Gender Inequality - Final Report

Tim Cross

July 16, 2025

Abstract

In this report I detail my analysis of World Bank data on Gender statistics to predict and model the level of gender equality for 82 countries using machine learning models. I chose the target variable 'Female share of employment in senior and middle management (%)' as an indicator for gender equality. For the vast majority of countries the target variable takes values below 40% (75th percentile) . In exploratory analyses, I find no strong correlations between the ∼160 variables and the target variable (maximum absolute correlation of ~ 0.3). The data is split into a train and test set (30-70 split), and multiple machine learning methods based on decision tree or linear regression are trained on the data. Most models have a tendency to over-fit the data, as they show very high accuracy and R^2 for the train set but much lower values for the test set. Even with hyper parameter tuning and data compression using principal component analysis (PCA) this issue was not resolved. Further research into this subject is needed, however the current analysis suggests that either the variables used to build the model are not good indicators for the target variable or a more sophisticated model is needed to describe the relation between them.

1 Data

The dataset, Gender Statistics, is taken from the World Banks databank and includes 1162 features on key gender topics. Themes include demographics, education, health, labour force, and political participation. The data also includes more general data such as Gross Domestic Product (GDP), Gross National Income (GNI), birth rate and population. Due to the scope of this project I decided to just use data from 2019. This year was chosen as it was the most recent year which had complete data for a large number of countries, and avoided the COVID period, which could have the effect of distorting some features. Once this year was selected it reduced the features to 790. To choose the target variable from this massive dataset I considered variables where a large number of country had data for and could be used to indicated gender inequality in a quantitative way. The variable that I chose is the percentage of female share of employment in senior and middle management, as traditionally these roles are dominated by men.

My initial hypothesis before any data analysis was that countries with higher values in the target variable generally have higher GDP, and have more socialist leaning policies e.g. Iceland, France and Scandinavia. But as we will see in the coming sections this hypothesis is not supported by the data.

2 Data Cleaning

I used the Pandas library within Python for data cleaning and to prepare the data for further processing and analysis. First I pivoted the data to allow the features to occupy columns, with the countries being the index rows. Next, I removed all countries with a population under 1 million. This was for a number of reasons: these countries generally had a lot of missing data, and there were many samples which were not countries e.g. principalities or overseas territories. Next I removed all countries without data in the target variable, which left 82 countries. Finally I removed any features which contained missing data, which reduced the features to 163. I decided against imputing the data, due to the quantity of missing data, and the difficulty in imputing from other countries data. The final cleaned data includes 82 countries with 163 features.

3 Exploratory Data Analysis

A histogram of the percentage of female share of employment in senior and middle management, the target variable, is shows in Figure 1. The middle 50% of the countries have between 28.09% and 39.16% share of female senior and middle management, and the mean value for this variable is 32.6%.

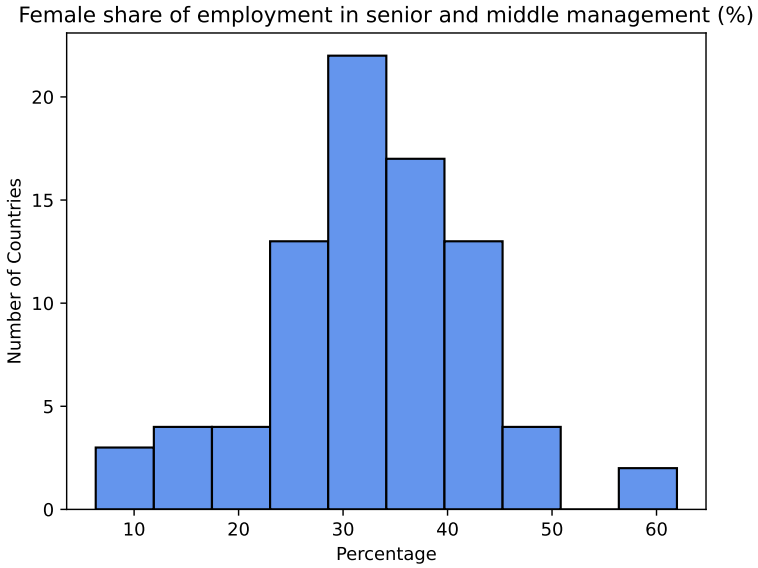

Figure 1: Percentage of female share of employment in senior and middle management in 82 countries in 2019.

To investigate my hypothesis I plotted GDP per capita against the target variable in Figure 2. Interestingly, this showed that the countries with the highest percentage as having a low GDP per capita. The countries with higher GDP per capita tend to aggregate towards the middle of the range for the target variable. This is contrary to what I expected and disproves my hypothesis.

Looking at the relation between the target variable and life expectancy at birth in Figure 3 we don't see a clear correlation either. Indicating that wealth and social care does not impact the target variable directly.

I investigated the correlations between the target variable and all other features using a correlation table. None of the features showed strong positive or negative correlations with the target, with the strongest being A woman can work in an industrial job in the same way as a man (1=yes; 0=no) at 0.33 and Population ages 05-09, female at -0.26. Other population related variables show a similar negative correlation with the target variable.

As the number of variables is very large, I do not include a correlation matrix here. However upon analysing correlations between the features I realised that some of them are highly correlated which may cause problems in training the models.

4 Methodology - Machine Learning Models

I used two groups of machine learning models based on decision tree and linear regression. The decision tree based models include:

- Decision Tree
- Random Forest
- Histogram-based Gradient Boost
- eXtreme Gradient Boosting: XG Boost

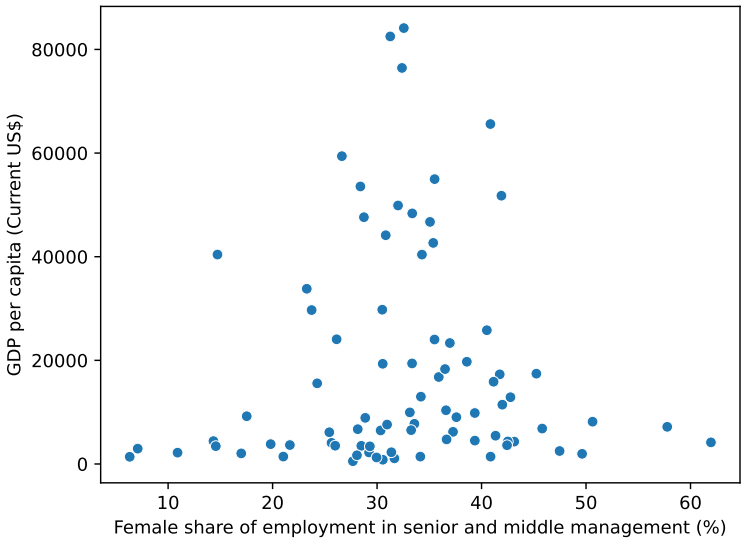

Figure 2: GDP per capita versus Percentage of female share of employment in senior and middle management in 82 countries in 2019.

and the regression based models are:

- Linear Regression
- Linear Regression - Lasso
- Linear Regression - Ridge

and a neural network model:

- Multi-layer Perception Regressor

I applied hyper parameter tuning with cross validation on the best models (Random Forest and XG Boost) to improve their performance. To sample the hyper parameters, I used Bayesian Hyper Parameter optimization, a method that is far more efficient than a grid search.

Due to the large number of features I apply Principal Component Analysis (PCA) for data compression. The data was standardised before PCA was applied to it. Figure 4 shows the variance associated with the first 15 principal components. As the figure shows the vast majority of the variance is captured by the first component. I choose to use the first 10 PCs as my compressed data as the variance is reduced substantially below this point (2.7 for the 10th PC versus 60.6 for the first PC).

Two data sets are used for my analysis:

- The original cleaned data with 163 features
- The compressed data including the first 10 principal components

In preparation for the modelling each data set was divided into a train-test split, selecting 30% for test and 70% for train using random selection.

5 Modelling Results

To evaluate the models I use R^2 (Table 1) and Mean Squared Error (Table 2). The tree based models generally perform better than the regression based ones. In particular, XG Boost is the best performing model against the original test data with an R^2 score of 0.42, followed by Random Forest with a score of 0.286 and Histogram-based Gradient Boost at 0.141. The training scores of some of the models are

Figure 3: Life expectancy at birth versus Percentage of female share of employment in senior and middle management in 82 countries in 2019.

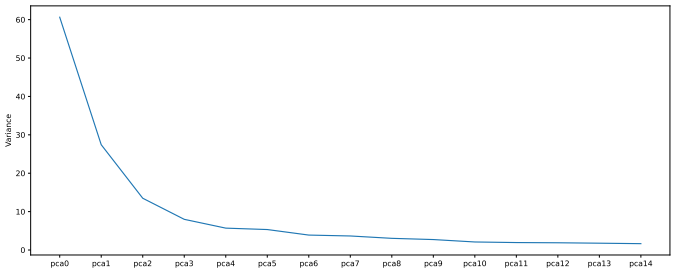

Figure 4: Variance captured by the first 15 Principal Components.

very close to 1, indicating that they have over fitted the data, especially when we see that the test values are much lower.

The regression based models performed very poorly on the original data, with the exception of linear regression all other models fail to fit the training set and show negative and in some cases extremely low scores against the test data.

Compressing the data with PCA improved the performance of the regression based methods but did not help with the tree based method except for decision tree.

Note that the MLP Regressor model will likely need hyper parameter tuning to perform better and was added as a last resort to find a more suitable method for modeling this data.

Hyper-parameter-tuning (HPT) was only done for Random Forest and XG boost the best performing models. However this did not improve the performance of the model. As can be seen the training scores for the tuned models are much lower than the original also solidifiying the conclusion that the models have over fitted the data.

6 Conclusion

In this report I show an analysis of gender statistics data by World Bank. I chose the 'percentage of female share of employment in senior and middle management' as my target variable to indicate

ML Model	Original (Train)	Original (Test)	PCA (test)
Decision Tree	1.0	-0.497	-0.164
Random Forest	0.840	0.286	0.133
Random Forest - HPT	0.280	0.138	
Hist Gradient	0.539	0.141	
XG Boost	1.0	0.420	-0.55
XG Boost - HPT	0.273	-0.028	
Linear Regression	1.0	-37688	-0.589
Lasso	0.707	-14.230	-0.010
Ridge	0.362	-198.144	-0.587
MLP Regressor	-3.000	-7.615	

Table 1: ML Model Results - R^2 Score. The training scores are only shown for the original data. The test scores are shown for both the original and the PCA data. The models with an HPT suffix show the results after hyper parameter tuning is performed. These results are only shown for the original test data.

ML Model	Original (Train)	Original (Test)	PCA (test)
Decision Tree	0	174	135
Random Forest	15	83	101
Hist Gradient	43	100	
XG Boost	2	67	123
Linear Regression	2	4403150	186
Lasso	28	1779	27
Ridge	60	23265	185
MLP Regressor	3e18	9e18	

Table 2: ML Model Results - Mean Squared Error. The training scores are only shown for the original data. The test scores are shown for both the original and the PCA data.

gender equality. After cleaning the data 82 countries were selected and the data was limited to 2019 to limit the scope of the project. Explo

The results from the models show that with this data the models did not do a good job of predicting the target variable. With quite a few models: Decision Tree, Random Forest, XG Boost and Linear Regression, the models performed strongly on the train data, but poorly on the test data indicating over-fitting.

The use of hyper-parameter-tuning on both Random Forest and XG Boost both had a negative effect on the results. This could be due to the cross validation that is carried out during the tuning, which would reduce the amount of data being worked with, and give the model less potential for a better prediction, even with improved hyper parameters.

The Principal Component Analysis also had a negative effect on the success of all models. I am currently unclear why this happened, and will need to carry out more investigation to come to a conclusion.

Looking at the results overall, it seems that either the features selected are not good indicators of the target variable, or the relationships between the features at the target are non-linear, and the models I selected are not suitable for modelling them. It could also be that factors linked equality and not so easily recorded by standard measurements used by countries, and could be more subtle and nuanced, connected to the culture and society of the countries in question. There may be factors connecting countries, which are either very hard to ascertain through data, or the factors may be different for each country.

Going forward with this project, and to try to make progress with modelling the relationships, I have the following ideas for future direction:

- Use data with different features. The data used was in the main related to gender, but it may be beneficial to find some more general data relating to the countries.
- Increase number of years, to increase the overall data range

- Use different machine learning models, which may deal better with the non-linear nature of the relationships.