US Unemployment Rate forecasting

CS501 Computing for Science and Engineering - Fall 2023

Russell Ho, Prathyush Kothari, Lucas Mori (team leader), Darshitaben Patel

September 1st, 2023

1 Problem Statement

Developing an accurate unemployment rate forecasting model is essential to address the challenges posed by dynamic labor markets and facilitate informed decision-making for individuals, businesses, and policymakers.

Why this problem is important
Unemployment rate forecasting is a critical component of economic planning and policy formulation. Accurate predictions of unemployment rates provide governments, businesses, and job seekers with essential insights for making informed decisions. However, the inherent complexity of labor markets, influenced by various economic, social, and political factors, presents a significant challenge for reliable forecasting.

Existing forecasting models often struggle to capture the intricate dynamics of unemployment rates. These limitations hinder the ability to anticipate changes in job markets effectively. In today's rapidly evolving economic landscape, characterized by globalization, technological advancements, and unforeseen disruptions like pandemics, the need for precise and adaptable unemployment rate forecasting models is more pressing than ever.

This research aims to address this challenge by developing an advanced forecasting model for unemployment rates. By integrating historical labor market data with relevant economic, demographic, and social indicators, this study seeks to create a model that can provide more accurate and timely forecasts. The primary objective is to equip policymakers, businesses, and individuals with the necessary tools to navigate labor market fluctuations and make proactive decisions.

In pursuit of this goal, this research paper aims to:

1. Develop two unemployment rate forecasting model that considers a comprehensive set of factors influencing the job market, using LSTM and ARIMA models.
2. Assess each models performance, highlighting its strengths and weaknesses.
3. Identify useful external data that can be used to improve the unemployment rate forecast.

Through these efforts, this research endeavors to contribute to the advancement of unemployment rate forecasting, facilitating more informed labor market strategies and policies for governments, businesses, and individuals.

2 Related Works

Asymmetric behavior in unemployment rates have been documented since the post-war era of the United States, increasing the complexity of forecasting techniques. Conventional linear forecasting methods are not able to accurately predict the myriad of patterns found within different macroeconomic indicators, hence suggesting that improvement over conventional linear forecasts can be made through the use of nonlinear time-series models (Rothman). Past research on forecasting performances for a variety of linear and nonlinear time series models have been done including various different factors, such as during periods of economic expansions/contractions by exploiting the asymmetric cyclical behavior of unemployment numbers. The vector models used in past research have incorporated factors such as using initial jobless claims as a leading indicator, utilizing monthly rates for forecasting quarterly rates, and focusing on features of trends evident during specific periods of time. Additionally, these forecasts of nonlinear models have been combined with consensus forecasts, showing significant improvements in prediction accuracy over other methods (Montgomery et al. Forecasting the U.S. unemployment rate). In general, although machine learning techniques have been favored for their accuracy over other conventional methods, these techniques has not been adopted widely yet. Exceptions are Two Sigma, who used taxi data to forecast the unemployment rate of the New York City region and studies like the one done by Xu, Li and Chen who used Google searches for the term unemployment rate to create a unique method for forecasting; these techniques although unconventional, have only performed slightly better than the mean SPF forecast. Speaking of more traditional computational methods proposed, such as utilizing machine learning to select specific factors to be taken into account (Kreiner and Duca Can machine learning on economic data better forecast the unemployment rate?). These techniques have been proven to outperform the SPF over one-, two- and four-quarter horizons suggesting ample room for further research in machine learning computational methods.

This study's primary objective is to predict economic indicators, specifically national unemployment rates, by employing an ARIMA (Autoregressive Integrated Moving Average) model for time series forecasting. To enhance the precision and efficiency of the time series analysis while minimizing data noise, both linear and nonlinear smoothers were employed. Utilizing Python, this research visualized the data through time-series decompositions, enabling the dissection of time series into trend, seasonality, and noise components. In forthcoming research, we plan to extend our analysis to encompass multi-class classification. To achieve this, it is essential to augment the number of employment status labels, with a specific emphasis on the "unemployed" category, to ensure dataset balance. By distinguishing between employed, non-employed, and retired classes, we aim to offer a more nuanced and comprehensive portrayal of an individual's unemployment status. This refined information holds potential value for relevant authorities and utility providers, showcasing the depth of insights that can be gleaned from the analysis of datasets generated by autonomous IoT systems in the context of the Industry 4.0 revolution.

3 Approach

For the purpose of this project, we will be utilizing the dataset available on Kaggle titled "USA Key Macroeconomic Indicators" that contains data retrieved from FRED (Federal Reserve Bank of St. Louis) that was released by the U.S. Bureau of Economic Analysis. This dataset provides a comprehensive collection of key macroeconomic indicators for the United States, spanning from 1960 to 2021. One of the primary indicators present in this dataset is the unemployment rate. The unemployment rate is a vital economic metric that represents the percentage of the labor force that is jobless but actively seeking employment. Historically, fluctuations in the unemployment rate have been influenced by various factors, including economic downturns, technological advancements, and policy changes. By analyzing this dataset, we aim to gain insights into the patterns and factors affecting the unemployment rate over the years, which will be instrumental in building and refining our predictive models.

The following step-by-step approach will be adopted:

Figure 1: Steps of our approach

1. **Data Collection and Cleaning**:
   - Obtain historical data for US unemployment rates. This data will preferably span several decades to account for various economic cycles.
   - Integrate external factors (e.g., economic indicators, policy changes) that could provide additional context and potentially improve the model's accuracy.
   - Clean the data to remove any outliers or missing values, ensuring that it is consistent and ready for analysis.
2. **Exploratory Data Analysis (EDA)**:
   - Visualize the data to identify patterns, trends, seasonality, and cyclic behaviors.
   - Determine statistical properties of the data like mean, variance, and autocorrelations.
   - Evaluate and choose potential external factors, such as GDP growth, inflation rates, and policy changes, based on their historical influence on the unemployment rate.
3. **Data Preprocessing**:
   - Split the data into training and test sets. The training set will be used to train the models, while the test set will be reserved for model evaluation.
   - Normalize or standardize data if necessary, especially for LSTM, which is sensitive to the scale of input data.
4. **Model Implementation**:

- *SARIMAX Model*:
  - Use the Seasonal Autoregressive Integrated Moving Average (SARIMAX) model, which is well-suited for time series data with trends and seasonality.
  - Identify the best parameters (p, d, q)(P, D, Q)s for the SARIMAX model using grid search or similar techniques.
  - Train the SARIMAX model on the training dataset.
- *LSTM Model*:
  - Use the Long Short-Term Memory (LSTM) model, a type of recurrent neural network (RNN), to model the unemployment rate.
  - Structure the data into suitable time steps for LSTM input.
  - Design the LSTM architecture, choosing the number of layers and units.
  - Train the LSTM model on the training dataset.

5. **Model Evaluation**:
   - Use the test set to evaluate the predictive performance of both the SARIMAX and LSTM models.
   - Compare the forecasts of both models with the actual values from the test set.
   - Compute performance metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and others to quantify the accuracy of each model.

4 Deliverables

Given the steps outlined for the project, these are the expected outcomes at each stage and the overall expectations from the project.

1. **Data Collection and Cleaning**
   - **Outcome**: An organized dataset containing relevant variables for unemployment rate prediction. The data should be devoid of any inconsistencies, missing values, or anomalies which could potentially affect the forecasting. This process would produce a 'clean' dataset, ready for exploratory data analysis.
2. **Exploratory Data Analysis (EDA)**
   - **Outcome**: Insights about the distribution, correlation, and patterns in the data. Through various visualizations like scatter plots, histograms, and heat maps, one should be able to discern underlying trends, seasonality, or any cyclic behaviors in the unemployment data. This step might also bring forth any outliers or unique data characteristics that were missed during the data cleaning process.
3. **Data Preprocessing**
   - **Outcome**: A dataset that's suited for training machine learning models. This would involve:
     - Normalization or standardization: Ensuring data is on a similar scale.
     - Feature engineering: Extracting or combining features to improve model performance.
     - Train-test split: Segregating data into training and testing sets to validate the forecasting accuracy.
     - Time series data preparation: For LSTM, this would involve creating sliding windows or sequences of data points to predict the next point in the series.
4. **Model Implementation**
   - **Outcome**:
     - **SARIMAX**: A configured SARIMAX model with optimized hyperparameters.
     - **LSTM**: A deep learning architecture designed to capture long-term dependencies in the unemployment data. The model should be set up with considerations for the number of layers, neurons, dropout rates, etc., based on the nature of the data.
5. **Model Evaluation**
   - **Outcome**: Quantitative metrics showcasing the performance of each model in forecasting the unemployment rate. Common metrics for this could include:
     - Mean Absolute Error (MAE)
     - Mean Squared Error (MSE)
     - Root Mean Squared Error (RMSE)
     - Forecasting accuracy

  - Visualization of actual vs. predicted values to understand the model's capability visually.

Overall Project Outcomes:

- A deep understanding of the dynamics of the unemployment rate based on historical and external data.
- Two robust forecasting models (LSTM & SARIMAX) capable of predicting future unemployment rates.
- A foundation upon which further modeling techniques or data sources can be integrated to refine the forecasting process.
- Recommendations and insights which could be valuable for policymakers, economists, and researchers interested in the unemployment rate's behavior and trajectory.

5 References

Kreiner, Aaron, and John Duca. \Can machine learning on economic data better forecast the unemployment rate?" Applied Economics Letters, vol. 27, no. 17, 2019, pp. 1434{1437, https://doi.org/10.1080/13504851.2019.1688237.

Montgomery, Alan L., et al. \Forecasting the U.S. unemployment rate." Journal of the American Statistical Association, vol. 93, no. 442, 1998, pp. 478{493, https://doi.org/10.1080/01621459.1998.10473696.

Rothman, Philip. \Forecasting asymmetric unemployment rates." Review of Economics and Statistics, vol. 80, no. 1, 1 Feb. 1998, pp. 164{168, https://doi.org/10.1162/003465398557276.

Sidhari Manasa1, M.Kalidas2 , \UNEMPLOYMENT RATE FORECASTING USING SUPERVISED MACHINE LEARNING MODEL" © 2022 IJCSPUB | Volume 12, Issue 4 October 2022 | ISSN: 2250-1770 | https://ijcspub.org/papers/IJCSP22D1055.pdf

Andrew Kyung1 , Steve Nam2 , \Study on Unemployment Rate in USA Using Computational and Statistical Methods" 978-1-4673-8457-5115/$31.00 ©2019 IEEE | https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8993016