OIM 454: Life Expectancy Prediction Project

Albert Wong, Connor Kucich, Christopher Norgren

Table of Contents

Project Description: 2

Data Preprocessing: 3-4

Models and Analyses: 5

Results and Discussion: 6

Summary: 7

References: 8

Project Description

This report aims to outline the data, models, analysis, and results of two machine learning models trained on World Bank data. The two machine learning models that were used are a Multiple linear regression model and a K-Nearest Neighbors (KNN) model. The issue that was addressed using the World Bank dataset and these models are determining what predictors influence life expectancy. Through our development and analysis, we aim to uncover actionable insights that policymakers and organizations could potentially use to improve health outcomes worldwide.

The dataset for this project is derived from the World Bank, containing various socio-economic, demographic, and geographic indicators from countries across the globe. The primary objective is to build predictive models that estimate the factors that influence life expectancy. The predictor variables include data such as income level, world region, GDP, population data, and many others.

According to a research article titled “The Economics of Longevity - An Introduction”, life expectancy has increased roughly 2.5 years every decade since 1840, reaching 71 today. This dramatic increase in life expectancy has brought with it numerous economic benefits and helped contribute to the economic prosperity the world has experienced over the past 100 years. The insights generated from this report can be used to help governments, nonprofits, and healthcare entities make decisions that have the most positive impact on life expectancy to continue its increase across the globe. Through the use of machine learning models, we are able to support data-driven decision making to improve public health.

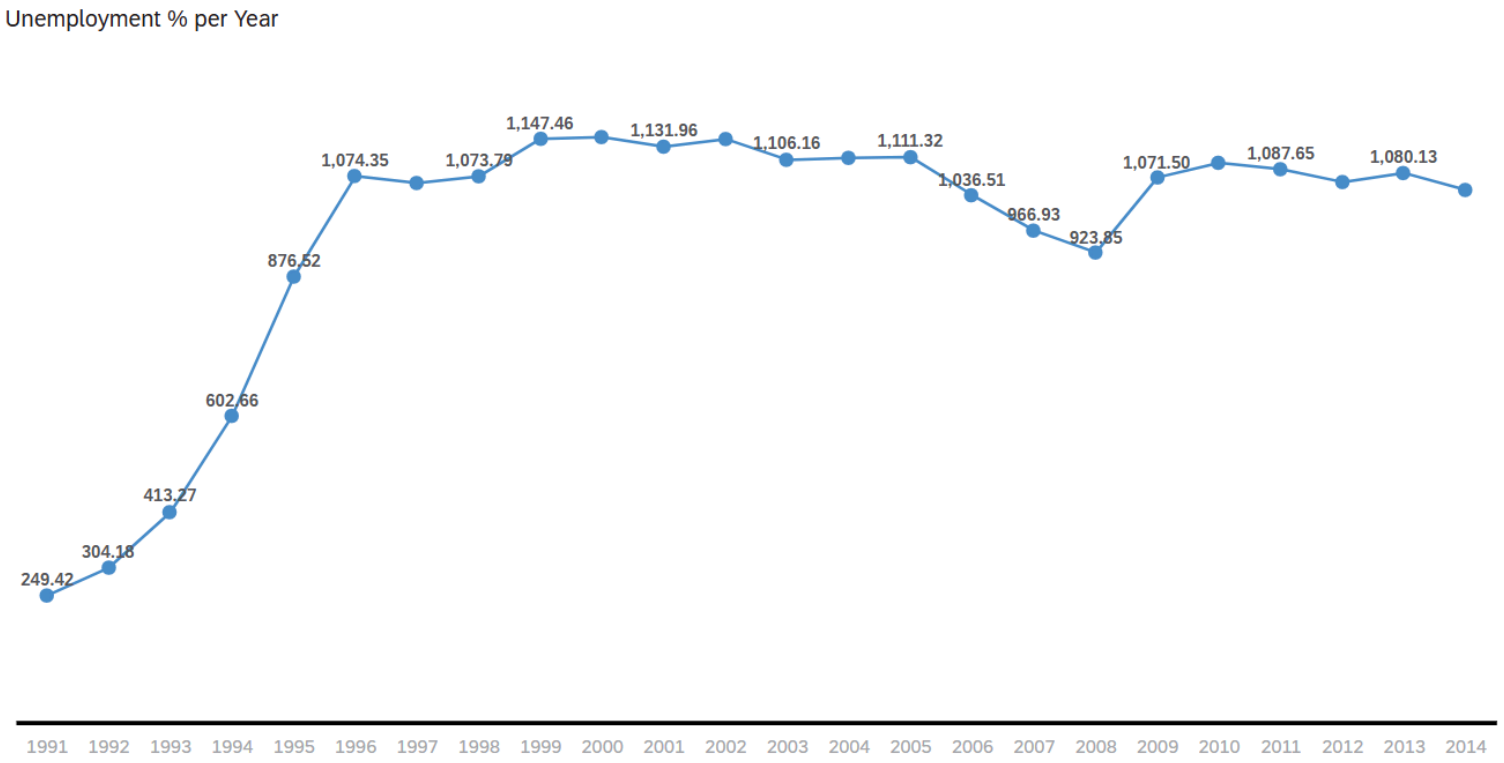
As stated previously, the two machine learning models that are used include a Multiple linear regression model and a K-Nearest Neighbors model.

Multiple Linear Regression - To establish a model that leverages interpretability to understand the relationship between predictor variables and life expectancy

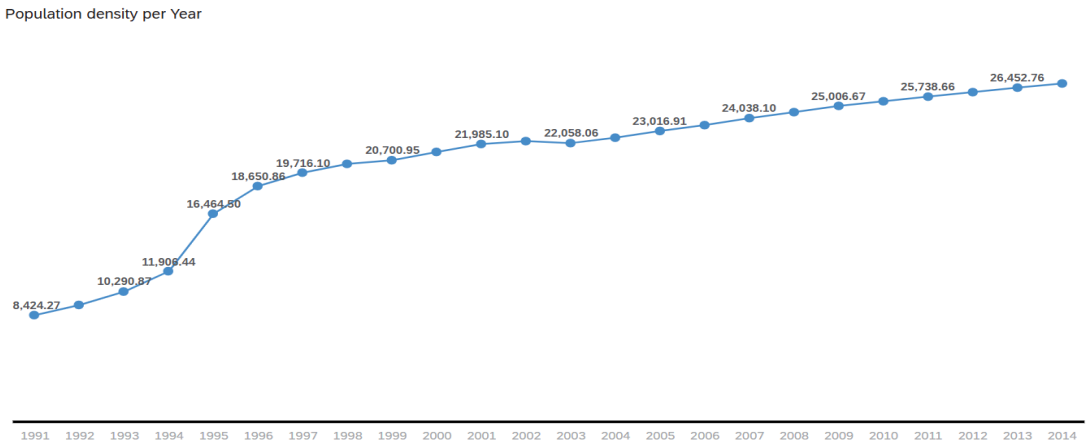
K-Nearest Neighbors (KNN) - To explore non-linear relationships and patterns that may not be captured by linear regression

Data Preprocessing

With missing data or data errors, we used power query to remove any cells that were blank or empty. The original data set had 12,450 rows. To gather a random sample, we used 60% of data. In excel, we used the =RAND() function, then we used the number filters, less than 0.6. This resulted in us having a total of 2,776 rows. Here are data visualizations for 2 important variables in the analysis with a description.

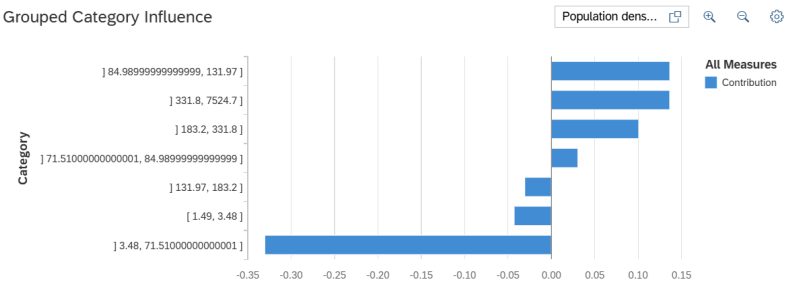


What this line chart shows is the total unemployment percentage per year. There was a great increase in unemployment starting from 1991 to 1996. After that it stays at a steady line until 2005 where it decreases a little to until 2008. After that, we can see it go back to the steady line like from 1996 to 2005.



With this chart, it is about total population growth density per year. Similarly to the chart above about the total unemployment percentage per year, the greatest amount of increase is from the years 1991 to 1995. What is evident from this chart is that almost every year, the population density is increasing with a positive growth rate.

We also created a correlation chart showing the relationship between a predictor variable and the dependent variable with a description.



This chart shows that countries with a higher population density tend to have a longer life expectancy. This can be seen as each bar on the graph represents a range of population density, while the x axis shows the influence it had on our outcome variable of life expectancy. As the chart demonstrates, the higher ranges of population density tend to have a positive life expectancy influence (leads to a higher life expectancy), while countries with a lower population density tend to have a negative life expectancy influence (leads to a lower life expectancy.

When cleaning our data, we transformed the region text data into 7 dummy variables for each region, each region being; East Asia & Pacific, Europe and Central Asia, Latin America & Caribbean, Middle East & North Africa, North America, South Asia, and Sub Saharan Africa. If the country was in that region, then it would be given a dummy variable of 1 and if not, it would be given a 0. Additionally we deleted the Country Code as it is the same thing as Country Name and done to avoid multicollinearity. We also transformed the income group to category scores as it was a text variable and would mess up our model; Low Income = 1, Lower middle income = 2, Upper middle income = 3, High income:OECD = 4, High income:nonOECD = 4.

Models and Analyses

The first technique used was a Multiple Linear Regression (MLR) model done in R. Multi-linear regression is a statistical technique used to model the relationship between one dependent variable and two or more independent variables. The purpose is to predict the outcome of a dependent variable using multiple independent variables. The independent variables can be continuous or qualitative, but the dependent variable must be continuous. The end result this model is able to produce is being able determine the variation of the model and the relative contribution of each independent variable. In our case multiple linear regression models the relationship between a dependent variable (life expectancy) and multiple independent variables by assuming a linear relationship between the predictors and the target variable.

The second technique used was using the algorithm of K-nearest neighbors (KNN) done in R. This algorithm is able to predict outcomes based on the similarity of data points. It works by finding the K closest data points to a new data point, measured by distance (e.g., Euclidean distance). The value of K determines how many neighbors are considered when making a prediction. K-NN is intuitive and effective, relying on the idea that similar data points tend to have similar outcomes. In our case, life expectancy is influenced by a range of factors, and KNN can leverage these factors to make predictions without assuming a specific functional form of the relationship. Therefore we can utilize how KNN makes predictions by identifying the closest data points to a given query point to make an informed prediction.

The predictors that were included in our techniques were:
Is.DRC + Is.China + Is.Russia + Is.USA + East.Asia.&.Pacific + Europe.and.Central.Asia + Latin.America.&.Carrebian + Middle.East.&.North.Africa + North.America + South.Asia + Sub.Saharan.Africa + IncomeGroupRanking + Year + Electric.Power.Consumption + GDP + GDP.per.Capita + Individuals.using.the.Internet + Population.density + Unemployment.%

Some locations have “Is.” in the beginning. This was used for us to highlight major geographic regions. IncomeGroupRanking ranked income from low=1 to high income=4. The year spanned from 1991 to 2014. The rest of the predictors varied in value as they were dependent variables.

Predictors that were disincluded were Life expectancy, Birth rate, Death rate, and Infant mortality rate. Life expectancy was not included as this was our outcome variable. Birth rate, Death rate, and Infant mortality rate were too obvious of predictor variables and would diminish the results because they obviously influence life expectancy.

Results + Discussion

As described above, the two models we used for this project were a multiple linear regression, which we completed in R, and a k-nearest neighbor model, which we also completed in R. We determined that the best metrics to evaluate the performance of these models were R squared, which measures the goodness of fit for the model, RSME (root mean squared error), which is an error based metric that reflects the quality of the predictions, and MAE, which is an error based metric that shows on average how far off predictions are from the actual value, regardless of sign. For both RSME and MAE a lower number is better, but for R squared a higher number would indicate a better goodness of fit.

For the multiple linear regression, in the training data the model provided us an R squared of 0.8493, however it did not provide us MAE or RSME. The multiple linear regression's validation data had a R squared value of 0.8250452, a RSME of 3.6045709, and a MAE of 2.7230056. The multiple linear regression's test data had a R squared of 0.8291443, an RSME of 3.5239836 and a MAE of 2.6616486.

Overall, the RMSE and MAE were very low for both test and validation which is good, however the r^2 is very high for both of them, which show the model is fit very well, but at the number its at (0.82) might be fit a little too well and be a sign of overfitting. This also could be because a majority of the outcome variable (life expectancy) is around the same number. The r^2 is about the same on the test and validation data as it is on the training data, so that is more evidence toward the model not being overfit.

For the k-nearest neighbors model, we were only able to extract the metrics from our k-nearest-neighbors prediction set, and when we did, we chose to extract the same metrics for consistency, and then compare the models to see which one performs better. For the k-nearest neighbors prediction data, we got a R squared value of 0.94645118, a RSME of 0.05042133, and a MAE of 0.03488461. Once again this is a very high R squared value suggesting that this model is very well fit, perhaps overfit. This combined with the extremely low RSME and MAE lead me to believe that this model is overfit to our dataset. This is especially true when compared to the results of the multiple linear regression.

This leads us to recommend that the multiple linear regression model be used instead of the k-nearest neighbor model while using a model to evaluate this dataset. Although the k-nearest neighbors metrics are more favorable (higher R squared, lower MAE and RSME), they are suspiciously high (in the case of R squared), or suspiciously low (in the case of RSME and MAE, both suggesting that the error between predicted and actual value is less than 0.1). While the performance metrics were also high/low for the multiple linear regression, we can compare the performance on the test/validation data to the training data, and see that it performs just about the same on all of them, meaning that it is not overfit to our data there. We can not say this about the k-nearest neighbors model. One explanation for the boosted metrics is that life expectancy doesn't have a huge variance in general, it can definitely vary, dropping as low as 43 and peaking as high as 83, however in most cases it sits around 70. This means that most

of the predictions will tend to be close to correct, as most of the correct values are around the same range. Because of all this, going forward, we recommend using a multiple linear regression model to evaluate this data.

Summary

This report explored the factors influencing life expectancy using machine learning techniques applied to World Bank data. Through data preprocessing, we addressed missing values, removed redundant variables, and transformed categorical variables into dummy variables to optimize model performance. Two models, a multiple linear regression and a k-nearest neighbors (KNN) model, were built and analyzed.

The linear regression model was selected for its interpretability, allowing us to understand the linear relationships between predictors such as GDP, unemployment, and population density, and their impact on life expectancy. The KNN model was used to explore non-linear relationships, offering an alternative perspective on the dataset. Both models were evaluated using R-squared, RMSE, and MAE metrics.

Results revealed that while the KNN model showed higher R-squared and lower error metrics, its performance raised concerns of overfitting, likely due to the minimal variation in life expectancy data. Conversely, the MLR model demonstrated more reliable generalization across training, validation, and test datasets, making it the better choice for predicting life expectancy, and what we would recommend to decision makers to use in this scenario.

Key insights from this project highlight the significant role of socio-economic factors, such as income level and geographic region, in shaping life expectancy. These findings provide actionable recommendations for policymakers and organizations aiming to improve global health outcomes. Overall, the project emphasized the importance of careful model selection, the value of interpretability in machine learning, and the potential of data-driven strategies to address complex societal issues.

References

Scott, Andrew J. (2023). The Economics of Longevity – An Introduction. *The Journal of the Economics of Ageing*, 24. https://www.sciencedirect.com/science/article/pii/S2212828X22000718