ARTICLE

Unemployment Rate Prediction Upon Gender

Xiwen Cao, Xu Chen

Equal Contribution

1. Data Analysis

1.1 Background

The problem of unemployment is a significant challenge for our organization and the wider US economy. High unemployment rates can lead to social and economic instability, as well as increased public spending on welfare programs. Therefore, understanding the factors that influence unemployment rates is crucial for developing effective policies to tackle this issue. Our project's goal is to understand the impact of macroeconomic factors on the unemployment rate for men and women separately in the US. Specifically, we would like to investigate how changes in factors like GDP, inflation, interest rates, wage, population, tax rate and CPI affect the unemployment rate for each gender.

1.2 Data Description

The data utilized in this study was obtained from FRED and spans the period from 1982 to 2023, with a monthly frequency. A total of 494 entries were collected, encompassing 8 explanatory variables that were used to evaluate the unemployment rate. Additionally, data was gathered for the unemployment rate for both men and women on a monthly basis, enabling the assessment of the impact of these macroeconomic factors on unemployment rates for each gender separately.

The 8 explanatory variables collected for the study include quarterly GDP, population, interest rate, wage, industrial production (total index), M2, and smoothed U.S. recession probabilities. All of these variables are continuous in nature. However, since U.S. recession probability is highly skewed, most of its value is very small, so we transform it into a categorical variable.(Which will be discussed in Data Cleaning)

The dataset was randomly split into 80% for training data and 20% for testing data. In the next step, to gain insight into the nature of the data, visualization techniques were employed.

1.3 Data Visualization

1.3.1 Correlation Visualization

Data visualization is a valuable technique to aid in the exploration of variables that may impact the unemployment rate. In this study, a correlation plot was utilized to identify which numerical variables may exert a distinctive influence on the unemployment rate. By analyzing the correlations between each pair of numerical variables, a more comprehensive understanding of their relationships can be obtained. This approach provides a visual representation of the strength and direction of the associations between variables, enabling the identification of potential causal relationships and aiding in the interpretation of the results.

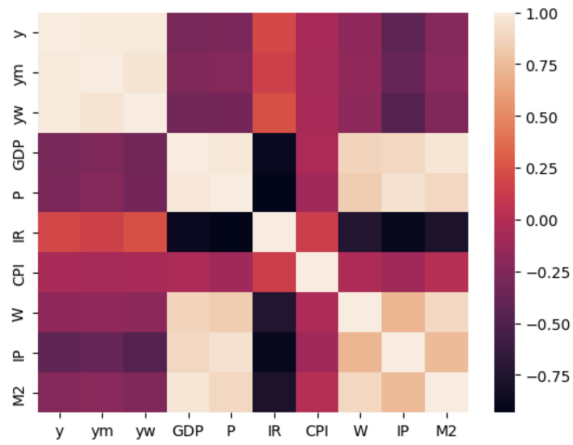

Figure 1. Correlation Visualization

As we can see from Figure 1, the unemployment rate is highly correlated with industrial production (total index), GDP and population.

1.3.2 UNEMPLOYMENT RATE VISUALIZATION

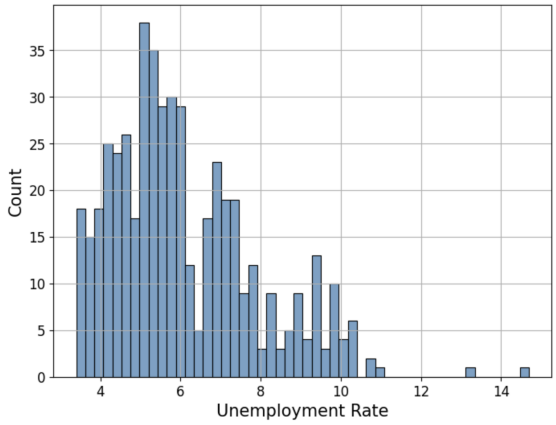

Figure 2. Unemployment Rate Visualization

We then plot the target variable "Unemployment rate", as shown in Figure 2.the original unemployment rate distribution is skewed. In order to reduce the skewness of data and make it easier to work with, natural log transformation will be applied on it at the part Data Cleaning.

1.4 Data Cleaning

Data cleaning involves addressing several issues that require attention. Firstly, the dataset contains a few missing values that require handling. These missing values may carry meaning or be inconsequential, necessitating different approaches to their treatment. Secondly, the U.S. recession probability is highly skewed, so we manage it into a categorical variable. By addressing these issues, the team can ensure the integrity and quality of the data used for subsequent analysis. Thirdly, since

the value of the features in our dataset vary a lot, so we standardized these features. Lastly, the unemployment rate distribution is highly skewed and we will use log transformation to deal with it.

1.4.1 NA Values

The GDP data obtained from FRED is updated on a quarterly basis, whereas the other data collected is available at a monthly frequency. To address this mismatch, the team has decided to use forward fill to fill in the missing values of the GDP data. This involves propagating the last observed value of the GDP data forward until the next observation, effectively filling in the missing values with the most recent available value. This approach helps to ensure that the dataset is complete and that all variables are aligned with the same temporal frequency.

1.4.2 Group US Recession Probability

The majority of values for the U.S. recession probability variable are below 0.2. In order to simplify the variable, we has chosen to categorize it by defining values less than 0.15 as 0 and values greater than 0.15 as 1. This approach allows for a clear distinction between low and high values of the variable, while still retaining its significance in the analysis. By categorizing the variable in this manner, we can more easily interpret its effects on the outcome variable.

1.4.3 Standardized Numerical Variables

Given the significant variance in the values of the numerical variables in our dataset, the team has decided to standardize them by subtracting their mean value and dividing by their standard deviation. By performing this operation, we have constrained the values of these features to the range of [-1,1]. This approach allows us to compare and contrast the impact of different variables on the unemployment rate more effectively, as it puts all features on the same scale.

1.4.4 Natural Log Transformation of Unemployment Rate

Figure 2 depicts the original distribution of the unemployment rate variable, which is characterized by skewness. To mitigate this skewness and improve the manageability of the data, the team has opted to apply a natural log transformation to the variable. Figure 3 demonstrates that this transformation has resulted in a more normally distributed variable. Additionally, the log transformation has reduced the impact of outliers, making the variable more representative of the data as a whole. By employing this approach, we can more effectively analyze the relationship between the unemployment rate and other variables, and draw more accurate conclusions from the data.

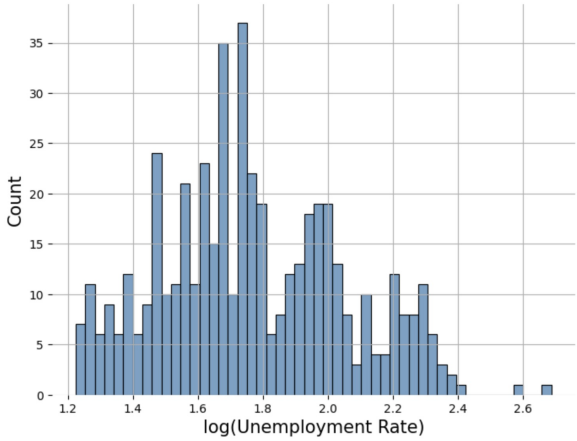

Figure 3. Log Unemployment Rate Visualization

2. MODEL SELECTION

2.1 Linear Regression

To establish a baseline for comparison with other models, we begin by creating a linear model that exclusively employs quadratic loss. The reasoning behind using a linear model with quadratic loss is straightforward: the explanatory variables in the dataset are expected to exhibit a linear correlation with the housing price. The goal function for this linear model is formulated as follows:

$$minimize \sum_{i=1}^{n} (y_i - w^T x_i)^2 \tag{1}$$

2.1.1 Linear Regression with PCA

Considering that some features are highly correlated, which might lead to multicollinearity in the linear regression, so we try to use PCA to mitigate this effect. It involves transforming a large set of correlated variables into a smaller set of uncorrelated variables called principal components, while retaining as much of the original variance in the data as possible.

PCA works by identifying the directions of greatest variance in the data and projecting the data onto these directions to form the principal components. The first principal component captures the most significant source of variation in the data, the second component captures the second-most significant source of variation, and so on. The goal function for PCA is formulated as follows:

$$minimize \|Y - XW\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (Y_{ij} - x_i^T w_j)^2 \tag{2}$$

2.1.2 Linear Regression with Bootstrap

We then try to use bootstrap, considering that there are limited data samples of our dataset, which might lead to the inaccuracy of linear regression. bootstrapping is a resampling technique that involves repeatedly drawing samples from our source data with replacement, often to estimate a population parameter. By “with replacement”, we mean that the same data point may be included in our resampled dataset multiple times. The procedure of Bootstrap is as follows:

1. Sample (x_i^k, y_i^k) with replacement from D, $i = 1, 2, ..m$, to form D_k

2. Estimate model $g_{D_k}: \mathcal{X} \rightarrow \mathcal{Y}$

3. Use it to make prediction for new input x

2.2 SVR

We then tried the SVR, since the linear regression assumes that there is linear relationship between the unemployment rate and the macroeconomic factors. But this assumption does not necessarily hold. SVR is particularly useful when dealing with non-linear data, as it allows for the use of non-linear kernel functions to transform the data into higher dimensions, making it easier to find a hyperplane that can accurately separate the data.

Additionally, SVR allows for the use of different types of loss functions to handle different types of errors, which makes it a versatile and flexible tool for regression analysis.

The objective function and constraints are as follows:

Minimize:

$$MIN \frac{1}{2} ||W||^2 \tag{3}$$

Constraints:

$$|y_i - w_i x_i| <= \epsilon \tag{4}$$

3. MODEL AND RESULTS

3.1 Linear Regression

3.1.1 Linear Regression Model

After data cleaning, we have 9 variables and 494 data entries. The variables are standardized and the unemployment rate is log transformed.

After running the linear regression, the R^2 is 0.55, with a MSE of 0.058. As Figure 4 shows, we can find that in general, the true unemployment rate vs predicted unemployment rate is linear correlated. But there exists many noise.

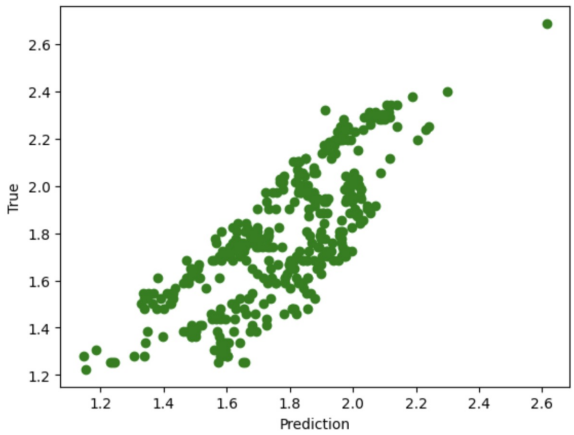

Figure 4. True value vs Predicted value-Linear

3.1.2 Linear Regression + PCA

After applying PCA and selecting the top 4 principal components, we attempted to improve our linear regression model by running a regression on the new variables. However, we did not observe any significant improvement in the performance of our model.

One possible reason for this lack of improvement could be that our dataset is relatively small, and we do not have enough features to capture the complexity of the underlying relationships between the variables.

It is worth noting that PCA is a powerful technique for multicollinearity, and can be particularly effective when applied to datasets with many features. However, its effectiveness can be limited in small datasets, as the amount of variance explained by the principal components may be relatively small.

3.1.3 Linear Regression + Bootstrap

Given the limitations of our small dataset, we applied bootstrap resampling to help us estimate the coefficients of the linear regression model. Specifically, we resampled our original dataset 1000 times, and for each resampled dataset, we ran a linear regression and estimated the coefficients. We then computed the mean of the coefficients across all resampled datasets as our final estimator.

Our approach using bootstrap resampling led to a slight improvement in the performance of our model. The R^2 value improved to 0.65 and the MSE also decreased to 0.028.

3.2 SVR

Given the limited performance of the linear regression model on our dataset, we explored Support Vector Regression (SVR). The results of our analysis, presented in figure 5, show that the SVR significantly outperforms the linear regression model.

Specifically, we observed that the true unemployment rate plotted against the predicted unemployment rate using the SVR model displays a strong linear correlation, indicating that the model is accurately capturing the underlying patterns in the data. The results of the four models is listed in the table below:

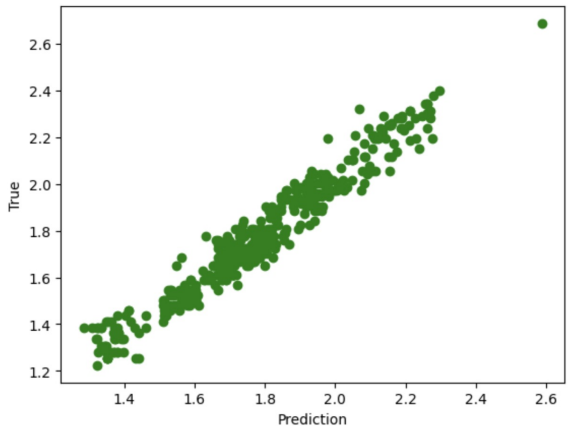

Figure 5. True value vs Predicted value-SVR

Model	R2	MSE	MAE
LR	0.520	0.038	0.174
LR+PCA	-12.632	3.121	1.748
LR+Bootstrap	0.653	0.028	0.153
SVR	**0.928**	**0.006**	**0.062**

We can see that the performance of SVR is significantly better than linear regression models.

4. Gender Analysis

We then examine the impact of macroeconomic factors on unemployment rates for men and women separately. To incorporate the gender aspect, we introduced a categorical feature, "Gender". We employed forward feature selection using both Linear Regression and SVR to identify the most influential features.

Forward selection is an iterative approach that starts with an empty model and gradually adds features that contribute the most to improving the model's performance. We continued this process until adding a new variable no longer enhanced the model's performance according to the metric, r-square score.

The results showed that Linear Regression achieved an R-square score of 0.52 with the inclusion of six best features, but gender was not among them. However, SVR performed significantly better with an R-square score of 0.937, and gender was among its top six features. This experimental outcome strongly suggests that gender indeed has an impact on the unemployment rate, a finding that is also visually evident from the accompanying figure 6. The rest best features are Population, Wage, Industrial Production Index, M2, OECD, and CPI, respectively.

5. Conclusion

In this study, we utilized Linear Regression and Support Vector Regression to examine the influence of macroeconomic factors on the unemployment rates for men and women in the United States. Our research expanded beyond market effects to incorporate sociological aspects, successfully demonstrating significant gender-based disparities in the unemployment rate.

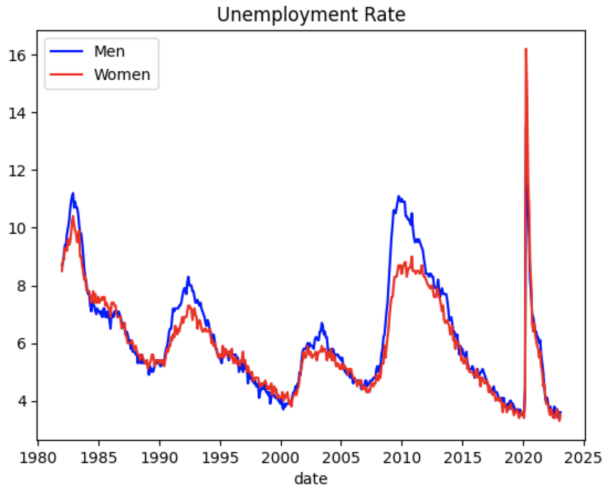

Figure 6. Unemployment rate: Men and Women

However, it is important to note that the correlation between gender and the unemployment rate is not consistently constant. Although men generally experience higher unemployment rates compared to women, we observed a striking exception during the COVID-19 pandemic. During this period, women faced unprecedentedly high levels of unemployment. We contemplated augmenting our original data set with additional features, such as a world pandemic uncertainty index. However, we encountered challenges in finding a suitable indicator that could provide long-term coverage compatible with our data set, which is already limited in size.

In the subsequent stages of our research, we intend to conduct sensitivity analysis to gain insights into how different input features contribute to the observed outcomes. Additionally, we will explore alternative models that may offer a better fit for our data set. Besides, to delve deeper into gender disparities, it will be crucial to incorporate a sociological perspective to elucidate the underlying causes of these differences.