Final Report: Women In Labor

Dalia Habiby, Frankie Tyndall, Karene C Matoka Nana, Daniel Bernal

2022-12-08

Contents

Executive Summary

It was reported by the World Bank that approximately 2.4 billion women do not possess the same economic opportunity as their male counterparts. As part of this number, we count 178 countries preventing women from fully participating in the labor force, 86 countries restricting certain jobs to women, and 95 countries where equal pay is not guaranteed ("Nearly 2.4 Billion", 2022). Those data were made available by the World Bank through their Women, Business, and Law (WBL) project. The goal behind this initiative was to inform and provide "data on the laws and regulations that affect women's economic opportunity"("About Us"). Since 2009, they have been collecting data and researching how to improve women's economic opportunities and empowerment. Our study was conducted using their dataset published in 2021.

We were first interested in knowing if we could predict a country's WBL (Women, Business and the Law) Index based on the length of paid maternity leave, the retirement age, the country's gross domestic product (GDP), the percentage of females in the population, and the total population. The WBL index is made of 35 questions which are divided into 8 categories. Each category is rated "based on the percentage of questions with no restrictions on women's rights" (Indicator: Gender in the Economy). The index is calculated by averaging the scores from those 8 categories. We additionally wanted to classify whether a woman can work in an industrial job in the same way a man can, based on a country's income level, the length of paid paternity leave, the length of paid maternity leave, the WBL Index, the retirement age, and the percentage of females in the population. We concluded that the ridge regression was able to best predict a country's WBL while logistic regression had the highest classification rate for our second research question. We decided to include all predictors variables in our ridge and logistic regression models.

Data and Methods

This research focuses on the analysis of women's global labor inequality measured by the Women, Business and the Law Index and it takes into account a variety of factors such as country's gross domestic product, whether women can obtain an industrial job, paid maternity, paid paternity, retirement conditions, country's population, and percentage of women's population. The data set used covers the most recent year reported (2021) and includes a total of 190 countries which represent the total number of observations. Country selection was based on data availability as many countries don't have data available for the variables used in this analysis. As the accuracy of the results is subject to the data quality, it was considered to obtain all the data from the World Bank's data repository as this is a reliable source that can provide information for all the factors included in this analysis.

In order to obtain only the relevant variables, two steps were executed. First, we considered the Women, Business and Law report for the year 2021. This data set contains the WBL index, the income group for each country, whether a woman can work in an industrial job as a man, length of paid maternity leave, length of paid paternity leave, whether a woman has the same mandatory retirement age as a man, retirement age for women, and retirement age for men. The second step included the addition of some socioeconomic variables obtained from the World Bank's World Development Indicators that include GDP, unemployment rate, total population, and percent of women population for each country.

Once these data sources were merged, the final data set contains the following information:

```
## Rows: 190
## Columns: 14
## $ ID              <chr> "AFG2021", "AGO2021", "ALB2021", "ARE2021", "ARG2021", ~
## $ Country         <chr> "Afghanistan", "Angola", "Albania", "United Arab Emirat~
## $ CountryCode     <chr> "AFG", "AGO", "ALB", "ARE", "ARG", "ARM", "ATG", "AUS",~
## $ Region          <chr> "South Asia", "Sub-Saharan Africa", "Europe & Central A~
## $ IncomeGroup     <chr> "Low income", "Lower middle income", "Upper middle inco~
## $ WBLIndex        <dbl> 38.125, 73.125, 91.250, 82.500, 76.250, 82.500, 66.250,~
## $ IndustrialJob   <chr> "No", "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "Y~
## $ PaidMaternity   <dbl> 90, 90, 365, 45, 90, 140, 91, 0, 112, 126, 84, 105, 98,~
## $ PaidPaternity   <dbl> 14, 1, 3, 0, 2, 0, 0, 14, 30, 0, 4, 14, 3, 3, 0, 21, 1,~
## $ SameRetirement <chr> "Yes", "No", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", ~
## $ GDP             <dbl> NA, 7.254699e+10, 1.826004e+10, NA, 4.914930e+11, 1.386~
## $ Unemployment    <dbl> 5.58, NA, NA, NA, 8.74, NA, NA, 5.11, 6.18, NA, NA, 6.2~
## $ PercPopFemale   <dbl> 48.70149, 50.52062, 49.11087, 31.05507, 51.20231, 52.96~
## $ TotalPop        <dbl> 39835428, 33933611, 2811666, 9991083, 45808747, 2968128~
```

While assessing the steps for our data collection, we identified ethical challenges that can arise from our own biases or the biases of others affecting our data, thoughts, and actions. We considered availability heuristic bias given the data sources that we encountered, our sources must have the information and richness required for our analysis. Furthermore, we are aware of the bandwagon effect, accountability is what matters most and all opinions are important at the moment of selecting our data, that is why a consensus was reached to use the World Bank data as our main source. Lastly, we took into consideration the rules and policies related to the American University's Student Code of Conduct to ensure transparency and student compliance with our deliverables.

Results

The first research question we investigated concerned predicting a country's Women, Business and the Law (WBL) Index through regression analysis. In order to find the best model, we applied ordinary least squares linear regression, ridge regression, and lasso regression. Table 1 above reports our findings from running a linear model on the five identified predictors as well as utilizing Leave One Out Cross Validation to find the prediction mean squared error of the model. Furthermore, in an effort to make sure our models are as robust as possible, we investigated multicollinearity through variance inflation factor as well as the correlation between our predictors.

Table 1: OLS Linear Regression on WBL Index

Predictor	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-28.3	18.5	-1.53	0.129
PaidMaternity	0.0726	0.0215	3.38	0.000899
SameRetirementYes	7.48	4.33	1.73	0.086
GDP	1.86e-12	6.1e-13	3.05	0.00269
PercPopFemale	1.82	0.362	5.03	1.3e-06
TotalPop	-1.82e-08	8.98e-09	-2.02	0.0449
Model Prediction MSE	232.9637			

All of these predictors are statistically significant at the 0.05 level, except the same retirement age variable, which is statistically significant at the 0.1 level. The variable standard errors are relatively small for the non-indicator variables, though the adjusted r-squared value was only 0.22, meaning that this model accounts for only 22% of the variation in WBL index. Furthermore, the calculated mean squared error of prediction was 232.9637, which is acceptable but not optimal.

Table 2: Correlation Between Predictors

	WBLIndex	PaidMaternity	GDP	PercPopFemale	TotalPop
WBLIndex	1.0000000	0.2448817	0.1231775	0.3977986	-0.0353495
PaidMaternity	0.2448817	1.0000000	-0.0920806	0.1833686	0.0733460
GDP	0.1231775	-0.0920806	1.0000000	0.0081424	0.6003541
PercPopFemale	0.3977986	0.1833686	0.0081424	1.0000000	-0.0339052
TotalPop	-0.0353495	0.0733460	0.6003541	-0.0339052	1.0000000

The only set of predictors with a worrying correlation is GDP and Total Population, which is to be expected. However, we decided that a value of 0.6 is not cause for taking out one of the variables before pursuing shrinkage methods.

Since the linear regression model included all five predictors in our full model, we decided to explore if they had multicollinearity before moving forward. Table 2 above indicates each variable's correlation with the other predictors. Table 3 below displays the variance inflation factors for our linear model. Since there was a moderately strong correlation between GDP and Total Population, we chose to move on to shrinkage methods, utilizing Ridge and Lasso regression to minimize the error sum of squares and reduce variance.

Table 3: Predictor Variance Inflation Factors

	PaidMaternity	SameRetirement	GDP	PercPopFemale	TotalPop
vif.reg.	1.131919	1.098248	1.624531	1.042813	1.635735

The variance inflation factors are not concerning, as they are all between 1 and 2.

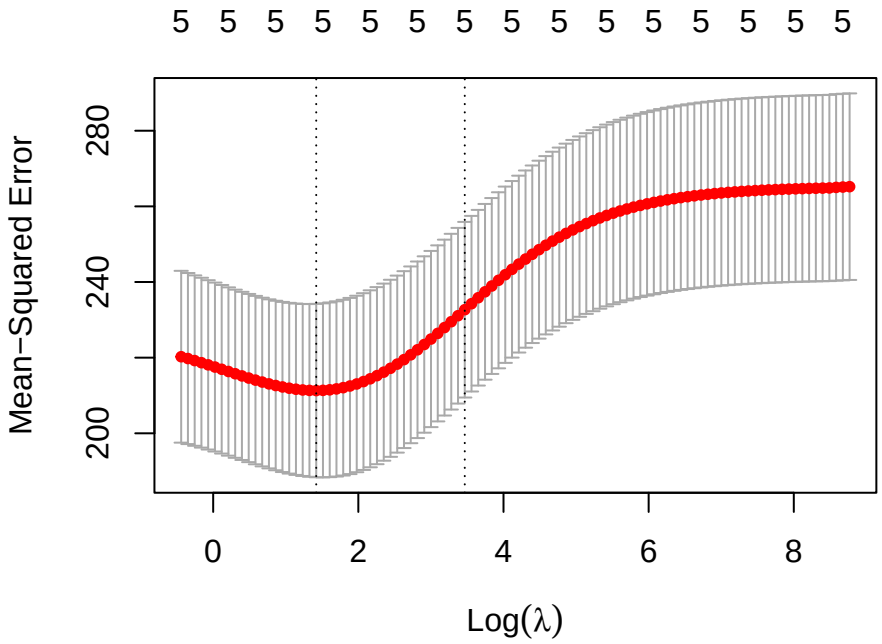

Figure 1: This graph represents the mean-squared error for the log of each value of lambda using Ridge Regression. Through cross-validation, we found that the best lambda value for minimizing the mean-squared error was 4.14. This lambda brought the mean-squared error down to 211.3. This is an improvement from the linear model results, and all five predictors are still in the model, as ridge regression cannot shrink parameters to zero.

Table 4: Ride Regression Coefficients

Estimate	Predictor
-8.344265e+00	(Intercept)
5.490192e-02	PaidMaternity
5.212539e+00	SameRetirementYes
1.155952e-12	GDP
1.501931e+00	PercPopFemale
-9.942644e-09	TotalPop

The Ridge coefficient for each predictor is notably smaller than the OLS estimate.

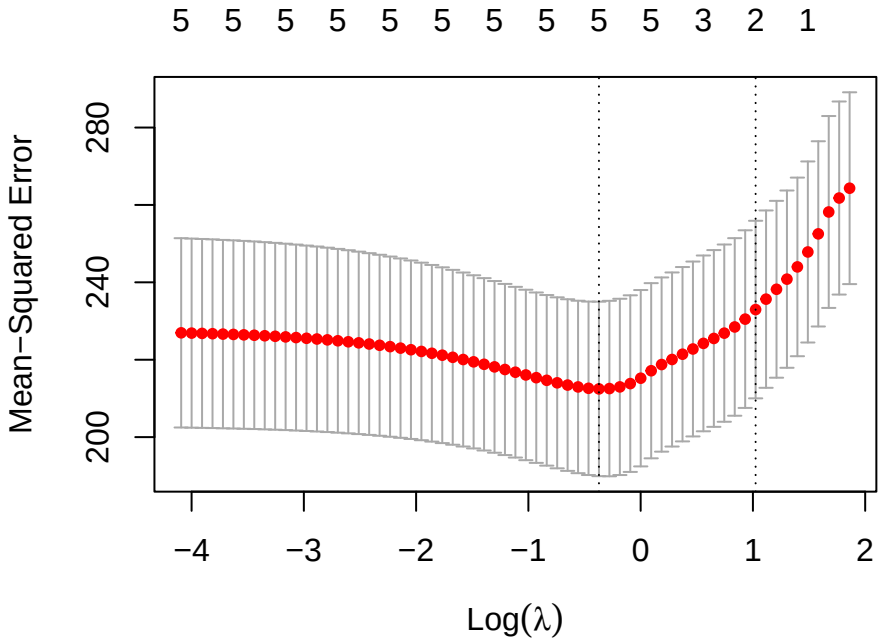

Figure 2: This graph represents the mean-squared error for the log of each value of lambda using Lasso Regression. Through cross validation, we concluded that the best lambda value was 0.6898. This lambda had a mean-squared error of 212.5. This is an improvement from the linear model results, though not quite as low as the ridge regression measure was. When analyzing the minimum lambda value, all five predictors are still in the model. However, the 1se lambda value of 2.7847 only includes 2 parameters: length of paid maternity leave and percent of the population that is female. Though, the mean-squared error of the model with lambda as 2.7847 would be notably higher, at 233.0.

Table 5: Lasso Regression Coefficients

Estimate	Predictor
-1.572546e+01	(Intercept)
5.214277e-02	PaidMaternity
4.126298e+00	SameRetirementYes
1.006970e-12	GDP
1.674420e+00	PercPopFemale
-6.969909e-09	TotalPop

The Lasso coefficient for each predictor is notably smller than the OLS estimate.

The second question of interest regarded classifying whether or not a woman can work in an industrial job in the same way a man can. When building the K Nearest Neighbors model, we accounted for the following variables: amount of days for paid paternity leave, WBL index, percent of women population, amount of days for paid maternity leave, whether the retirement age is the same for men and women, and country income level. To execute the model, two things were necessary; First, we performed a one-hot encoding for same retirement age, as this predictor provides only "Yes" and "No" values. Second, we performed a label encoding for income group and assigned values from 1 through 4 for each of the levels: "Low income", "Lower middle income", "Upper middle income", "High income". The data was divided in half for training and testing sets, and since the initial nearest neighbors were randomly selected, it was necessary to tune the parameters of the model to evaluate if the accuracy could increase. Here are the results:

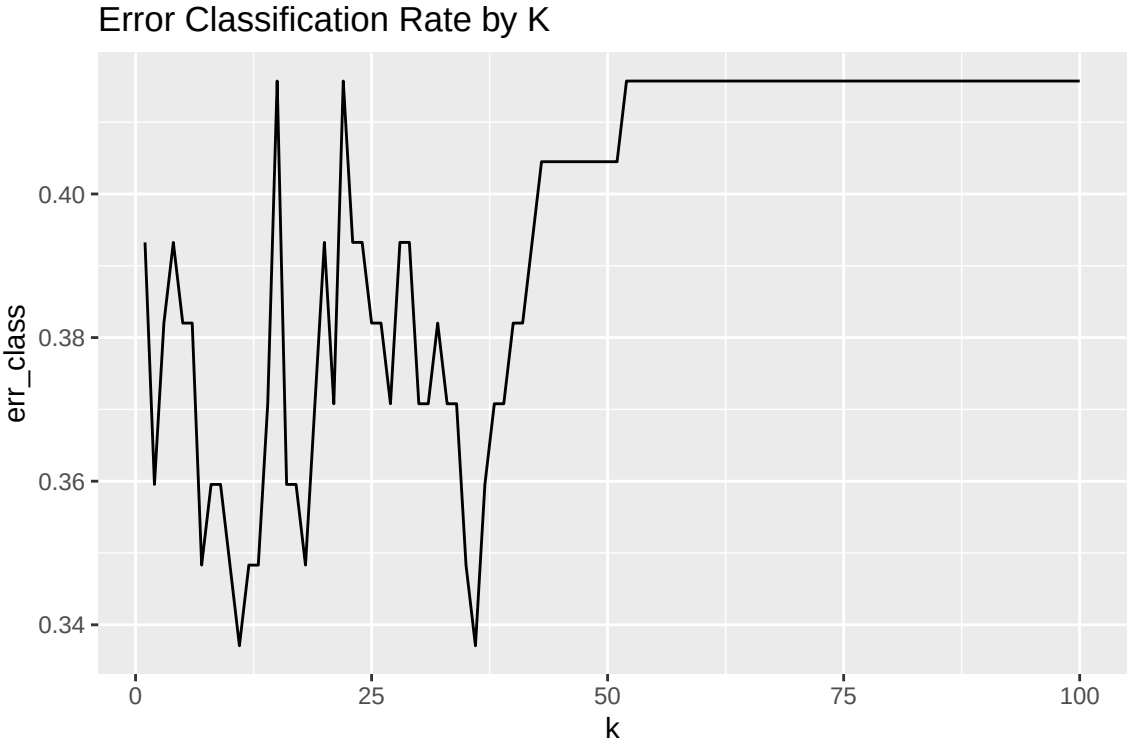

Figure 3: This plot displays that through cross validation, the lowest error classification rate occurs when k nearest neighbors is 11. The value of this error classification rate is 0.3370787, which means that the correct classification rate is 0.6629213.

Our second classification approach was the logistic regression model. It included the same variables as the KNN model: amount of days for paid paternity leave, WBL index, percent of women population, amount of days for paid maternity leave, same retirement age, and country income level. Similarly, data was divided in half (50%) for train and test sets. We opted to use the original data values and properly classify them based on their data classes. Here are the results for the logistic model on the training data set:

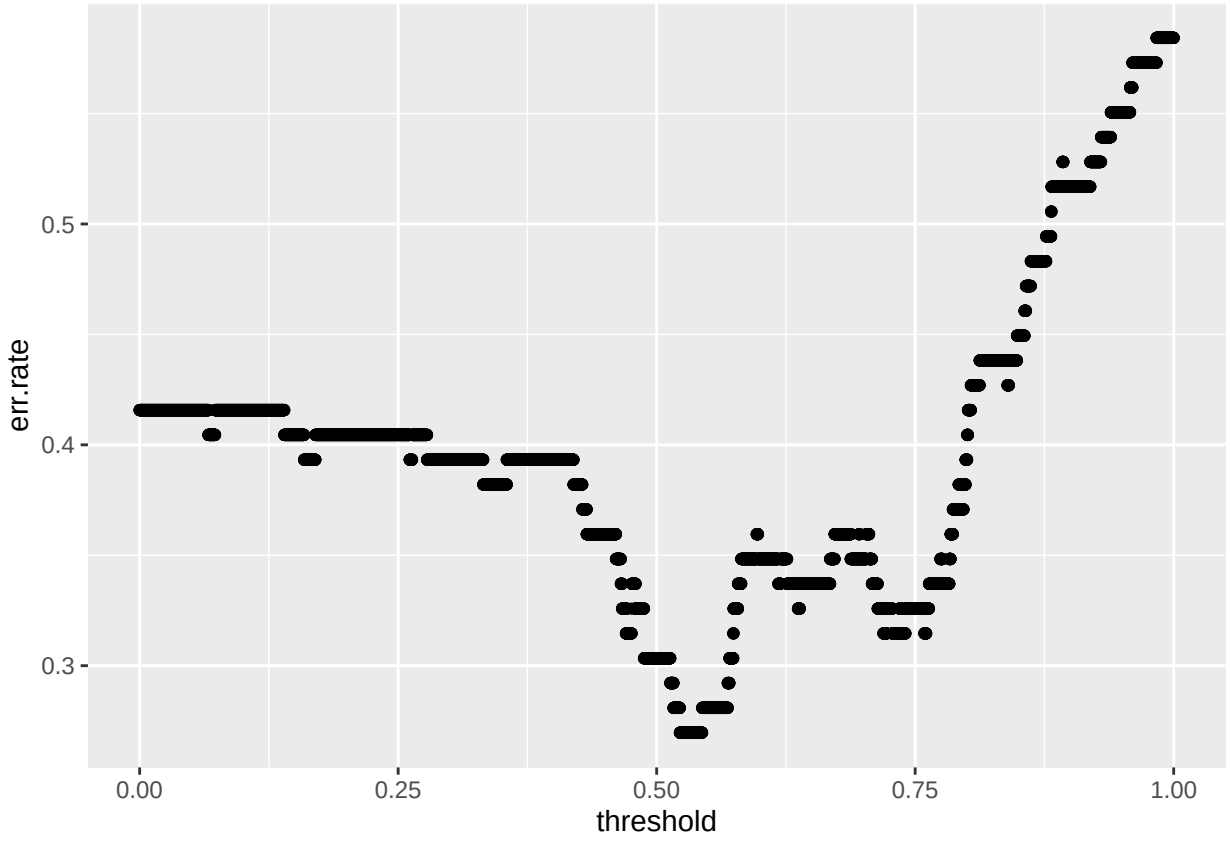

Figure 4: According to cross-validating our full logistic regression model, the lowest error rate occurred when the threshhold was 0.5225, with an error rate of 0.2696629. This means that the correct classification rate is 0.73033

Table 6: Confusion Matrix for LDA

	No	Yes
No	17	20
Yes	6	46

Table 7: Confusion Matrix for QDA

	No	Yes
No	24	13
Yes	15	37

Our third and fourth aproaches were Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). The initial LDA model included all the six variables used in the KNN and the logistic regression models. Since the variables (SameRetirement and IncomeGroup) are both categorical, we decided to convert those variables to a factor. The first step consisted of running LDA on our training dataset, then we used the testing data for prediction, and we finished by providing the confusion matrix above. However, while running the full model, we also found out that the removal of the variable IncomeGroup resulted in a higher classification rate compared to the full model. The model without IncomeGroup resulted in a classification rate of 0.708, which is more accurate than the KNN model but not as accurate as the logistic regression.

The QDA model also included all six of the variables in this section. We followed the same steps as for the Linear Discriminant Analysis, used the same training data set to run QDA, used the same testing dataset for prediction and provided a confusion matrix above to show the performance of this model. Although the LDA performed better without the IncomeGroup variable, removing it had no effect on the QDA results. The full QDA model had a classification rate of 0.685. Thus, the linear discriminant analysis was more appropriate for this question, but the logistic regression was the overall best.

Discussion

In many nations, there have been significant legal rights advances for women in recent years. Other places continue to firmly establish limits on women's autonomy. We used the linear, ridge, and lasso approaches to test the best model and see if there are any links between a country's WBL Index and labor measures including paid maternity leave length, retirement age, country's gross domestic product (GDP), the percentage of females in the population, and the total population. According to the article "Gender-Discriminatory Laws and Women's Economic Agency," the WBL Index appears to be positively correlated with paid maternity leave, but the researchers discovered that there is a strong correlation between the WBL Index and equal labor market outcomes rather than with retirement age (Htun et al., 2019).The paper also discovered that limitations on women's legal competence predict their possession of assets and involvement in the labor force, while discrimination at work, parental leave, and the magnitude and direction of salary gaps had positive associations with the WBL Index (Htun et al., 2019). The significance of defining and quantifying legal rights and their potential impacts as multidimensional effects is highlighted by these findings.

For the first part of our study, we used different regression methods to discover from our data

if there is a relationship between the WBI Index as a response variable and paid maternity leave as well as age of retirement as the response variables and which of these methods was best. The first method used to analyze the relationship was linear regression. Using linear regression, we discovered that of the predictor variables only paid maternity leave length, GDP, and percent of population that is female had any significant relationship with the WBL Index. These results correlate with the findings from the 2019 research paper mentioned above. Being that our data was collected in 2021, it is not surprising that over the span of only 2 years that there is still no relationship with the WBL Index and retirement age. What we did find that the above paper did not consider is the relationship to the response variable and the percent of women in each country. That predictor produced a strong p-value, similarly to how the paid maternity leave length predictor did, which leads to them both being very significant in predicting the WBL Index response variable.

Our current model shows us that only about 22% of the variation in the WBL Index can be explained by our model with the 5 predictors. Using our linear regression model, we found the adjusted R-squared, BIC, and CP, both the adjusted R-squared and the CP found that our best model included all of the predictors while the model with the lowest BIC would only have two predictors, which are length of paid maternity leave and percent of the female population. These two variables being significant in correlation with the WBL Index makes sense as the WBL index examines how laws impact women at various stages of their careers, concentrating on those laws that are relevant in the major business city. Thus, the length of paid maternity leave as well as percent of women in the population would work as the best model.

In order to determine the best predictors for the model when studying the WBL Index we also looked at the multicollinearity and variance reduction methods. Ridge regression and lasso regression were employed on the model to determine the best parameters through process of elimination. When we found the variance inflation factors and they were all around 1 indicating that there is minimum variance among the predictors. The predictors GDP and total population had higher VIF values, indicating that having the two predictors in the model will inflate the variances of all the other variables by around 2. The ridge regression that was conducted to help reduce the chances of the variance inflation being due to multicollinearity. Using ridge regression, all of the predictors are kept in our model, as shown in the above in Figure 1. For the lasso, only the paid maternity leave and the percent of female population were kept as supported by the BIC we calculated in from our linear model.

For the linear model the calculated predicted mean squared error, which is the minimized sum of the squared errors, is 232.9637, for the ridge regression the predicted mean squared error is 211.3, and for lasso 212.5. The Ridge regression has the lowest mean squared error with 5 predictors. From the WBL Index, we also wanted to classify whether a woman can work in an industrial job in the same way a man can based on a country's income level, length of paid paternity leave, length of paid maternity leave. While we are looking at classifying based on income level, length of paid paternity leave, length of paid maternity leave, the WBL Index, retirement age, and the percentage of females in the population, the article "Women in Male-Dominated Industries and Occupations (Quick Takes)" (n.d., 2021) discusses the already existing barriers that women face working in industrial jobs, such as societal expectations and beliefs about women's leadership abilities, pervasive stereotypes, such as that of the "caring mother" or office housekeeper (n.d., 2021). Also, according to a Pew Research Center study, 28% of women working in male-dominated industries have personally experienced sexual harassment, compared to 20% of women working in female-dominated industries (Parker, 2018.).

To be able to classify whether a woman can work in an industrial job in the same way as a man based on six variables, we employed KNN classification, logistic regression, linear discriminant analysis, and quaratic discriminant analysis. Being that the response variable is a qualitative variable,

with values of either no or yes as levels, we needed to find the values with the k-nearest neighbors to produce the best estimate of the response. As shown in the Figure 3 above, the calculated K is 11 which is somewhat large given how relatively small our dataset is. This indicates that or model might be too restrictive, meaning that the estimate of Y-hat (whether a woman can work in an industrial job in the same way as a man) may be based on data or points that are irrelevant. This is why we conducted a logistic regression, to be able to observe how the response variable responds to each of the predictor variables in our classification model. From the logistic regression model, we find that only the length of paternity leave and the WBL Index are significant; all the other predictors in our model do not have significance. This conclusion does correlate with our finding with the KNN classification method that we might have data in our model that is irrelevant to estimating the response variable.

Finally, the LDA and QDA methods included the same six variables as KNN and logistic regression: amount of days for paid paternity leave, WBL index, percent of women population, amount of days for paid maternity leave, same retirement age, and country income level. However, the income variable was ultimately removed from the linear discriminant analysis. The LDA model performed better than the QDA model, with a classification rate of 0.703 while the QDA had a classification rate of only 0.685. Despite this, the best model for predicting whether or not a woman could work in an industrial job in the same was as a man was the logistic regression.

Contrubutions and Recommendations

For additional research on our first question, "Can we predict a country's WBL (Women, Business and the Law) Index based on the length of paid maternity leave, retirement age, the country's gross domestic product (GDP), the percentage of females in the population, and the total population?", we would recommend creating a model with predicting the WBL Index using paid maternity leave, percentage of female population, and the equal labor market outcomes variable, that was found to be significant in the 2019 research article, and observe if a better model, with a stronger predictive power can be produced.

For the second question, "Can we classify whether a woman can work in an industrial job in the same way a man can based on a country's income level, length of paid paternity leave, length of paid maternity leave, the WBL Index, retirement age, and the percentage of females in the population?", we recommend that additional research be conducted looking at not collective country income level and paternity leave, but the income levels of women specifically in industrial jobs and other variables such as rate of sexual harassment and access to mentorship be added to the data to better classify whether or not a woman can work in an industrial job in the same way as a man.

Our research contributes to the existing literature by bringing focus to the percentage of the population that are women and by providing analysis that can be utilized by many stakeholders. This study is targeted towards governments, officials, organizations, and activists who are interested in the research of women labor inequality. We believe that being able to predict a country's WBL Index as well as a woman's ability to work in a similar manner to a man is a rich source of information. Our findings can potentially provide a baseline to develop further studies that can lead to policy recommendations to fight gender inequality across the world.

References

About Us. (n.d.). [Text/HTML]. World Bank. Retrieved December 8, 2022, from https://wbl.worldbank.org/en/aboutus

Indicator: Gender in the Economy. (n.d.). Millennium Challenge Corporation. Retrieved December 8, 2022, from https://www.mcc.gov/who-we-select/indicator/gender-in-the-economy-indicator

Nearly 2.4 Billion Women Globally Don't Have Same Economic Rights as Men. (2022). Retrieved December 8, 2022, from https://www.worldbank.org/en/news/press-release/2022/03/01/nearly-2-4-billion-women-globally-don-t-have-same-economic-rights-as-men

Women in Male-Dominated Industries and Occupations (Quick Take). (n.d.). Catalyst. Retrieved December 8, 2022, from https://www.catalyst.org/research/women-in-male-dominated-industries-and-occupations/

Htun, M., Jensenius, F., & Nelson-Núñez, J. (2019). Gender-Discriminatory Laws and Women's Economic Agency. Social Politics, 26. https://doi.org/10.1093/sp/jxy042

Parker, K. (2018). Women in majority-male workplaces report higher rates of gender discrimination. Pew Research Center. Retrieved December 8, 2022, from https://www.pewresearch.org/fact-tank/2018/03/07/women-in-majority-male-workplaces-report-higher-rates-of-gender-discrimination/

Appendix

Daniel Bernal: I collaborated with the data collection and preprocessing such as adjusting the format, filtering, and correction of invalid entries. I developed the data, methods, and ethics discussion, and coordinated the development and maintenance of the GitHub repository. Contributed with the KNN and Logistic Regression models and their respective outputs and overview. I also obtained additional variables to add onto the original dataset.

Dalia Habiby: I collected the original dataset as well as the total population and percent female population variables. I also assisted in data cleaning and combining. Primarily, I worked with Frankie on the linear, lasso, and ridge regression analyses. I wrote the results section of the report and formatted the body of the report and all graphs and tables in Rmarkdown.

Karene Matoka: I worked with Daniel on classification analysis (second research question). We both worked on KNN and logistic regression and created graphs to showcase our results. I worked on the LDA analysis on my own. I tried running LDA after removing one variable to see if the classification rate would increase. I also wrote the executive summary as well as reported our findings under the results section.

Frankie Tyndall: I wrote up the discussion and recommendations, as well as cited references from articles pertaining to the results that were found when studying the WBL Index dataset. I also helped with data cleaning procedures on the dataset such as removing unwanted variables, duplicated information, and NA values. Finally, I specifically worked on the linear, lasso, and ridge regression methods to get the necessary information needed for our results and discussion/recommendation sections.