GDP per capita growth, a variable that shows economic growth in countries.

This file contains time series data on Gross Domestic Product (GDP), Population, Employment, Total Hours Worked, Per Capita Income and Labor Productivity (measured as GDP per Person Employed and GDP per Hour Worked). Data is available for 130 countries, plus a second version of Chinese data based on alternative data and a second version of US GDP data based on alternative ICT investment price deflators, covering the period 1950-2021.

The downloaded excel document was converted to a csv file and the first three rows were eliminated so that only important data remained. The resulting data set's shape is (2128, 77) comprising 163,859 data points.

Data cleansing

- Transformation to a panel dataset: using the pandas melt and pivot functions the dataset was converted into a panel data set (9576, 16) in which each row represents a combination of country-year and each column is one of the 16 macroeconomic variables.
- Handling of missing values: a row-wise removal was done in order to remove all of the missing values (50% of rows contained at least one missing value) and avoid inserting bias. This left a dataset of 4,814 rows and 16 columns[1].
- The dataset then was divided into the explanatory variables X and the dependent variable Y (which is GDP per capita growth).

Data transformation: a principal components analysis (PCA)[2] was performed on the dataset to use the multicollinearity present (shown in the correlation plot) to produce a set of uncorrelated variables and to reduce the number of predictors from 16 to 4 (explaining 87% of the variance), using the Kaiser's criterion line as a threshold (see graphs in Jupyter Notebook).

[1] If there is at least one missing value in one of the 16 variables in a specific year-country observation, the entire row information was removed.

[2] PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

II. Titanic Data

Source: Titanic - Machine Learning from Disaster competition hosted on Kaggle.
https://www.kaggle.com/competitions/titanic/data?select=train.csv

Dataset Description: This dataset features 12 variables which describe the individuals onboard Titanic. The target variable being the Survival status of the individuals. Here we are using only the training dataset provided by the hosts of the competition on kaggle, since the test data does not have the target variable in it. So, we are assuming that the given training dataset as the complete dataset and then perform our analysis by splitting the dataset into train and test datasets. The dataset has the data for 890 individuals.

The variables are as follows:

- Survival = Survival, [0 = No, 1 = Yes]
- Pclass = Ticket class, [1 = 1st, 2 = 2nd, 3 = 3rd]
- Sex = Sex [Male, Female]
- Age = Age in years
- Sibsp = # of siblings / spouses aboard the Titanic
- Parch = # of parents / children aboard the Titanic
- Ticket = Ticket number
- Fare = Passenger fare
- Cabin = Cabin number
- Embarked = Port of Embarkation, [C = Cherbourg, Q = Queenstown, S = Southampton]

Data Cleansing

The dataset which we downloaded was pretty clean and no major cleaning was required to create the final dataset. While examining the variables, we transformed some of the numerical variables to categorical to help us improve the prediction. So, we transformed the Sex variable which was an object to a dummy variable with 1 being female and 0 being male. The Age variable had around 177 missing values. So, we used the KNN Imputer to impute the missing data. This algorithm tries to fill the missing values in a specific row by computing the euclidean distance between the row and remaining rows in the dataset. And the row which is the nearest is used to fill in the missing values for Age.

Data Transformation

We transformed the Age and the Fare variables into categorical variables. In order to transform the variables, we first examined the percentage of people who survived for different age and fare groups. Then we categorized the Age into 5 different groups:

- Age < 16: 0
- 16 < Age <= 32: 1
- 32 < Age <= 48: 2
- 48 < Age <= 64: 3
- Age > 64: 4

And for fares, we examined the percentage of people who survived for different intervals of Fares and transformed the fare variable as follows:

- Fare <= 14.454: 0
- 14.454 < Fare <= 128.082: 1
- Fare > 128.082: 2

The code for the cleaning and transformation parts is available in the **Jupyter Notebook submitted along with the Data Report**.

Modeling Approach for the Datasets

Macroeconomic Country Data	Titanic Data
Ridge, Lasso, Elastic net , Decision tree regressor, KNN regressor and Random Forest regressor, and Deep learning algorithms like CNN, MLP.	Logistic Regression, Support Vector Classifier, KNN Classifier, Random Forests Classifier, Decision tree classifier, and the Naive Bayes Classifier, and classification using the Deep learning algorithms like MLP, and other Neural Networks.