Prediction of Unemployment Rates in the United States by K-Nearest Neighbor Regression Analysis

Yifu Tang [1], Jinghan Feng [2]

[1] Lehigh University, Pennsylvania, 18015, USA

[2] Faculty of Economics, Minzu University of China, Beijing,10081, China

Abstract. Unemployment remains a pervasive challenge on the global stage, bearing significant social, psychological, and economic ramifications. This issue touches the core of human existence, influencing not just the fundamental necessities of life but also personal aspirations and leisure activities. For nations on the cusp of economic growth, like the United States, unemployment is a formidable obstacle to achieving their growth aspirations. Addressing this concern requires a forward-looking approach, where predictions about future unemployment rates are made based on accessible data. This study embarks on such a mission, focusing on the unemployment rates in the United States. It employs the k-nearest neighbor regression (kNNR) to make these estimations. To enhance the precision of these forecasts, we have assembled a new dataset, including factors presumed to impact the behavior of unemployment. Encompassing factors believed to have a bearing on unemployment dynamics. To assess the effectiveness of the kNNR, we have juxtaposed its performance with that of another prominent machine learning contender: linear regression and decision trees. The empirical outcomes are telling with a coefficient of determination (R2 value of 0.936; the kNNR not only showcased its predictive prowess but also outshone the other algorithms in the fray. These findings underscore the potential of the kNNR algorithm as a potent tool in unemployment rate prediction endeavors.

Keywords: Unemployment rate, machine learning, regression, k-neighbors regression.

1. Introduction

The unemployment rate is a macroeconomic aggregate reflecting the state of the labor market [1]. It is a significant economic indicator for a country [2], and it affects a community's ability to cope with a disaster [3], international trade [4] thus leads to economic insecurity [5] and health and subjective well-being [6]. By taking a deeper dive into the macroeconomic vantage point, it's evident that unemployment generates a cascade of challenges, including a notable shrinkage in savings, a decline in social security inflows, and a ballooning of health-related expenditures. Beyond these immediate implications, long-term reverberations also manifest countries contend with deferred technological advancements, and their position in the global competitive market begins to wane, both in the current scenario and future projections. Besides, unemployment is affected by many factors, such as government size or expenditure measured as a percentage of national income[7], business cycles[8], and technology[9].

Given the overarching significance of unemployment, policymakers must base their economic strategies for the labor market on an authentic and accurate understanding of unemployment metrics. In this vein, illuminating the multifaceted characteristics of unemployment and charting out efficacious policies becomes an urgent need. Enter the realm of machine learning, which emerges as an advanced toolset promising to provide insightful estimations of unemployment rates, harnessing the power of relevant, granular data.

Pinpointing every single determinant of unemployment might seem like a Herculean task. Yet, some pivotal elements, like the annual growth rate of a population, undeniably intertwine with unemployment trajectories. Drawing from the example of the United States in 2019, where the population increased by 0.49%, a parallel uptick was observed in unemployment a rise of 4.38% [10]from the preceding year. This observation underscores a troubling reality: the growth curve of employment opportunities often lags behind that of population increments. Another beacon in the study of unemployment determinants is the industrial index. As a barometer for gauging economic

growth achieved through relentless industrialization, this index highlights the nexus between industrial growth and employment opportunities. As illustrated by crises, economic upheavals cast another long shadow, wielding direct and nuanced indirect influences on variables like unemployment.

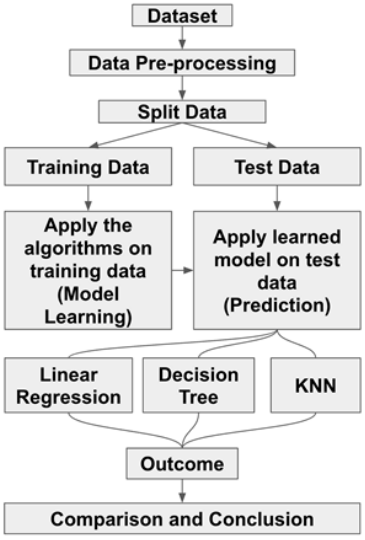

Fig. 1. Architecture Overview.

This process begins with the Dataset, which we subject to data pre-processing to clean and prepare the data. After pre-processing, we split the data into training data and test data. We then apply linear regression, decision trees, and kNN to the training data for model learning. In parallel, we apply the learned models from these algorithms to the test data for prediction. The results from these algorithms lead to the outcome. Finally, we compare the results of these methods and conclude to determine the best-performing method.

Research on the unemployment rate and macroeconomics started thriving since the mid of the mid-1990s[11]. A Yale Professor, Arthur Okun, found the relationship between economic growth and unemployment for the first time in 1962, and later, the upgrade to Okun's Law incorporated additional independent variables named the dynamic and the production function model. Over the years, the academic and research fraternity has embarked on a quest to decode and predict unemployment rates. This journey has been navigated using time-honored techniques and avant-garde methodologies. Mulaudzi and Ajoodha, for instance, amalgamated an array of methods from vector autoregression to a plethora of machine learning algorithms like SVR, applying them to data gleaned from South Africa. Their findings heralded the traditional approach of VAR, coupled with machine learning stalwarts like LTSM and GRU [12], as the most productive. Anjar Wanto and Irfan Sudahri Damanik use an artificial neural network application using a combination of the Levenberg-Marquardt Algorithm with bipolar sigmoid function to measure the open unemployment rate in indonesia[13].

Table 1. Details of the variables presented.

Variable	Mean	Std. Dev.	Min	Max
tvs	16.6	1.7	8.9	18.7
fsod	-105853.9	165826.1	-864074.1	308215.1
mvg	23494.8	4550.3	17241.5	30781.6
clf	160693.8	3244.1	154673	167839
p	327865.5	5175.5	317397	335501
dpi	15869.7	2361.3	12441.9	21858.1
ip	100.6	3.0	84.6	104.1

While many studies have cast their analytical nets over global unemployment landscapes, there needs to be more research on delving deep into the United States unemployment metrics through the lens of machine learning. This manuscript endeavors to plug this gap. At its core, it harnesses the k-

nearest neighbor regression (kNNR) algorithm, setting its sights on deciphering the United States monthly unemployment rhythms. His exploration further enriches by juxtaposing the kNNR's prowess with other contenders in the machine learning arena: ridge regression and linear regression. The dataset, meticulously curated and sourced from the official bureaucratic corridors, is a melange of factors surmised to be instrumental in sculpting unemployment contours.

To navigate through this comprehensive study, the paper is architecturally designed as follows: Section 2 unravels the intricacies of the dataset in question; Section 3 offers a panoramic view of the machine learning models at play; Section 4 delves into the empirical results, weaving together a narrative that contrasts the algorithmic models and their respective performances, culminating in the final section that encapsulates conclusions and proffers potential directions for prospective research endeavors.

2. Dataset

The dataset consists of 72 samples and seven features: total vehicle sales (tvs), federal surplus or deficit (fsod), the market value of gross federal debt (mvg), civilian labor force level (clf), population (p), disposable personal income (dpi) and industrial production (ip). The dataset consists of 121 data points, each corresponding to a specific month. Data are collected monthly from reliable and official government resources such as the U.S. Bureau of Economic Analysis and the Federal Reserve Bank of St. Louis between August 2013 and August 2023. Details of the features are presented in Table 1.

3. Methodology

Dataset $D = (x_i, y_i)$ where $i \in \{1, 2, \ldots, N\}$ with N samples is given. Here, x_i is the $n \times m$ input vector (predictor) and y_i is the ith value of the target variable.

3.1. Page Numbers

Linear regression is a statistical method employed to eluci-date the association between a solitary dependent variable (re-ferred to as y) and multiple independent variables (commonly denoted as x_1, x_2, x_1, ..., x_t). Linear regression is defined in Equation 1

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{it} + \epsilon_i \tag{1}$$

where β are the coefficients of the features and ε_i is the error term. The values of β are calculated by the analysis of the residual sum of squares (RSS) formula in Equation 2

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

The β coefficients are the measure of how much the output variable changes for each unit change in the predictor variable.

3.2. Decision Tree

Decision tree analysis is a method used for classification, em-ploying a divide-and-conquer strategy. By dissecting large databases, it identifies crucial features and patterns for dis-crimination and prediction. With roots in machine learning and artificial intelligence, decision trees are increasingly rel-evant in the chemical and biochemical domains. Their abil-ity to transparently depict decision-making processes makes them popular in various scientific domains.

Various metrics have been formulated to assess decision tree splits. In this brief overview, we'll concentrate on two key metrics: Information Gain (InfoGain) [14] and Gini Index (Gini) [15]. The Info concept is defined in Equation 3, where N_j represents the count of samples in class *j*, N(t) signifies the total samples in node j, and N_j(t) denotes the instances of class t in node *t*.

$$Info = -\sum_{j}\left(\frac{N_j(t)}{N(t)}\right)\log_2\left(\frac{N_j(t)}{N(t)}\right) \tag{3}$$

The value optimally enhancing the *information difference* is chosen and termed *InfoGain*. *InfoGain* is determined using Equation 4. Here, *Info(q)* denotes the information related to the feature subset *q*, while *pk* represents the fraction of sam-ples directed to the *kth* subset.

$$InfoGain = Info(Parent) - \sum_{k} p_k Info(Child_k) \tag{4}$$

The Gini Index quantifies the decrease in class impurity when partitioning feature space, as described in Equations 5 and 6. In a training set, $p(j)$ is the prior probability of a sample being in class *j*. The notation $||g||$ signifies the nor-malization of the vector g to a unit length. In the microarray context, proportional priors align with Matlab's default deci-sion tree settings, where $p(j) = \frac{N_j}{N(1)}$.

$$impurity = 1 - \sum_{j}\left|p(j)\frac{N_j(t)}{N_j}\right|^2 \tag{5}$$

$$Gini = impurity(Parent) - \sum_{k} p_k impurity(Child_k) \tag{6}$$

By comparing InfoGain and Gini scoring criteria for fea-ture selection, both scores reveal similar trends but can influ-ence which features are prioritized. While exact score values vary, their movements in binary classification are alike. The choice of criterion affects tree complexity and feature impor-tance, influencing model accuracy [16].

3.3. K-Nearest Neighbors Regression

In the k-nearest Neighbors Regression (kNNR) context, the target value is determined by averaging the values from the k-nearest neighboring samples. These nearest neighbors are identified based on a chosen distance metric. One commonly employed distance metric is the Minkowski distance, defined as in Equation 7

$$D(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{\frac{1}{p}} \tag{7}$$

where x_i is the predicted value. Here, x_i represents the pre-dicted value, and the Minkowski distance metric encompasses specific cases, such as Manhattan distance for $p = 1$, Eu-clidean distance for $p = 2$, and Chebyshev distance for $p = \infty$. Following the selection of the distance metric, the kNNR model computes the target value using the Equation 8

$$y = \frac{1}{k}\sum_{s=1}^{k} y_i \tag{8}$$

An essential parameter in kNNR is the user-defined k value, which signifies the number of nearest neighbor's con-sidered. However, determining the optimal k value remains challenging. For minimal values of k, such as *k=1*, the model may excessively fit the training data, resulting in elevated errors in the validation dataset. Conversely, when using high values of *k*, the model may fail to effectively fit either the training or validation data. This occurs because the sam-ples are typically distributed across the input data, and the *kth* nearest neighbor could be pretty distant from the query sample, thus significantly impacting the resulting mean. Ulti-mately, identifying the optimal *k* parameter is essential, as it reduces error rates in both the training and validation sets.

4. Results & discussion

4.1. K-Performance Criteria

Numerous performance metrics have been introduced for the assessment of machine learning techniques. Evaluating a model's performance hinges on its ability to provide predic-tions that closely align with the observed or actual values. Among these metrics, the coefficient of determination R2, mean squared error (MSE) and mean absolute error (MAE) are extensively employed in the machine learning domain.

The coefficient of determination, denoted as R2, indicates the accuracy with which the predicted values align with the true target values. Higher R2 values signify superior model performance. It is mathematically defined as Equation 9.

$$R^2 = 1 - \frac{\sum_{i=1}^{P}(y_i - \hat{y}_i)^2}{\sum_{s=1}^{P}(y_i - \bar{y})^2} \tag{9}$$

where $\bar{y}$ is the mean of the target, y_i is the actual value and $\hat{y}_i$ is the predicted value.

MAE is the average distance between each actual data point and the corresponding predicted value. MAE is defined as Equation 10.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

(best value = 0; worst value = +∞)

MAE is effective when outliers are due to data corruption, as it doesn't excessively penalize them. The L_1 norm moder-ates the impact of potential outliers, offering a balanced per-formance metric for the model. However, if the test set con-tains numerous outliers, the model's performance might be average [17].

MSE measures the closeness of the regression line and the actual data points. This approach regulates large forecasting errors [18]. MSE is computed in Equation 11.

$$\text{MSE} = \frac{1}{n}\sum_{s=1}^{n}(y_i - y_i)^2 \tag{11}$$

(best value = 0; worst value = +∞)

MSE is suitable when detecting outliers is essential. Due to the L_2 norm, MSE emphasizes larger weights for these points. Notably, if the model produces a significantly poor prediction, the squared aspect of the function amplifies the error [17].

MAE and MSE are metrics to gauge the errors in a model's predictions, with lower values indicating a more effective model. Data is typically partitioned into training and testing subsets to assess the efficacy of machine learning methods. The model is constructed using the training data, while its performance is evaluated using the test data. In conventional practices, this data separation is accomplished randomly, which may not always yield a reliable measure of the model's success.

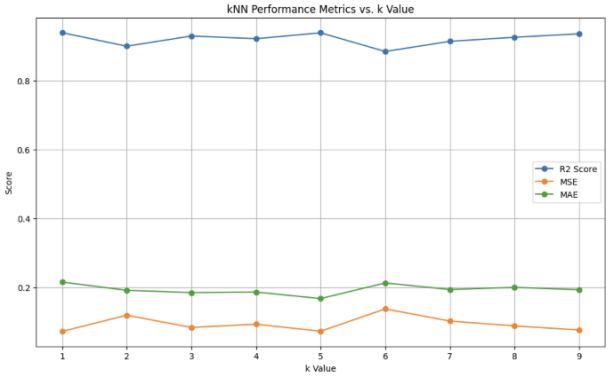

Fig. 2. Performance Metrics of kNN Across Varying k Values.

The chart depicts the kNN algorithm performance across k values from 1 to 9. The chart helps assess optimal k values. To address this limitation, k-fold cross-validation has been developed to assess the performance of regression methods independently of the specific data split [19]. In k-fold cross-validation, the dataset is randomly divided into k equally sized groups or folds. One-fold is reserved for testing, while the remaining k-1 folds are used for training the model. This training and testing process is iterated k times using a different fold as the test set. The method's performance can be more accurately and robustly assessed by averaging the results obtained from these iterations.

4.2. Experimental Results

This research study aims to estimate unemployment rates in the United States using the k-nearest neighbors' regression (kNN) algorithm and subsequently compare the results with outcomes from two distinct machine learning techniques. The dataset employed comprises seven attributes and a total of 72 samples. To assess the performance of these models, three evaluation metrics are used: R-squared (R^2), mean absolute error (MAE), and mean squared error (MSE).

To ensure robust and reliable results, a k-fold cross-validation method with 20 iterations, where $k = 5$, is applied [19]. This approach allows for a thorough assess-ment of model performance. Figure 2 depicts the changes in R^2, MAE, and MSE concerning the choice of k for the kNNR algorithm. Consequently, the optimal k value is 5.

The performance outcomes of the models are summarized in Table 2, with the best results highlighted in bold. Re-markably, kNNR consistently outperforms the benchmarking methods across all three-performance metrics. This supe-rior performance can be attributed to the inherent advantage of kNNR, which is particularly beneficial when the relationship between the target and predictor variables is uncertain, whether linear or nonlinear. In cases where the predictor-target relationship is nonlinear, kNNR surpasses linear regres-sion, demonstrating its effectiveness in handling such data. Hence, kNNR emerges as the most suitable choice for this specific dataset.

Table 2. Comparison of Model Performance Metrics

Methods	R-Squared	MAE	MSE
Linear Regression	0.615	0.570	0.461
Decision Tree	0.723	0.256	0.331
kNN	**0.936**	**0.180**	**0.069**

5. Conslusion

Unemployment is a significant concern for nations due to its implications on human existence's societal, emotional, and fi-nancial facets. In this research, the kNNR algorithm was em-ployed to predict unemployment rates in the United States, and the outcomes were juxtaposed with time series analyses. A rigorous 5-fold cross-validation was performed over 20 it-erations to enhance the reliability of the results. The kNNR regression demonstrated superior efficacy, with an R^2 value of 0.936, outpacing both linear regression and decision tree in terms of performance. Thus, kNNR stands out as a viable technique for forecasting unemployment rates. Future stud-ies could focus on two primary areas: incorporating alterna-tive machine learning approaches or devising novel prediction methods and delving deeper into the causative factors of un-employment to enrich the dataset.

Acknowledgement

The two authors work together and contribute equally to this paper.

References

[1] H.A. Ahmed and T. Nasser, "Long-run relationship between the unemployment rate and the current account balance in the United States: An empirical analysis," *Port Econ J*, vol. 22, pp. 397–416, 2023.

[2] Olivier J Blanchard and Daniel Leigh, "Growth forecast errors and fiscal multipliers," *American Economic Review*, vol. 103, no. 3, pp. 117–120, 2013.

[3] Shichao Tang, Libby Horter, Karin Bosh, Ahmed M Kassem, Emily B Kahn, Jessica N Ricaldi, Leah Zilversmit Pao, Gloria J Kang, Christa-Marie Singleton, Tiebin Liu, et al., "Change in unemployment by social vulnerability among united states counties with rapid increases in covid-19 incidence—july 1–october 31, 2020," PLoS *One*, vol. 17, no. 4, pp. e0265888, 2022.

[4] Jonathan Eaton*, Samuel Kortum**, and Brent Neiman***, "On deficits and unemployment," *Revue économique*, vol. 64, no. 3, pp. 405–420, 2013.

[5] Martin Ehlert, "Job loss among rich and poor in the United States and germany: who loses more income?" *Research in Social Stratification and Mobility*, vol. 32, pp. 85–103, 2013.

[6] Heikki Ervasti and Takis Venetoklis, "Unemployment and subjective well-being: An empirical test of deprivation theory, incentive paradigm and financial strain approach," *Acta Sociologica*, vol. 53, no. 2, pp. 119–139, 2010.

[7] Damodar Nepram, Salam Prakash Singh, and Samsur Jaman, "The effect of government expenditure on unemployment in india: A state level analysis," *The Journal of Asian Finance, Economics and Business*, vol. 8, no. 3, pp. 763–769, 2021.

[8] Naceur Khraief, Muhammad Shahbaz, Almas Heshmati, and Muhammad Azam, "Are unemployment rates in oecd countries stationary? evidence from univariate and panel unit root tests," *The North American Journal of Economics and Finance*, vol. 51, pp. 100838, 2020.

[9] Florent Bordot, "Artificial intelligence, robots and unemployment: Evidence from oecd countries," *Journal of Innovation Economics & Management*, , no. 1, pp. 117–138, 2022.

[10] Mark Mather, Linda A Jacobsen, Beth Jarosz, Lillian Kilduff, Amanda Lee, Kelvin M Pollard, Paola Scommegna, and Alicia Vanorman, "America's changing population: what to expect in the 2020 census.," *Mycosystema*, vol. 38, no. 8, 2019.

[11] Despina Tumanoska, "The relationship between economic growth and unemployment rates: validation of okun's law in panel context," *Research in Applied Economics*, vol. 12, no. 1, pp. 33–55, 2020.

[12] Rudzani Mulaudzi and Ritesh Ajoodha, "Application of deep learning to forecast the south african unemployment rate: a multivariate approach," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, 2020, pp. 1–6.

[13] Zulaini Masruro Nasution, "Levenberg-marquardt algorithm combined with bipolar sigmoid function to measure open unemployment rate in indonesia," 2021.

[14] S.L. Salzberg, "C4.5: Programs for machine learning by j. ross quinlan," *Mach Learn*, vol. 16, pp. 235–240, 1994.

[15] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1984.

[16] Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004.

[17] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, pp. e623, 2021.

[18] S Prayudani, A Hizriadi, YY Lase, Y Fatmi, et al., "Analysis accuracy of forecasting measurement technique on random knearest neighbor (rknn) using mape and mse," in *Journal of Physics: Conference Series*. IOP Publishing, 2019, vol. 1361, p. 012089.