Life expectancy prediction Summary report

Michael Leung

michaelleung341@gmail.com

Life expectancy prediction Summary report

Michael Leung

Introduction

My project is life expectancy prediction, which I want to find out what are the key factors that affect the life expectancy through machine learning.

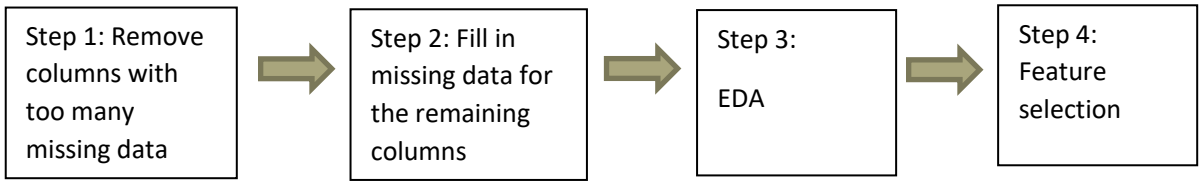
I think the project has its unique value because it can affect us on different level.

- For personal, we can plan for our future and maintain our quality of life.
- For company or organization, they can provide suitable products and services such as investment plan or saving plans.
- For governments, they can implement policies base on those features to extend the life expectancy of their citizens or deal with the aging population problem.

Details on dataset

- Source of the dataset: The world bank
- Name of the dataset: World development indicator 2022(It will update new data annually)
- File type: csv file (also with csv files that are the footnotes about the dataset)
- Reliability: High (It is from known world class organization and the data of all features are citated and with footnotes on other files.

Summary of cleaning and preprocessing



Step 1: I removed columns with over 40% missing values because I think the model cannot learn much from those data.

Step2: I grouped the data rows by the GDP per capita into 10 groups, then I filled in the missing data of the median of each column of each group.

Step 3: Exploratory data analysis (DEA)

From the EDA, there are some insights from the data:

- The distribution of life expectancy is slightly left skewed, this may due to the life expectancy data were collected annually and the its average is shifting every year. (Refer to graph 1)
- The average life expectancy over the world increases steady for the past 60 years with around 40% increase. (Refer to graph 2)
- The group with higher GDP per capita usually has longer life expectancy. (Refer to graph 3)
- There are some features that have linear correlation with life expectancy such as death rate, year, rural/urban population, etc.

Step 4: After EDA, I only selected the environmental features and removed the features that may be used to calculate the life expectancy (e.g. death rare, population of different age, etc.)

Modeling process

I split the data set into train (for model building), validation (for hyperparameter optimization) and test (to get unbiased predicting power of the model) set.

Although I did not treat the project as a time series question, I will use the old data as the train set, latest data as the test set and the data in the middle as the validation set as the data from the future may have some predictive power for the past data.

To be short, this is the accuracy score of all the models I had built for this project.

Model	R^2(with all features)	R^2(With 10 features)
Linear regression	N.A.	0.366
Ridge	N.A.	0.364
KNN regression	0.229	0.608
SVR	N.A.	0.428
Decision tree regression	0.670	0.675
Random forest regression	0.733	0.714

- For the first model that I built was the linear regression, which was the simplest machine learning model that will find out a regression formula that has the least R^2

with all the data points. I treated it as a base line of the prediction power of all the models. The R^2 score for the model was 0.366.

- For the ridge model, it is a linear regression model with regulation, I tried to modify the regularization strength of the model, but it has nearly the same R^2 as the linear regression model.
- Since the linear regression model didn't perform well on predicting the life expectancy, I tried to build some non-linear regression models such as KNN regression and SVR models. They had higher R^2 than linear regression model but didn't perform well when I put all the features to the model.
- After that, I built a decision tree regression model. it had significant higher R^2 (0.670) than the above models and it had similar predicting predict power for the model with all features and only 10 features.
- The final model I built was random forest model. It has the best R^2 (0.733) of the project after hyper. Even with 10 features, it can still have 97% R^2 as the original model.

Model evaluation

- Form the model with highest performance (random forest model), the top 10 important can be summarized in to three categories – infrastructure, population and economy.
- Also, I tried to visualize the prediction of model for different countries. I found that the model can predict the life expectancy within 5 years of the real data (Refer to graph 4).
- The model seems to have higher accuracy on predicting the life expectancy on the less developed countries.

Conclusion

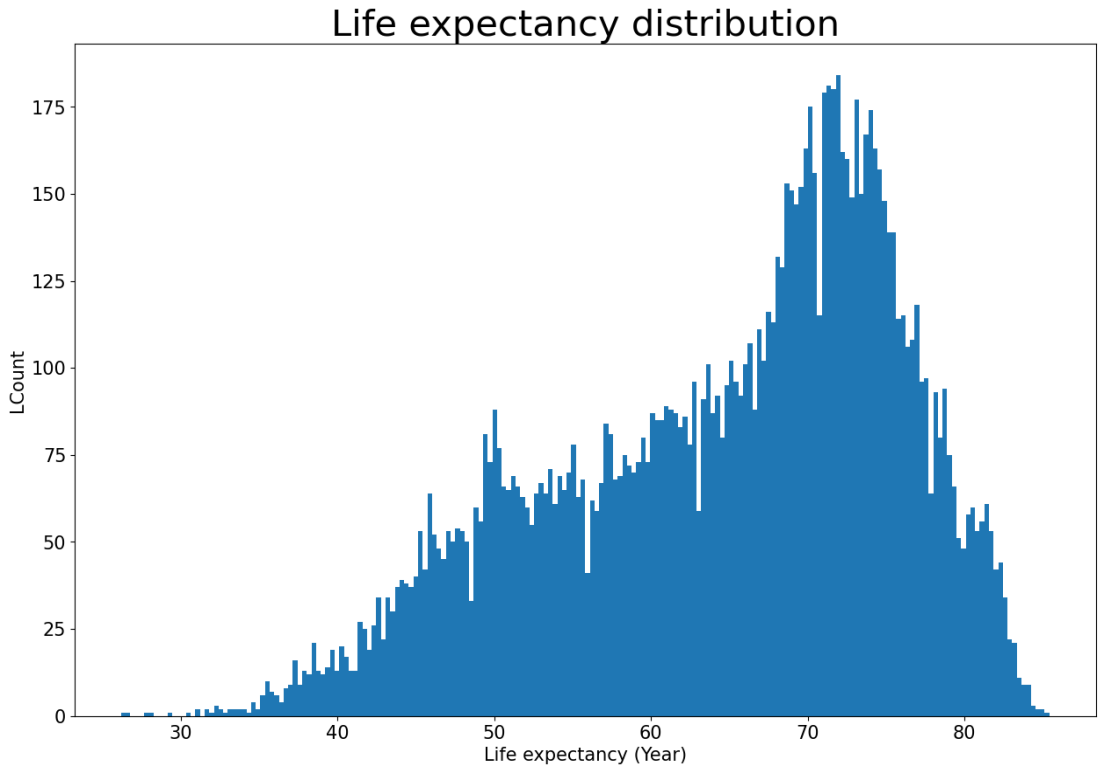
- The life expectancy increases over the past 60 years.
- Countries with higher GDP per capita usually have longer life expectancy.
- For the most accurate model, random forest regression model, the most important features are from infrastructure, population and economy categories.

Future plan

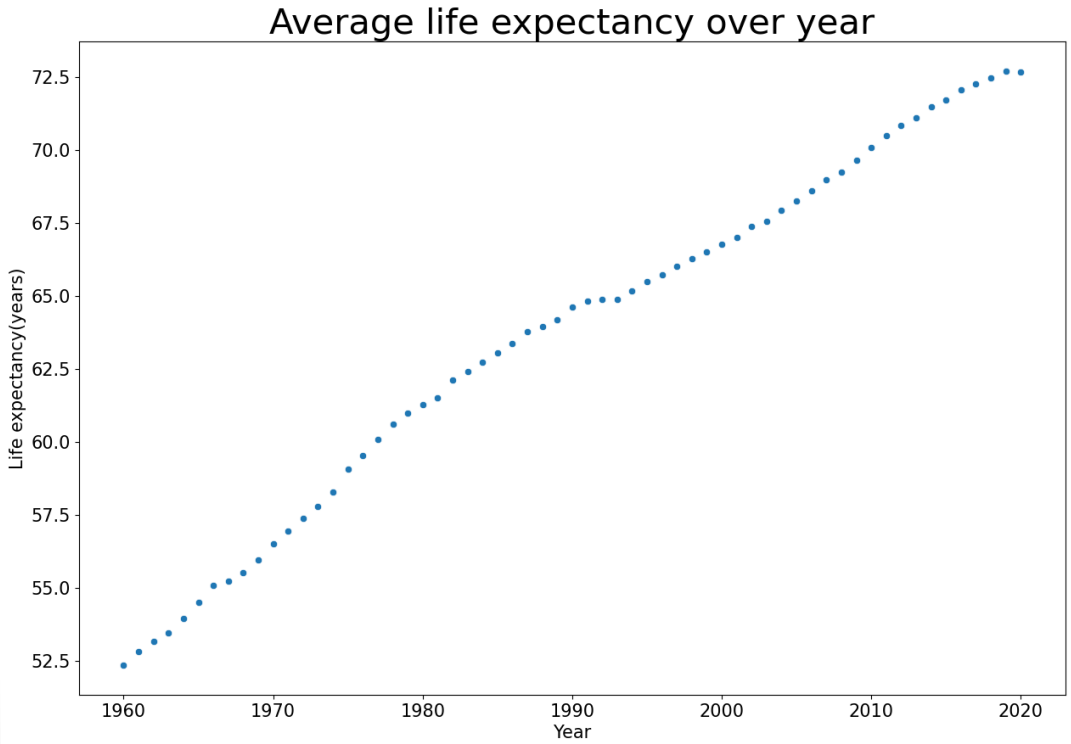
- Find out a better dataset, which has fewer missing data and is a larger dataset to build a more accurate model.
- Instead of the country level dataset, find out some datasets to predict the life expectancy in personal level.
- Try to build some deep learning models to see the predictive power.
- Search for more information on how infrastructure, population and economy affect the life expectancy and try to come out of some ideas on lengthen life expectancy.

Graph

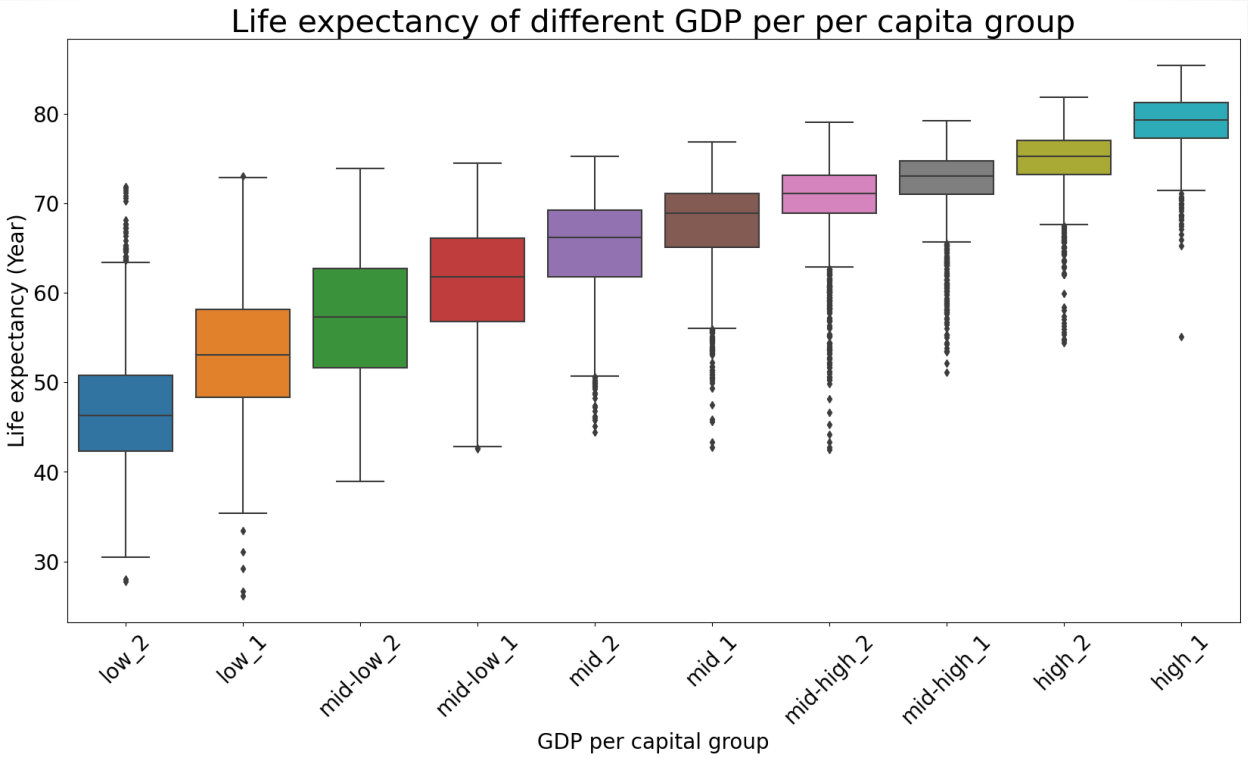
Graph 1: Life expectancy distribution



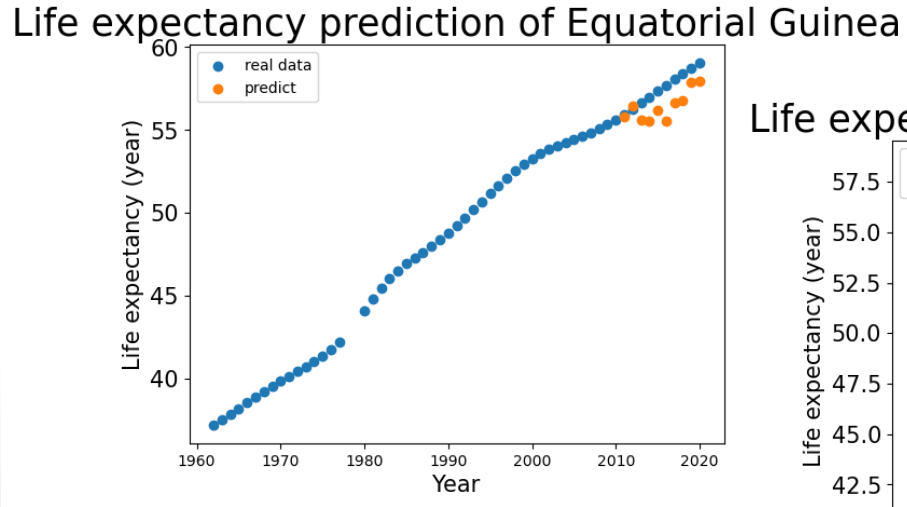
Graph 2: Average life expectancy over year

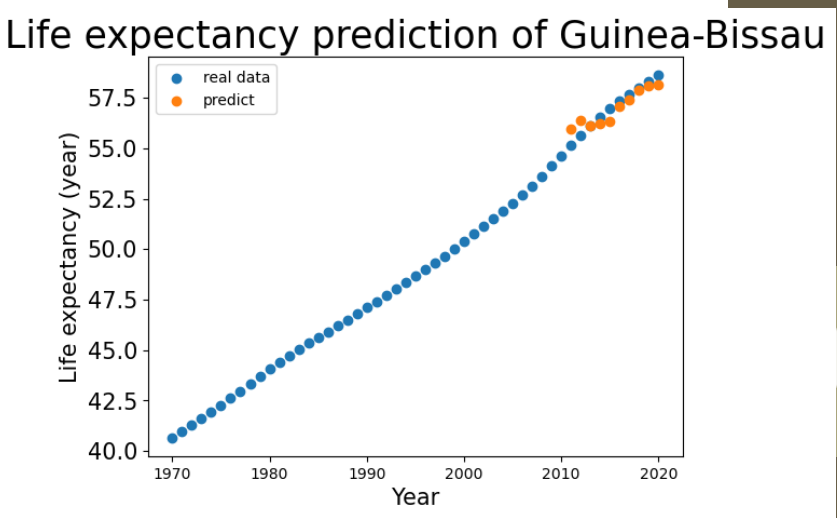


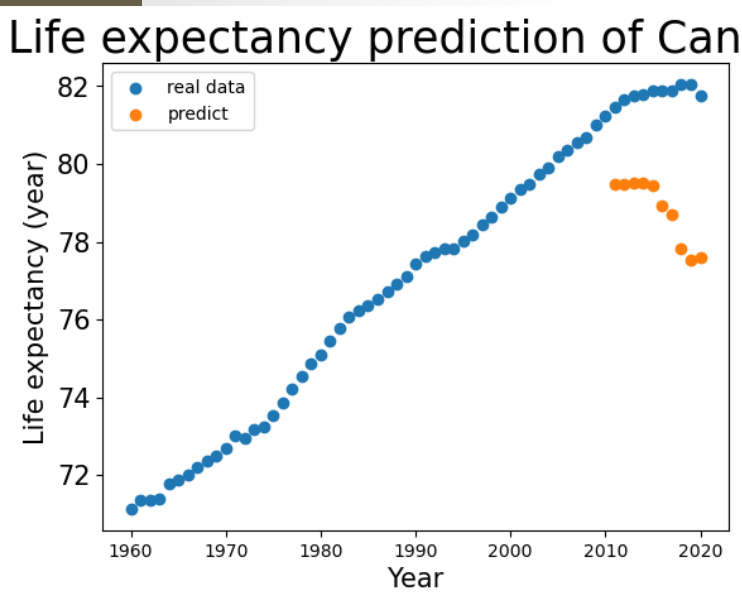
Graph 3: boxplot of life expectancy of different GDP per capita group

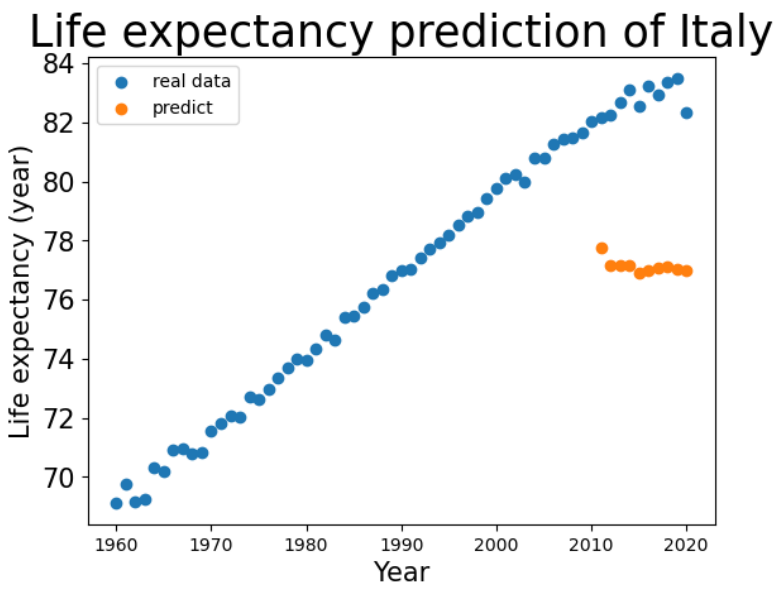


Graph 4: Life expectancy of some countries









Reference

Reference 1: World Development Indicators

https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators

Reference 2: How to select features after running DecisionTreeRegressor algorithm?

https://stackoverflow.com/questions/68914158/how-to-select-features-after-running-decisiontreeregressor-algorithm

Reference 3: Feature Selection Using Random forest

https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f