Predicting the need for COVID-19 Welfare Programs based on Monthly Unemployment Rates in the USA

I. EXECUTIVE SUMMARY

This paper explores the relationship between unemployment rates and welfare assistance programs, specifically during the COVID-19 pandemic. The main goal of the paper is set in a hypothetical situation where the unemployment rate per month determines if the welfare programs during the COVID-19 pandemic should pass. The hypothetical situation also extends to months before the COVID-19. The motivation behind this is to avoid dependency on welfare programs, which will be expanded on in the background and introduction section.

For this project, I used two machine learning techniques - one parametric and one non-parametric - to create a predictive model for my analysis. These models were tested on data from the U.S. Bureau of Labor Statistics for the years 2015 to 2022. From my analysis, I discovered that the models were not the best fit to make these predictions of the rollout of the welfare programs based on the monthly unemployment rates.

II. BACKGROUND

Past research has shown that welfare programs in the United States has reduced the poverty rate and offered assistance to the disadvantaged in society. Such welfare programs are provided by the government, and they include food assistance programs, housing subsidies, unemployment benefits, among others. Many of these programs are aimed towards low income families/individuals, people living with disabilities, and the elderly. Evidence shows that these welfare programs have cut poverty rates in the country by 4 percent in 1967 and by 43 percent in 2017. (Trisi & Saenz, 2019). This shows a positive impact of welfare programs on poverty reduction. With the COVID-19 pandemic, there was an increase in the need for food assistance due to a rise in the unemployment rate from 4.4% in March 2020 to 14.7% in April 2020. (Hodges, Jones & Toossi, 2021). Research has also shown that welfare pays better than work for millions of Americans. (Dublois & Inglam, 2021). This may lead to a dependency on welfare

programs. Looking at this argument, people with little marketable skills for work may deem it fit to live off government transfer payments and assistance, rather than work a low wage job. (Kenworthy, 1999). With this information, my research question is to predict the need of social welfare programs based on unemployment rates. This is a hypothetical situation that highlights the importance of the COVID-19 pandemic on unemployment rates, and uses the possibility of welfare dependency to predict the need for these welfare programs.

Following the COVID-19 pandemic, there was a decrease in employment opportunities due to instability in the economy. This led to multiple job layoffs and reduced rates of hiring. To recover the effects of the pandemic, the U.S. government introduced economic relief packages, like the American Rescue Plan, which comprised of unemployment benefits, support for small businesses, and general economic compensations. (U.S. Department of Treasury, 2022). From this information, this research paper will infer the relationship between unemployment rates before and after the pandemic and a model to predict the rollo out of the welfare programs based on the unemployment rate. This question arises from different ways to reduce dependency on welfare programs, and find alternatives to boost employment rates. It puts different factors into account. These include the need to increase employment rates in the country, increase in federal and state's minimum wage, and also to expand on improving welfare programs for them to be accessible to those who really need them. The research question also weighs in factors that influence individuals to obtain welfare assistance.

III. DATA

The data that will be used in this paper is data from the U.S. Bureau of Labor Statistics. This data from this source will highlight the unemployment rates and job openings from January 2015 to September 2022. This will show the rate of unemployment before and after the COVID-19 pandemic and the number of job openings in the country before and after the pandemic. There will also be a binary variable representing the introduction of the welfare programs.

Prior to this memo, the Survey of Income and Program Participation data was to be used for the analysis. However, further analysis of the datasets showed a lot of

limitations to the dataset, such as insufficient data collected from surveys, numerous null values in the dataset and in important variables for the analysis. The dataset from the U.S. Bureau of Labor Statistics in this case is more suitable and efficient for this topic of analysis.

Table 1: Descriptive Statistics of Variables

Variable Name and Label	Description	Mean	Std. Dev	Min	Med	Max
Total Unemployment Rate	Total U.S. unemployment rate per month	4.94	1.89	3.5	4.4	14.7
Total nonfarm job openings	Total number of nonfarm job openings per month	7,274,022	1,902,163	4,709,000	6,818,000	11,855,000
Total nonfarm hires	Total number of nonfarm hires per month	5,770,710	562,138.8	4,031,000	5,704,000	8,145,000
Total private job openings	Total number of private/non-governmental job openings per month	6,583,731	1,733,625	4,068,000	6,120,000	10,812,000
Total private hires	Total number of private/non-governmental job hires per month	5,411,333	552,992.1	3,805,000	5,335,000	7,890,000
Government job openings	Total number of government job openings	690,279.6	182,170	470,000	658,000	1,105,000

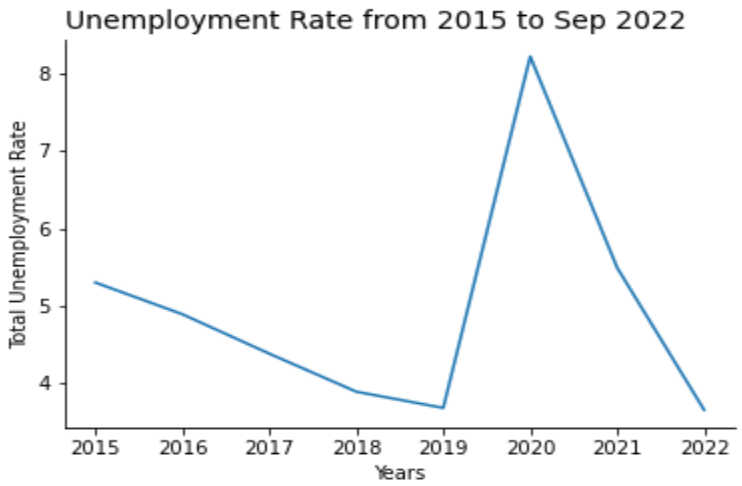
Government hires	Total number of government hires	359,430	40,197	227,000	355,000	576,000

The table shows descriptive statistics for the important variables, which are for unemployment rate, nonfarm job openings and hires, and private job openings and hires. The observations for this dataset are limited to 93. While this is a small number, it is important to highlight that the unemployment rates are counted per month and cannot be counted per day, hence the small number of observations. This dataset is also extracted from government data, which in this case, is limited to the years 2015 to 2022. This is important to highlight the effect of the pandemic overtime on the unemployment rate and employment opportunities in the country.

In the table, we see average unemployment rate from years 2015 to 2022 is 4.94%. The maximum unemployment rate in this case is 14.7%, which hypothetically occurred during the COVID-19 pandemic. There is also a reduction in job openings as we see a huge difference between the minimum and maximum values of nonfarm and private job openings in the table. We also see the gaps between the job openings and hires. For example, there was a maximum of 10,812,000 total private job openings, however, there was only a maximum of 7,890,000 total private job hires. This leaves 2,922,000 private jobs that were not filled. These may have been due to different reasons such as cost reasons for the firm/company, or inability to fill the role.

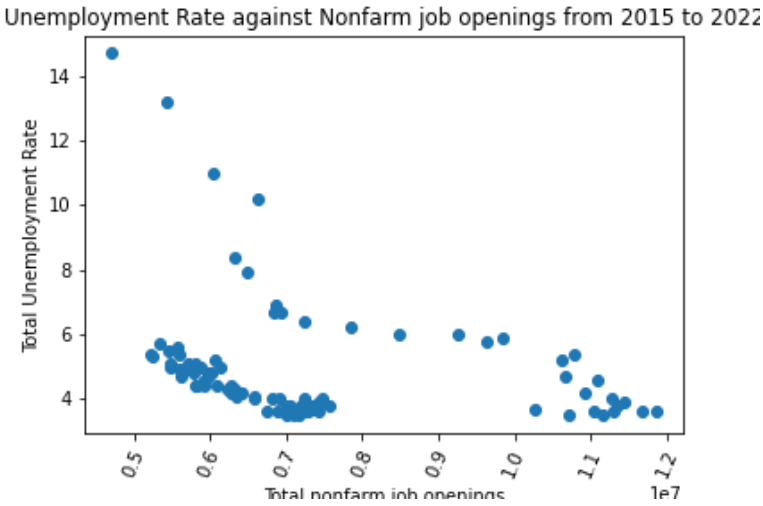
Data Visualization

PLOT 1



This plot shows the rate of unemployment from January 2015 to September 2022. The unemployment rate is at its minimum point below 4% in the year 2019. At the height of the COVID-19, there is a spike in the unemployment rate at 14%, which goes down slowly overtime. This shows the severity of the pandemic on the labor market.

PLOT 2



This plot shows the unemployment rate against total nonfarm job openings in the US, from January 2015 to September 2022. It shows that a reduction in job openings at about 300,000 jobs matched a high unemployment rate of more than 14%. A limitation to this graph is that it does not show all job openings in the US, but it still shows a good representation of the state of the labor market during this time period.

IV. METHODOLOGY

The variables that will be in the feature matrix will be unemployment rate, total nonfarm job openings and hires, total private job openings and hires, government job openings and hires. The target array will be the binary variable representing the rollout of welfare programs, 'Welfare_01.'

The two machine learning techniques that will be used for this project are Logistic Regression and the K-Nearest Neighbors.

Modeling Techniques

Logistic regression is used when modeling a dependent variable 'Y', in this case, the welfare variable, and independent variables 'X'. The dependent variable in this case is usually categorical in nature. To evaluate performance of logistic regression on test data, we use a confusion matrix. A confusion matrix will help determine the accuracy of the logistic regression model. It also measures the rate of error in the model. I chose this technique because it will allow me make a possible prediction of the roll out of the COVID welfare programs. This variable is binary and categorical, and is also based on the healthy unemployment rate index of the United States by economists, which is 5% (Hankin, 2020). The variable will be '0' if the unemployment rate is less than 5%, which means that there should be no rollout of the welfare programs, and the variable will be '1' if the unemployment rate is more than 5%, which means there should be a rollout of the welfare programs.

It will also show relationships between the unemployment rate and COVID welfare programs in the U.S. An example of how logistic regression has been applied is making predictions on the classification of 'Survival or deceased' from the Titanic dataset. Here, the logistic regression model was used to determine if someone survived or died from the sinking of the Titanic. (Remanan, 2018). The data will show a hypothetical situation, where the predictions will assign the best period for the COVID-19 welfare packages, based on the unemployment rate in the country. This shows that logistic regression will be the best option, because it predicts categorical variables. In this project, the unemployment rate for assigning the welfare program will be at 5, since that is suggested as a healthy unemployment rate for the US, according to economists. (Hankin, 2020).

The second model I will be using will the the K-Nearest Neighbors model. The K-Nearest Neighbors is used to estimate the likelihood of a datapoint being a member of a group, based on the other datapoints that surrounds it. It is non-parametric in nature and it is a supervised algorithm. It also requires assumptions such as defining metrics to calculate distance between data points. The most used distance metrics are Euclidean distance and Manhattan distance.

For the project, this can be used to predict the likelihood of the COVID-19 welfare variable (Y) belonging to a classification, based on the unemployment rate and other variables in the feature matrix. An example of K-Nearest Neighbors is predicting whether a customer will purchase a product or not based on information about the customers such as age, gender and salary. This uses data from Social Network Ads. After splitting into training and test data, the model showed that it performed well in classifying the labels, but a few were wrongly classified. (Zoltan, 2018).

V. ANALYSIS

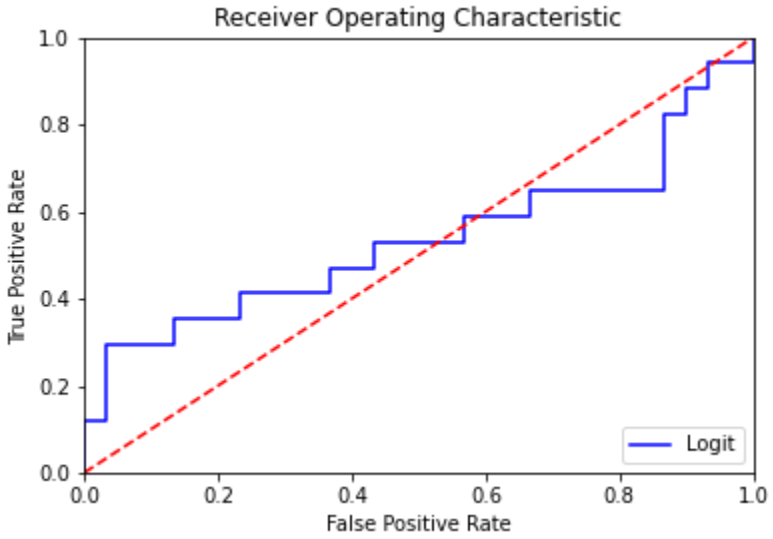
Logistic Regression

For the logistic regression, I first split the data into training and test data, holding out 50% for the test size, leaving 50% for the training size. I then made predictions with the test data using 'predict_proba'. To check the performance of my model, I constructed

a confusion matrix, where 29 positive class data points were correctly classified by the model and 2 negative class data points were correctly classified by the model. Only 1 negative class data point was incorrectly classified as belonging to the positive class by the model and 15 positive class data points were incorrectly classified as belonging to the negative class by the model.

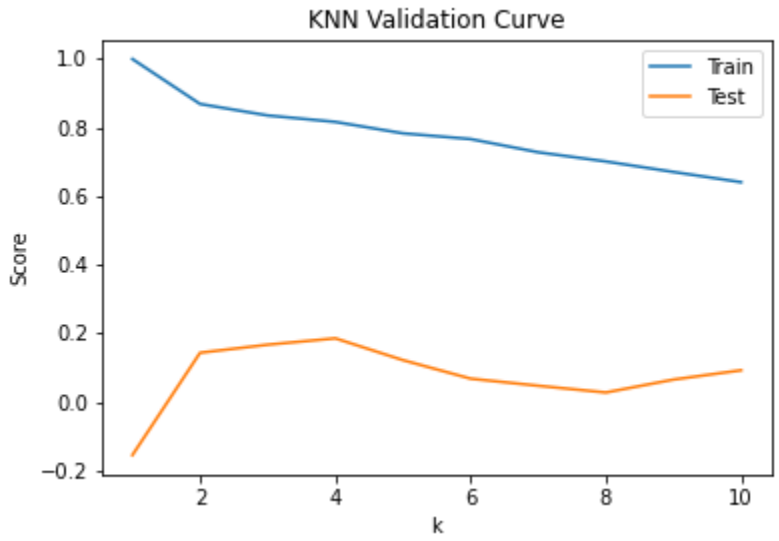
Confusion Matrix	Positive	Negative
Positive	29	1
Negative	15	2

I then got the accuracy score of the test and training data, which was about 66% for both. This shows the model does a fair job of predicting the binary variable, but possible improvements can be made to this, which will be expanded in the conclusions section of the paper.



K-Nearest Neighbors

For K-Nearest Neighbors, I first scaled the data in the dataframe since KNN uses Euclidean points to find the nearest neighbors. This is done to make the numbers in the dataframe of equal magnitude in order to find the nearest neighbors. After scaling, I split the test and training data, with a test size of 20%. The accuracy score of the model was 65%, which is fair.



This shows that the train data did somewhat better than the test data, since the train score is higher than 0.6, and the test score is below 0.2.

Both the logistic regression model and K-Nearest Neighbors model have similar accuracy scores. This shows that the two models did not do as well in predicting the binary variable.

VI. CONCLUSION

This analysis shows that there is indeed a relationship between the unemployment rate and the economic state of the U.S., in this case, the COVID-19 pandemic. The models in this analysis were to predict the need and rollout for necessary welfare programs in order to reduce dependency on these programs. Researchers have found that people tend to depend on welfare programs, which may be bad for the economy and also unfair for people that really benefit from

these welfare programs. The models therefore were to predict the rollout of welfare programs based on the unemployment rate for that month. This is a hypothetical situation that could possibly be used to help policymakers reduce dependency on welfare programs, as well as reduce cost attached to welfare programs. The validity of the models were not as strong and this may be for a number of reasons.

For the ROC curve showing the logistic regression model, the curves do not lean towards the true positive rate. This is because there were few data points for the model to predict. As stated earlier, the nature of this analysis needed a few data points, since the analysis looks at each month from 2015 to 2022. To fix this issue, a different modeling technique may be suitable to use for these predictions, especially a modeling technique fit for a few data points. Another alternative improvement to the logistic regression would be to include more variables in the analysis and in the feature matrix. This will ultimately allow for more data points to be analysed for the prediction model.

For the K-Nearest Neighbors, a possible improvement would be to rescale the data using 'Robust Scaling'. This will increase the accuracy score of the predictive model. This is best to overcome outliers in the data. Another form of scaling can be the min-max scaling method, which uses the minimum and maximum values of a feature to rescale the values within a range. (Martulandi, 2019). Finally, like the improvement mentioned for the logistic regression, new variables could be added to the feature matrix in order to add more data points for the analysis.

BIBLIOGRAPHY

Dublois, H., & Ingram, J. (2021). How the new era of expanded welfare programs is keeping Americans from working. *Foundation for Government Accountability.* https://thefga.org/wp-content/uploads/2021/11/Benefit-Stacking-paper-11-11-21.pdf

Hankin, A. (2020). The Downside of Low Unemployment. *Investopedia.* https://www.investopedia.com/insights/downside-low-unemployment/

Hodges, L., Jones, J. W., & Toossi, S. (2021). Coronavirus (COVID-19) Pandemic Transformed the US Federal Food and Nutrition Assistance Landscape. *Amber Waves: The Economics of Food, Farming, Natural Resources, and Rural America*, *2021*(1490-2021-1567). https://www.ers.usda.gov/amber-waves/2021/october/coronavirus-covid-19-pandemic-transformed-the-u-s-federal-food-and-nutrition-assistance-landscape/

Kenworthy, L. (1999). Do social-welfare policies reduce poverty? A cross-national assessment. *Social Forces*, *77*(3), 1119-1139. https://www.jstor.org/stable/3005973

Martulandi, A. (2019). Increase 10% Accuracy with Re-scaling Features in K-Nearest Neighbors + Python Code. *DataDrivenInvestor.*

https://medium.datadriveninvestor.com/increase-10-accuracy-with-re-scaling-features-in-k-nearest-neighbors-python-code-677d28032a45

Remanan, S. (2018). Logistic Regression: A Simplified Approach Using Python. *Towards Data Science.*https://towardsdatascience.com/logistic-regression-a-simplified-approach-using-python-c4bc81a87c31

Trisi, D., & Saenz, M. (2019). Economic Security Programs Cut Poverty Nearly in Half Over Last 50 Years. *Center on Budget and Policy Priorities, updated November*, *26*. https://www.cbpp.org/research/poverty-and-inequality/economic-security-programs-cut-poverty-nearly-in-half-over-last-50

U.S. Bureau of Labor Statistics: https://data.bls.gov/apps/covid-dashboard/home.htm

U.S. Department of Treasury. (2022). Fact Sheet: The Impact of the American Rescue Plan after One Year. https://home.treasury.gov/news/press-releases/jy0645

Zoltan, C. (2018). KNN in Python. *Towards Data Science.* https://towardsdatascience.com/knn-in-python-835643e2fb53