# Predicting the Women Entrepreneurship Index

## Problem

We are looking to analyze how different factors affect the Women Entrepreneurship Index. Based on that, we are trying to predict the Women Entrepreneurship Index around the World.

Our goal is to determine what factors influence the Women Entrepreneurship Index and how it can potentially be increased overall and in countries with lower Women Entrepreneurship Index. We can evaluate these factors through a correlation matrix. We will look at feature correlation to analyze the most contributing factors to the Women Entrepreneurship Index.

We chose to predict the Women Entrepreneurship Index and not the Female Labor Force Participation Rate because it is a unique skill to be an entrepreneur. Almost anyone can be a labor force participant but not an entrepreneur. Besides, an entrepreneurial lifestyle brings a certain level of independence, responsibility, and freedom. This is important for women because women's rights were suppressed for a long time. Women were highly dependent on men as a source of income, which was limiting women's freedom and ability to live a desired life. However, the World has changed, and equal opportunity is presented to anyone brave enough to take a chance and build a successful company or capitalize on one's passion through building a business.

Needless to say, the development of the Entrepreneurship Index as a whole is a complex subject depending on a particular country's laws, regulations, values, culture, financial system, and the ease of setting up a business legal entity. We can try to find additional data to include in our initial dataset to make the model more accurate and representative of reality to figure out what factors drive an increase in the Women Entrepreneurship Index. For example, [gender equality index](https://eige.europa.eu/gender-equality-index/2020), [ease of doing business](https://www.doingbusiness.org/en/rankings), business financing availability, or [corporate tax rates](https://taxfoundation.org/corporate-tax-rates-europe-2019/#:~:text=All%20European%20countries%20tax%20corporate,global%20average%20(21.4%20percent).). However, it will solely depend on public data availability.

We will measure success by ensuring all the necessary data preprocessing steps were taken, with feature normalization and dropping unnecessary features that do not bring value to the results. Ensure that our model is as accurate as possible in our training and testing phases with calculating performance metrics for regression. Also, we can train and compare metrics for five different models to choose the most accurate one for our prediction.

## [Dataset](https://www.kaggle.com/babyoda/women-entrepreneurship-and-labor-force) (link attached)

We will be using the *Women Entrepreneurship and Labor Force* dataset for our final project*.* This data was obtained from the Women Entrepreneurship Index and Global Entrepreneurship Index report published in 2015. The dataset itself was downloaded from Kaggle. The research is limited to OECD countries where all data for 2015 are available simultaneously in the database.

This dataset has eight features:

No (or Country ID, which we will remove)

Country - Country Name

Level of development - Nominal variable for Developing Index

European Union Membership - Nominal variable for EU Membership

Currency - Currency

Entrepreneurship Index - Entrepreneurship Index 2015

Inflation rate - Inflation Rate 2015

Female Labor Force Participation Rate - Female Labor Force Participation Rate 2015

The label for this dataset is:

Women Entrepreneurship Index - Women Entrepreneurship Index 2015.

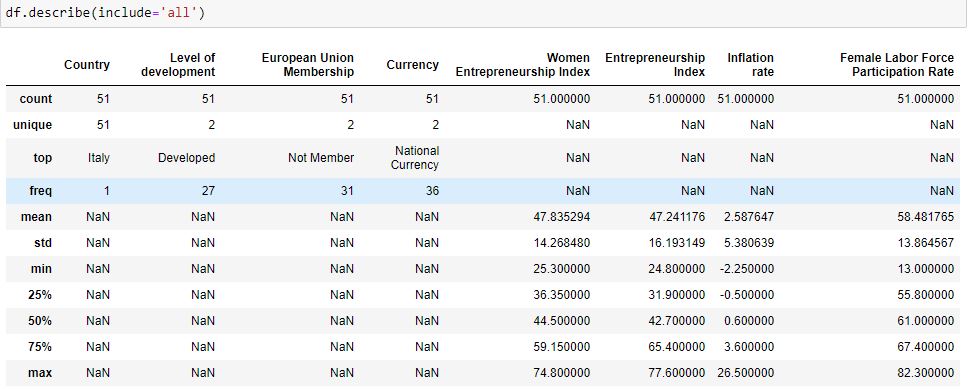


Figure 1. Dataset Statistics

We do not have any missing values in our original dataset. There are 51 distinct countries. The mean Entrepreneurship Index is 47.2 (min 24.8 and max 77.6), while the mean Women Entrepreneurship Index is 47.8 (min 25.3 and max 74.8). The mean Inflation rate is 2.6 with a standard deviation of 5.4. Average Female Labor Force Participation 58.5 (min 13 and max 82.3 and std 13.9).

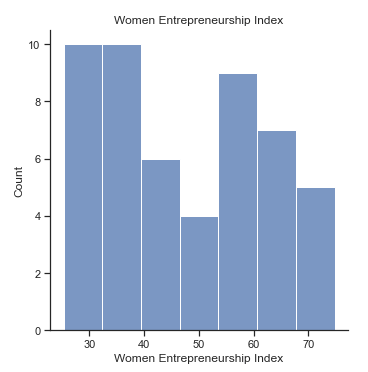
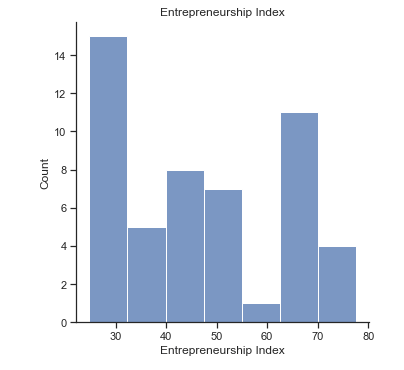


Figure 2. Entrepreneurship Index Histogram Figure 3. Women Entrepreneurship Index Histogram

Based on these graphs alone, the Women Entrepreneurship Index is lower in the count but more densely distributed than the overall Entrepreneurship Index.

### Features added to the dataset:

[Gender Inequality Index](https://en.wikipedia.org/wiki/Gender_Inequality_Index#Rankings)

[Ease of Doing Business](https://www.doingbusiness.org/en/rankings)

[Corporate Tax Rates](https://taxfoundation.org/publications/corporate-tax-rates-around-the-world/) 2020

[Schooling](https://www.kaggle.com/kumarajarshi/life-expectancy-who) (Years)

## Proposed Solution

#### Explain how you plan to solve this problem.

We will try three to five different models to predict the Women Entrepreneurship Index and choose the one with the best performance metrics. Since Women Entrepreneurship Index is a numerical value, it is a regression problem. We will use Regression Models such as SGDRegressor, Linear Regression, Lasso, Decision Tree Regressor, and Voting Regressor.

#### Splitting the Data

We will split data 80/20, and we split the test and validation set 50/50. Since it is not a time series dataset, we do not worry about what data goes into the test, train, or validation set. The split is randomly initiated. We do not touch the X\_test set until the very end to avoid data leaks.

After splitting the data and looking at the dataset statistics, we noticed that feature 'Gender Inequality Index' highly correlates with a few other features such as 'Entrepreneurship Index,' 'Schooling (years),' and a target variable 'Women Entrepreneurship Index.' Therefore, we had to drop it to account for multicollinearity.

#### Performance Metrics

We will use MSE or Mean squared error, R2 score, MAE or Mean absolute error, and MaxError. These are the best metrics for a regression problem:

* MSE is one of the most preferred metrics for regression tasks. It is simply the average of the squared difference between the target value and the value predicted by the regression model. It is preferred more than other metrics because it is differentiable and hence can be optimized better.
* MAE is a more suitable metric for measuring outliers. However, we do not have extreme outliers in our dataset.
* R2 metric helps us compare our current model with a constant baseline and tells us how much our model is better.

## Modeling

Describe the models you have decided to use to solve your problems.

We have decided to use several of the most common linear models from the sci-kit-learn library. They are:

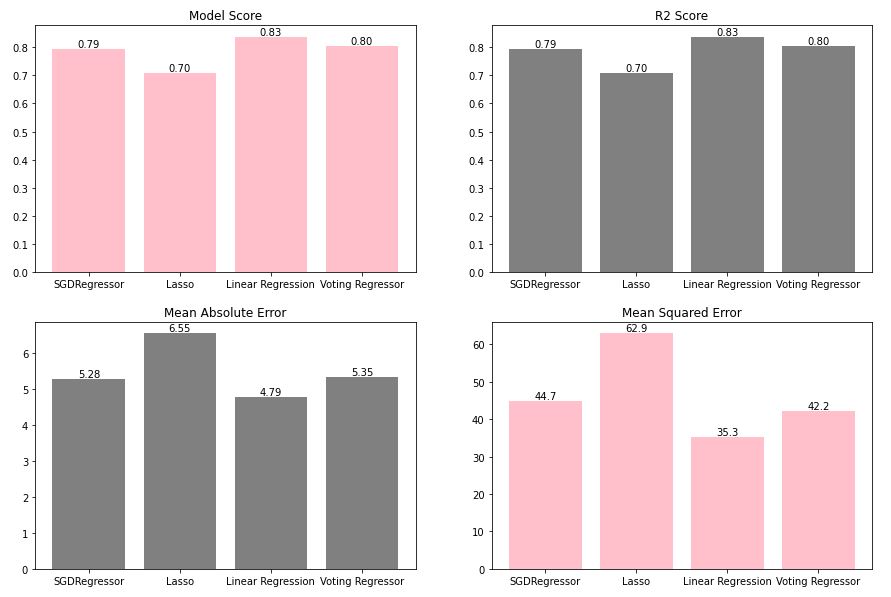
* Linear Regression: Here, we are examining two things: First, does a set of predictor variables do a good job predicting an outcome (dependent) variable? Second, which variables, in particular, are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.
* Decision Tree Regressor: Here, we are using one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application. It is a tree-structured classifier with three types of nodes.
* Linear Lasso: The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters).
* Stochastic Gradient Descent (SGD) Regressor: Stochastic Gradient Descent (SGD) is a simple yet efficient optimization algorithm used to find the values of parameters/ coefficients of functions that minimize a cost function. In other words, it is used for discriminative learning of linear classifiers under convex loss functions such as SVM and logistic regression. It has been successfully applied to large-scale datasets because the update to the coefficients is performed for each training instance rather than at the end of instances. Here we are implementing a simple SGD learning routine supporting various loss functions and penalties to fit the linear regression models.
* Voting Regressor: The idea behind the VotingRegressor is to combine conceptually different machine learning regressors and return the average predicted values. Such a regressor can be useful for a set of equally well-performing models to balance out their weaknesses.

We have selected these models as they are the most common and best performing models for a regression problem predicting numerical values. The scenario(s) that could limit these models' success would be multicollinearity, a higher number of hyperparameters, and if we did not perform our data transformation correctly.

## Results on the Train Set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metrics | **R2** | **MSE** | **MAE** | **Max Error** |
| *Linear Regression* | 0.83615667815756 | 35.3027051209182 | 4.79031580004211 | 18.2520238101492 |
| *Lasso* | 0.82555695725401 | 37.5865871687985 | 6.55803030303030 | 17.6588308109373 |
| *SGDRegressor* | 0.78626116961009 | 46.0535029276333 | 5.28329057157982 | 21.0713300618435 |
| *Decision Tree Regressor* | 0.99996171093969 | 0.00825000000000 | 0 | 0.25 |
| *Voting Regressor* | 0.83350674793166 | 35.8736756328883 | 5.35874273644460 | 17.9554273105433 |

Table 1. Performance Metrics on the Train Set

Figure 4. Performance Metrics Histogram

The trade-offs are as follows:

* Decision trees are constantly overfitting the data. Therefore, they are not as accurate.
* In the Linear Regression case, the model suits the data the best as there are few to no outliers. However, if we were to add more complexity to the dataset, the error rates would increase.
* Lasso, however, struggles with some types of data. Suppose the number of predictors is greater than the number of observations. Lasso will pick at most n predictors as non-zero, even if all predictors are relevant. Lasso will also struggle with collinear features (if they are related or correlated strongly), in which it will select only one predictor to represent the full suite of correlated predictors. This selection will also be made in a random way, which is terrible for reproducibility and interpretation.
* A voting ensemble can offer lower variance in the predictions made over individual models. This lower variance may result in a lower mean performance of the ensemble. However, a limitation of the voting ensemble is that it treats all models the same, meaning all models contribute equally to the prediction. This is a problem if some models are good in some situations and deficient in others.
* While SGDRegressor is very efficient and offers ease of implementation, Stochastic Gradient Descent (SGD) requires several hyperparameters like regularization parameters, and it is sensitive to feature scaling.

## Conclusion

Linear Regression should be used to solve this problem since it has the best performance results.

While Decision Tree Regressor ran faster at 269 µs ± 9.45 µs per loop and Linear Regression ran at 286 µs ± 8.15 µs per loop, it would make sense to pick Linear Regression regardless because the Decision Tree model tends to overfit and would make for a bad predictor.

The Linear Regression model was marginally larger (672 bytes) in size than the Lasso model (661 bytes), but given that it had significantly better performance. It would still make sense to choose that as our final model to train and test our data on.

Based on the trade-offs mentioned in the previous section, we can hypothesize the following best use cases for the other models:

* If there are a high non-linearity and complex relationship between the dependent and independent variables, a Decision Tree model will outperform a classical regression method. Therefore, if we add more new features to the dataset, the Decision Tree model can have a better performance.
* If we were to introduce low levels of multicollinearity in the dataset, the Lasso regression would be better suited for the analysis.
* The SGDRegressor is particularly useful when the number of samples (and the number of features) is very large. SGDClassifier and SGDRegressor provide the functionality to fit linear models for classification and regression using different (convex) loss functions and different penalties.
* Voting Regressor would be our second model choice. It has similar results to the Linear Regression since we included Linear Regression and Lasso into our Voting Regressor model.