PREDICTION OF THE PROPORTION OF WOMEN IN MANAGEMENT POSITIONS IN UPCOMING YEARS

Fatma Betül FİŞNE

040200204

EHB420E

10342

Assis. Prof. Dr. Onur ERGEN

21.01.2024

Repository: https://github.com/betulfis/project-ANN

ABSTRACT

Despite the increasing education levels of women and their growing participation in the workforce, the representation of women in managerial positions remains disproportionately low. This study aims to estimate the future number of female employees in managerial roles, with the anticipation that observing an increase in women's managerial roles will motivate those still in education. While various factors influence the representation of women in managerial positions, obtaining comprehensive data is challenging. Thus, the study utilizes a simplified approach, relying on the percentage of women in managerial positions from diverse countries over several years.

Data was collected from the Kaggle dataset "Women Managers," encompassing percentages of women in managerial roles across 128 countries from 1991 to 2022. Since Türkiye was not included, official data for 2020, 2021 and 2022 was added. Following data cleaning, the dataset was split into features (country and year) and output (percentage of women in management). Machine learning models were employed to predict future percentages.

Linear regression, polynomial regression, decision tree regression, random forest regression, XGB regression, lasso regression, and ridge regression were considered. The XGBoost regression model exhibited the lowest mean squared error, outperforming other models. The study concludes that ensemble learning methods, particularly XGBoost, provide more accurate predictions for this dataset. The trained XGBoost model was used to predict the percentage of female managers in various countries for the year 2023.

INTRODUCTION

Although the education level of women and their participation in the workforce have increased over the years, the rate of women working in managerial-level jobs still remains low [1]. It is aimed to estimate the number of female employees working in managerial positions in the following years. Seeing the increase in the number of women in managerial positions will increase the motivation of women still receiving education. There are many factors that affect the rate of women working in managerial positions. For an accurate prediction, we need to have data such as the qualification level of women who worked in these positions in the past years, the education and development levels of the countries, and the distribution of sectors in the country in those years, etc. However, these data are difficult to obtain. For this reason, the issue was simplified and the percentage of women working in managerial positions for different years from many countries was used as data.

DATA COLLECTION

First, appropriate data had to be found. For this, the "Women Managers" data set from the Kaggle site was used [2]. This data set was created by combining data from many sources. The data set includes the percentage of women in managerial positions in different years of countries from many different regions of the world. There are around 1500 data from 128 different countries. The oldest data is from 1991 and the newest data is from 2022. Since Türkiye is not included in the data set, it was added by the research, but there is only official data for Türkiye for the years 2020, 2021 and 2022. For some countries, data was only available for a single year, so these countries were removed from the data set. The countries and average proportion of women in management positions according to the country can be seen from Figure 1.

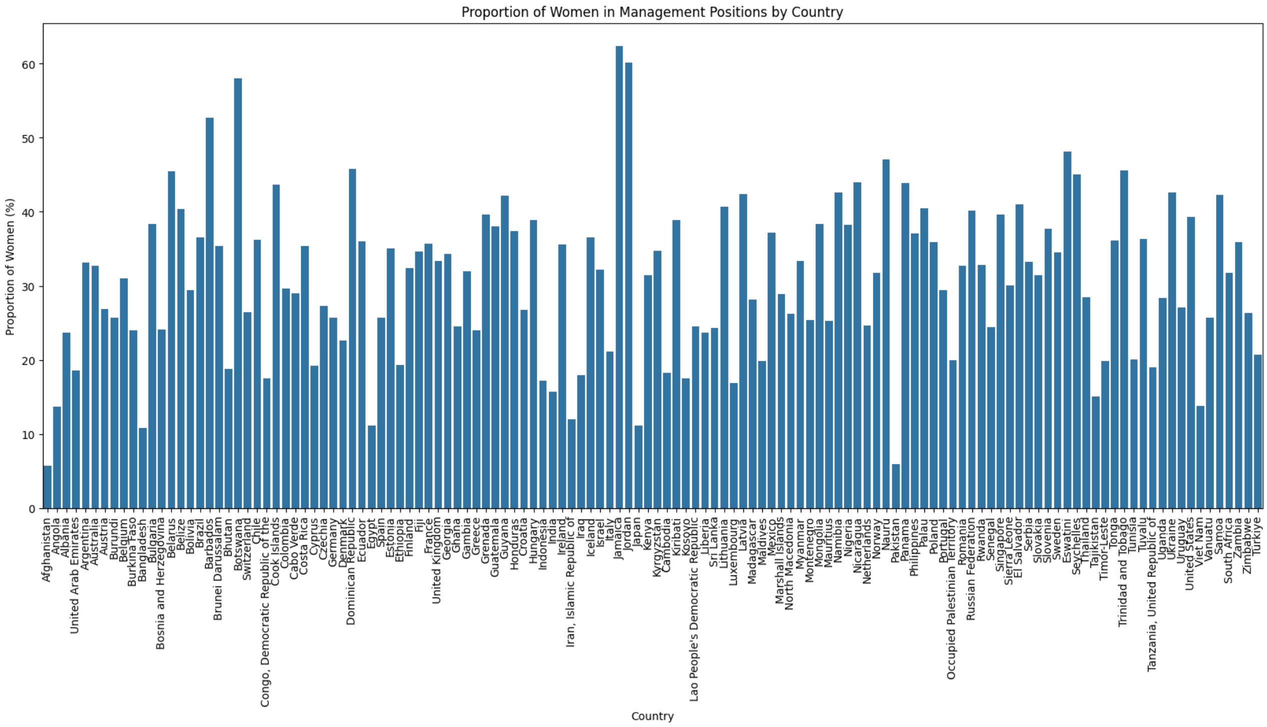

Figure 1: Percentage of women in management positions according to the country (averaged over the years)

After the dataset was finalized, the data was divided into features and output. Two features were used: country and year. Output is the percentage of women in management. The data was divided into two as train and test at a ratio of 0.2. The country feature was encoded because it is not numeric.

CHOOSING THE MODEL

There are many options for the model to be used. Some of them are linear regression, polynomial regression, decision tree regression, random forest regression, XGB regression, lasso regression and ridge regression. In order to choose among these models, it was necessary to determine which one gave more accurate results. For this reason, it was decided to test all of them and make predictions with the model that gives the least error.

The fundamental idea of linear regression is to find the best-fitting straight line (or hyperplane in higher dimensions) through the data points, such that the sum of the squared differences between the observed and predicted values is minimized. When linear regression was applied, the mean squared error was calculated as 26.11.

In polynomial regression, relationship between the independent variable and the dependent variable is modeled as an n-th degree polynomial. When second-degree polynomial regression was applied, the mean squared error is calculated as 26.14. Which is not very different from the mean squared error of the linear regression model. When third-degree polynomial regression is applied the mean squared error is 26.18. This error is still very close to error of linear regression and second order polynomial regression models.

The decision tree algorithm predicts the target variable's value by learning simple decision rules inferred from the training data. When the decision tree model is applied mean squared error is approximately 34.07.

Random forest regression is an ensemble learning method that extends the idea of decision tree regression. It selects subsets of features and data points to train each individual tree randomly. When this algorithm is applied mean squared error is approximately 25.59 for this data set.

Lasso regression (Least Absolute Shrinkage and Selection Operator Regression) model, is a linear regression technique that incorporates regularization to prevent overfitting and promote feature selection. When this regression technique is applied the mean squared error is 64.91.

Ridge regression (Tikhonov regularization or L2 regularization), is a linear regression technique that introduces a regularization term to the standard linear regression objective function. When this technique is applied, a mean squared error of 26.11 is obtained.

XGBoost or eXtreme Gradient Boosting is an ensemble learning method that combines the predictions of multiple weak learners (usually decision trees) to create a strong predictive model. When this model is used mean squared error is calculated as 23.50 which has the smallest error among the tried models.

The error rates of the applied models can be listed from least to most as follows: XGB regression, random forest regression, ridge regression, linear regression, second-order polynomial regression, third-order polynomial regression, decision tree and lastly lasso regression. It seems that ensemble learning methods give more accurate predictions for this data set. Therefore, it was decided to use the eXtreme Gradient Boosting model.

TRAINING THE MODEL AND PREDICTIONS

The XGBoost model is ready to make predictions after it is trained with train data and tested with test data. It can predict the rate of female managers in the coming years for each country in the data set for the desired year. For some countries, the proportion in 2021, 2022, and the proportion predicted for 2023 is tabulated. See Table 1.

	2021	2022	2023 (prediction)
United Arab Emirates	23.57	23.46	24.61
Argentina	37.36	37.91	34.29
United States	43.20	42.83	40.93
Mexico	38.72	39.07	36.98
Austria	34.31	31.53	31.47
Bosnia and Herzegovina	20.18	24.09	24.09
Belarus	43.99	42.36.	41.86
Iran	19.89	19.18	20.17
Brazil	37.16	38.29	36.87
Greece	29.44	33.74	33.05
India	16.96	16.21	18.01
Dominican Republic	56.11	58.71	52.4
Switzerland	30.28	29.88	31.76
France	36.78	29.29	36.24
Türkiye	19.30	20.70	23.10

Table 1: Predicted proportion of women in management positions in 2023

CONCLUSION

In conclusion, a model was created that predicts the proportion of women in management positions for the coming years. As a personal prediction, it was expected that the rates would increase in the future in each country, but according to the model created, the increase or decrease varies from country to country. Additionally, there appears to be a lack of data on this subject. For example, for Turkey, TUIK does not have data on the proportion of women in management positions before 2020. The years for which data are available vary for each country.

FUTURE WORK

In future studies, more detailed data should be sought, data should be collected from the official institutions of the countries, and if data has been collected for previous years, these should be accessed. Classifications can also be made according to the development levels and education levels of countries. As a different study, if characteristics such as gender, sector of employment, education level and age of people currently in managerial positions are obtained as data, a model can be created to predict whether a person is likely to become a manager or not, based on their characteristics.

REFERENCES

[1] T. Hanna, C. Meisel, J. Meyer, G. Azcona, A. Bhatt, S. D. Valero, A. Meagher "Forecasting Women in Leadership Positions", UN Women, University of Denver.

[2] Z. Shahzahi (2023, January 13). Women Managers. Kaggle. https://www.kaggle.com/datasets/zahrashahzahi/women-managers/data