Forecast Unemployment Insurance Claims in the US Using Economic and Labor Market Indicators

Introduction

This project investigates the application of statistical and machine learning methods to forecast **Initial and Continued Unemployment Insurance (UI) Claims** in the United States. Accurate forecasting of UI claims is critical for informing public policy, enabling timely allocation of resources, and assessing labor market conditions. The objective of our work is to evaluate and compare various modeling approaches that incorporate a wide set of economic, labor market, and financial indicators to improve the predictive accuracy of UI claims.

The target variables for this study are seasonally adjusted Initial Claims and Continued Claims, obtained from the U.S. Bureau of Labor Statistics. These weekly data series were aggregated and aligned to a **monthly frequency** to facilitate integration with monthly economic indicators. The explanatory variables were grouped into four categories:

- **Labor Market Indicators**: unemployment level, number of job losers, and discouraged workers
- **Inflation Indicators**: Consumer Price Index (CPI), Core CPI, Personal Consumption Expenditures (PCE), and the PCE chain-type price index
- **Income Indicators**: average hourly earnings of production and nonsupervisory employees
- **Market Indicators**: University of Michigan Consumer Sentiment Index, Brave-Butters-Kelley (BBK) Real GDP estimate, federal funds rate, and NASDAQ Composite Index (monthly average)

All variables were preprocessed and merged into a unified panel dataset. Missing values were handled, and all variables were aligned temporally to ensure consistency across the modeling framework.

Our modeling process was structured in multiple stages:

1. Baseline Modeling: We first implemented Ordinary Least Squares (OLS) regression models for both Initial and Continued Claims, evaluated using five-fold cross-validation to assess out-of-sample prediction accuracy.

2. Regularization: To mitigate overfitting and enhance robustness, we employed Lasso and Ridge regression, tuning regularization parameters through cross-validation.

3. Dimensionality Reduction: Recognizing potential multicollinearity among predictors, we applied Principal Component Analysis (PCA) and selected components based on predictive performance.
4. Ensemble Learning: We fitted Boosting models, which provided improved accuracy on validation sets, particularly in modeling complex nonlinear interactions.
5. Generalized Linear Models (GLMs): We compared two GLM specifications—a Gaussian model with an identity link and a Gamma model with a log link. These models were evaluated based on

goodness-of-fit criteria such as the Akaike Information Criterion (AIC), as well as residual diagnostics.

6. Time Series Modeling: Given the temporal nature of UI claims, we implemented AutoRegressive Integrated Moving Average (ARIMA) models to capture autocorrelation and enhance short-term forecasting capability.

7. Machine Learning Models: To further improve prediction performance, we trained Random Forest and K-Nearest Neighbors (KNN) models using the same set of predictors.

In summary, this project presents a comprehensive modeling framework that integrates traditional statistical techniques with modern machine learning methods. Through comparative evaluation, we aim to identify models that offer not only strong predictive performance but also interpretability and statistical validity. The subsequent sections of this report present detailed methodology, results, and insights derived from each modeling approach, culminating in a discussion of implications and recommendations for future research and application.

Variables

The dependent variable we are studying is the seasonally adjusted data from the Unemployment Insurance Weekly Claims Data. We analyze Initial Claims (IC): the total number of new applicants for unemployment benefits, and Continued Claims (CC): the average number of people still receiving unemployment benefits, to investigate whether there are differences in influencing factors. The data sources are the U.S. Department of Labor, the Bureau of Labor Statistics, and the Federal Reserve Economic Data (FRED).

Since most other data is on a monthly basis, we convert them to monthly data.

The independent variables we selected cover the following aspects, with data from February 1971 to February 2024, spanning 637 months of U.S. monthly data:

Unemployment

Unemployment Level: Refers to the proportion of people who are actively seeking work and willing to work, but are temporarily unemployed in the labor market.

Job Losers: Since this variable is highly correlated with the differenced sequence (correlation coefficient = 0.9770221 and 0.9796417) and with the Unemployment Level, we consider using one of them. This refers to individuals who lose their jobs due to layoffs, closures, or other reasons.
Both of these variables measure the number of unemployed individuals, and we hypothesize they are positively related to both types of unemployment insurance. However, Job Losers represents passive unemployment, while Unemployment Level measures all unemployed individuals, so the influence might differ.

Discouraged Workers: Refers to individuals who have given up searching for work due to the difficulty in finding employment. Because of the restrictions on unemployment benefit applications, an increase in discouraged workers might reduce the number of people applying for unemployment insurance. Data for

this variable is available only after 1995.
For these variables, we consider both the original series and the differenced series, as the changes might be more related to IC.

Inflation

Consumer Price Index (CPI): A measure of changes in the prices of a basket of goods and services, reflecting changes in the cost of living.

Core CPI: Excludes food and energy prices and provides a better reflection of long-term inflationary pressures.
Since we are interested in the change, we use the percentage difference, denoted as $\triangle r$. Since these two variables are highly correlated (correlation coefficient = 0.9988891), we use one of them in the regression.

Personal Consumption Expenditures (PCE): A measure of total consumer spending on all goods and services.

Personal Consumption Expenditures Chain-type Price Index: A chain-weighted method of calculating PCE that considers changes in consumption patterns over time.
Since these are highly correlated (correlation coefficient = 0.9648003), we use one of them in the regression. We also focus on percentage differences, as this is what matters for the analysis. Even though the original series are highly correlated (correlation coefficient over 0.9), the percentage differences for consumption and inflation are not highly correlated (correlation coefficient below 0.5), so both might enter the regression.

Similarly to CPI, PCE reflects broader consumer behavior, and it might have a different impact than CPI.

Income

Hourly Earnings: Average hourly earnings of production and nonsupervisory employees, total private, seasonally adjusted, with CPI used for inflation adjustment.

Market

Real GDP: Brave-Butters-Kelley Real Gross Domestic Product. A monthly GDP measure adjusted for inflation.

Interest Rate: Federal Funds Effective Rate, with the differenced series considered.

NASDAQ Composite Index: Since we focus on changes, we use the percentage difference in the regression.

Consumer Sentiment: Data was only available quarterly before 1978, so linear interpolation is applied.

After preprocessing, no strong correlations are found among variables, except for the specified ones. All variables are standardized, including IC and CC.

The mean of IC is 1,633,120, with a standard deviation of 965,775.9. The mean of CC is 2,871,793, with a standard deviation of 1,583,808. Around 2020, due to the COVID-19 pandemic, both IC and CC saw unprecedented peaks. The last major peak occurred during the 2008-2009 economic crisis.

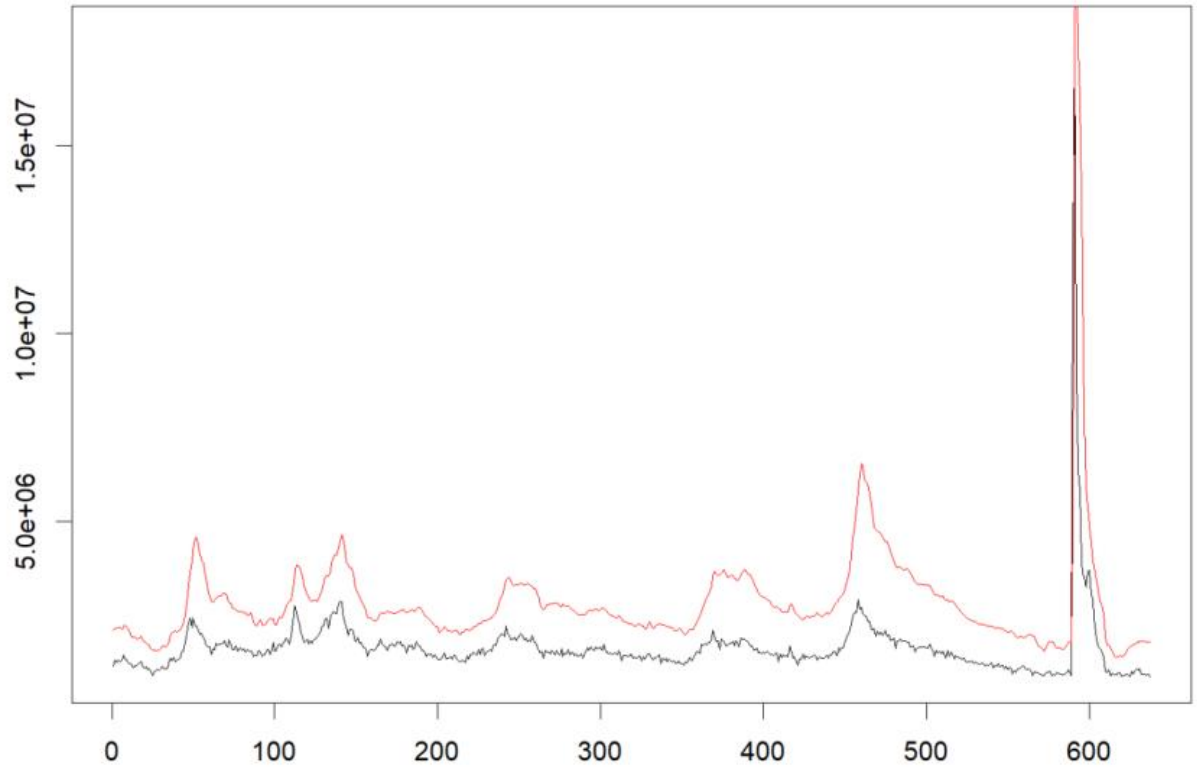


Figure. IC (Black line) and CC (Red line)

1. Multinomial Regression

First, using data after 1995, we include Discouraged Workers along with all other variables for regression. The results show that this variable has a significant negative impact on both IC and CC. This confirms our hypothesis that, due to restrictions on unemployment benefit applications, an increase in Discouraged Workers may lead to a decrease in those receiving unemployment benefits. The negative impact on CC is stronger, as long-term unemployed individuals may have to exit the unemployment insurance program. However, △Discouraged Workers is positively correlated with UI. This may be because △Discouraged Workers and both IC and CC are similarly influenced by the employment environment. The high Adjusted R-squared value reflects the model's strong explanatory power.

IC Regression Model:
IC = -0.14921 + 1.40301*Unemployment Level + -0.95516*Discouraged Workers + 0.02742*△rCPI + 0.43012*Adjusted Hourly Earnings + 0.27911*Consumer Sentiment + -0.29745*Adjusted real GDP + 0.41402*Interest Rate + -0.06197*△rNASDAQ Index + 0.13135*△Unemployment Level + 0.13959*△Discouraged Workers + -0.45704*△Interest Rate

Table. IC initial model coefficients

	coef	P-value
Intercept	-0.14921302	3.22E-02
Unemployment Level	1.40300736	3.80E-68
Discouraged Workers	-0.95516392	1.06E-31
△rCPI	0.02741772	4.65E-01
Adjusted Hourly Earnings	0.43012102	3.22E-16
Consumer Sentiment	0.27910859	3.12E-10
Adjusted real GDP	-0.29745441	2.09E-09
Interest Rate	0.41402274	1.11E-04
△rNASDAQ Index	-0.06197413	1.55E-01
△Unemployment Level	0.13135357	1.74E-03
△Discouraged Workers	0.13959008	1.06E-04
△Interest Rate	-0.45704389	6.50E-05

CC Regression Model:
CC = -0.16485 + 1.89264*Unemployment Level + -1.30340*Discouraged Workers + -0.01527*△rCPI + 0.40625*Adjusted Hourly Earnings + 0.21309*Consumer Sentiment + 0.08093*Adjusted real GDP + 0.33853*Interest Rate + -0.03063*△rNASDAQ Index + -0.10104*△Unemployment Level + 0.17687*△Discouraged Workers + -0.19773*△Interest Rate

Table. CC initial model coefficients

	coef	P-value
(Intercept)	-0.1648521	1.06E-03
Unemployment Level	1.89264268	3.45E-135
Discouraged Workers	-1.30339675	1.22E-77
△rCPI	-0.01527346	5.72E-01
Adjusted Hourly	0.40625148	2.98E-25

Earnings		
Consumer Sentiment	0.2130917	2.85E-11
Adjusted real GDP	0.08092712	2.05E-02
Interest Rate	0.33852698	1.19E-05
△rNASDAQ Index	-0.03062759	3.29E-01
△Unemployment Level	-0.10103809	8.26E-04
△Discouraged Workers	0.17687152	2.45E-11
△Interest Rate	-0.19773331	1.55E-02

Next, using the full dataset, we perform regression on the remaining variables and compare the effects of Unemployment Level vs. Job Losers, CPI vs. Core CPI, and PCE vs. PCEPI. The results show that for both IC and CC, the model using Job Losers has a stronger explanatory power, especially for CC, where there is a significant improvement. This may be because passive unemployed individuals are more likely to claim unemployment insurance for a longer period.

Table. Choose between correlated variables

adjusted R^2	IC	CC
Unemployment Level	0.5762	0.5625
Job Losers	0.6745	0.7203
△rCPI	0.6745	0.7203
△rCore CPI	0.6744	0.7205
△rPCE	0.6735	0.7255
△rPCEPI	0.6745	0.7202

Inflation has little effect on IC, with all four inflation variables being insignificant. However, for CC, PCE is significant. While changes in inflation have little effect on new claims for unemployment insurance, they may have a positive impact on CC. Inflation can increase the demand for CC.

Using Job Losers and the appropriate inflation indicator, we fit the models and use CPI for IC and PCE for CC.

IC Regression with Job Losers and CPI:
IC = 0.7105*Job Losers + 0.03567*△rCPI + 0.1732*Adjusted Hourly Earnings + 0.1826*Consumer Sentiment + -0.1944*Adjusted real GDP + 0.2745*Interest Rate + -0.00243*△rNASDAQ Index +

0.1787*△Unemployment Level + 0.1207*△Job Losers + -0.06797*△Interest Rate
Adjusted R-squared: 0.6745

Table. IC model coefficients

	coef	P-value
(Intercept)	0.00000	1.00000
Job Losers	0.71054	0.00000
△rCPI	0.03567	0.18293
Adjusted Hourly Earnings	0.17319	0.00000
Consumer Sentiment	0.18255	0.00000
Adjusted real GDP	-0.19442	0.00000
Interest Rate	0.27446	0.00000
△rNASDAQ Index	-0.00243	0.91662
△Unemployment Level	0.17868	0.11810
△Job Losers	0.12070	0.20941
△Interest Rate	-0.06797	0.00393

CC Regression with Job Losers and PCE:
CC = 0.9087*Job Losers + 0.1136*△rPCE + 0.1501*Adjusted Hourly Earnings + 0.1070*Consumer Sentiment + 0.02444*Adjusted real GDP + 0.1768*Interest Rate + 0.01050*△rNASDAQ Index + 0.2435*△Unemployment Level + -0.1700*△Job Losers + -0.04363*△Interest Rate
Adjusted R-squared: 0.7255

Table. CC model coefficients

	coef	P-value
(Intercept)	0.00000	1.00000
Job Losers	0.90869	0.00000
△rPCE	0.11355	0.00019
Adjusted Hourly Earnings	0.15014	0.00000
Consumer Sentiment	0.10701	0.00004

Adjusted real GDP	0.02444	0.42429
Interest Rate	0.17682	0.00000
△rNASDAQ Index	0.01050	0.62456
△Unemployment Level	0.24348	0.02082
△Job Losers	-0.17003	0.11231
△Interest Rate	-0.04363	0.04331

Removing the insignificant variables, we perform regression with the remaining variables:

IC Simplified Regression:
IC = 0.7688*Job Losers + 0.2009*Adjusted Hourly Earnings + 0.2206*Consumer Sentiment + -0.3889*Adjusted real GDP + 0.3278*Interest Rate + -0.06194*△Interest Rate
Adjusted R-squared: 0.6286

CC Simplified Regression:
CC = 0.9146*Job Losers + 0.1265*△rPCE + 0.1500*Adjusted Hourly Earnings + 0.1086*Consumer Sentiment + 0.1786*Interest Rate + 0.05966*△Job Losers + -0.04467*△Interest Rate
Adjusted R-squared: 0.7242

Table. CC Simplified model coefficients

	coef	P-value
(Intercept)	0.00000	1.00000
Job Losers	0.91459	0.00000
△rPCE	0.12645	0.00001
Adjusted Hourly Earnings	0.15002	0.00000
Consumer Sentiment	0.10863	0.00002
Interest Rate	0.17864	0.00000
△Job Losers	0.05966	0.03866
△Interest Rate	-0.04467	0.03634

At this point, all variables are significant at the 0.05 level. However, the Adjusted R-squared decreases, and IC does not pass the simplified model's ANOVA test, with an F-value less than 0.001. Therefore, we consider retaining some variables with moderate significance. For IC, after retaining △Job Losers, it passes the test (F-value = 0.2292).

Final IC Regression:
IC = 0.7085*Job Losers + 0.1783*Adjusted Hourly Earnings + 0.1721*Consumer Sentiment + -0.1947*Adjusted real GDP + 0.2910*Interest Rate + -0.06806*△Interest Rate + 0.2910*△Job Losers
Adjusted R-squared: 0.6738

Table. CC Simplified model coefficients

	coef	P-value
(Intercept)	0.00000	1.00000
Job Losers	0.70851	0.00000
Adjusted Hourly Earnings	0.17829	0.00000
Consumer Sentiment	0.17209	0.00000
Adjusted real GDP	-0.19472	0.00000
Interest Rate	0.29101	0.00000
△Interest Rate	-0.06806	0.00372
△Job Losers	0.29102	0.00000

It can be seen that the variables with a significant impact on both IC and CC are similar. In addition to PCE mentioned earlier, there is a difference in Adjusted real GDP, which has a significant negative impact on IC but does not significantly affect CC. This might be because initial claims for unemployment reflect new layoffs, and GDP growth leads to fewer new layoffs. However, continued claims for unemployment reflect persistent unemployment, and early-stage GDP growth may still leave many unemployed individuals who have not yet reentered the job market, weakening the relationship between GDP and CC.

As expected, Job Losers and △Job Losers both have a positive impact. The more unemployed people there are, and the more they increase, the more individuals apply for unemployment benefits. △Job Losers has a greater effect on IC, while Job Losers has a greater effect on CC. Hourly earnings are positively related to both IC and CC, which may be due to the following reasons: economic lag effects; in economics, wages are considered difficult to cut, and many companies, when facing deteriorating operations, prefer to lay off workers instead of directly cutting wages; wage increases are often concentrated in specific industries or high-income groups, while lower-income jobs deteriorate. Consumer Sentiment also has a positive correlation, likely due to economic lag effects, and survey results are skewed by higher-income groups. Interest rates are positively correlated, but △Interest Rate is negatively correlated. This could be because monetary policy has a lagging effect on the economy. High interest rates may slow down the economy and increase unemployment, but when the Federal Reserve begins raising rates, unemployment may not be high. When the Fed begins cutting rates, it may be during an economic slowdown with higher unemployment.

An Adjusted R-squared of around 0.7 suggests that the model has good explanatory power.

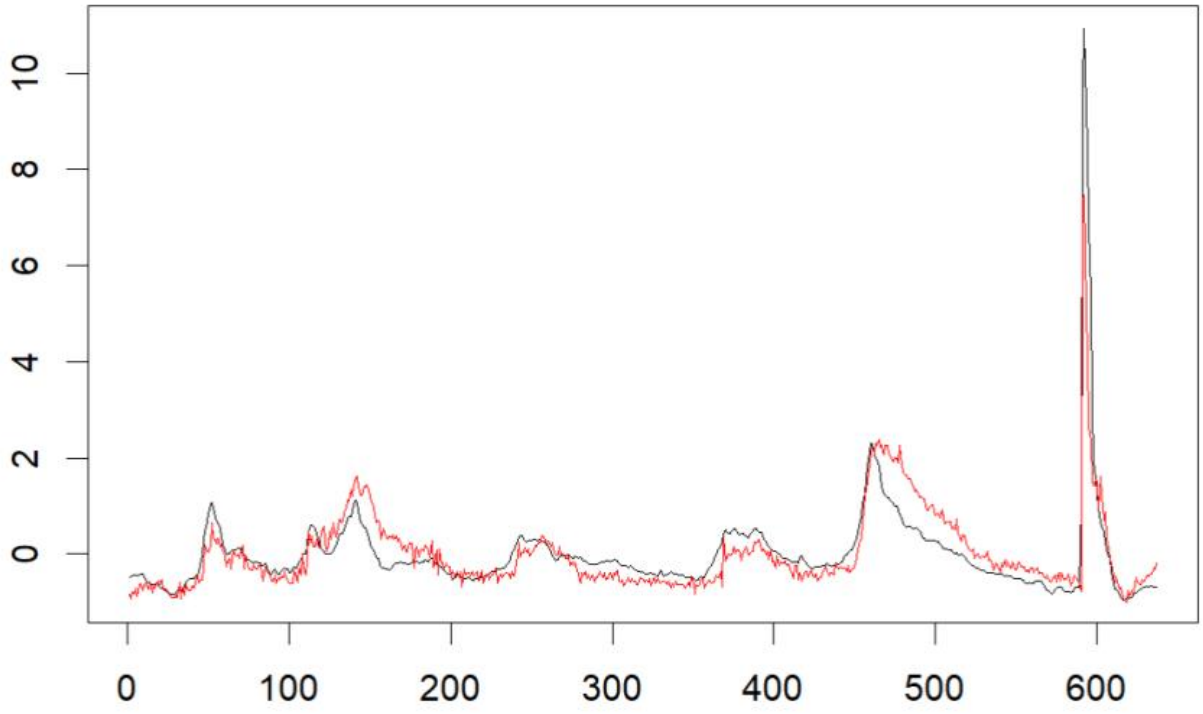


Figure. Prediction on whole dataset

Using five-fold cross-validation, we estimate the prediction errors for the two models. The prediction error for the IC model is estimated at 0.5596423, and the prediction error for the CC model is 0.7507784.

2. Regularization

We attempt to improve the model using regularization. Cross-validation is performed using a grid for alpha and lambda, including all variables except for those with strong correlations.

IC Model:

The optimal hyperparameter is alpha = 0.9, meaning that the penalty consists of 90% L1 regularization (LASSO) and 10% L2 regularization (Ridge). The optimal lambda for IC is 0.0869749, and the estimated prediction error is 0.5274346. This significantly improves the model. At this point, the model parameters are as follows:

IC = 0.50431*Job Losers + -0.10936*Adjusted real GDP + 0.05814*Interest Rate + 0.29863*△Job Losers

Adjusted Hourly Earnings, Consumer Sentiment, and △Interest Rate exit the regression, as they are likely variables influenced by the same factors, rather than directly causing changes in IC.

CC Model:
The optimal hyperparameter is alpha = 0.9. The optimal lambda for CC is 0.1747528, and the estimated prediction error is 0.3288756. This significantly improves the model. The model parameters are as follows:
CC = 0.65771 * Job Losers
All variables except for Job Losers are excluded, suggesting that Job Losers may be the fundamental variable influencing CC.

3. Principal Component Analysis

Next, we attempt PCA using all variables.
After plotting the cumulative variance explained ratio, we observe that the curve flattens, indicating that PCA may not be suitable. Checking the principal components with eigenvalues greater than 1, we determine that the first five principal components should be retained, explaining about 78% of the variance. When regressing using the first five components, the fifth component is not significant for both models. Thus, we only use the first four principal components for regression. The resulting Adjusted R-squared values are 0.5345 for IC and 0.5661 for CC, with estimated prediction errors of 0.6788507 and 0.876993, respectively. These errors are higher than those of OLS. Even when adjusting the number of retained components, the prediction errors are never lower than those from OLS, so we discard PCA.

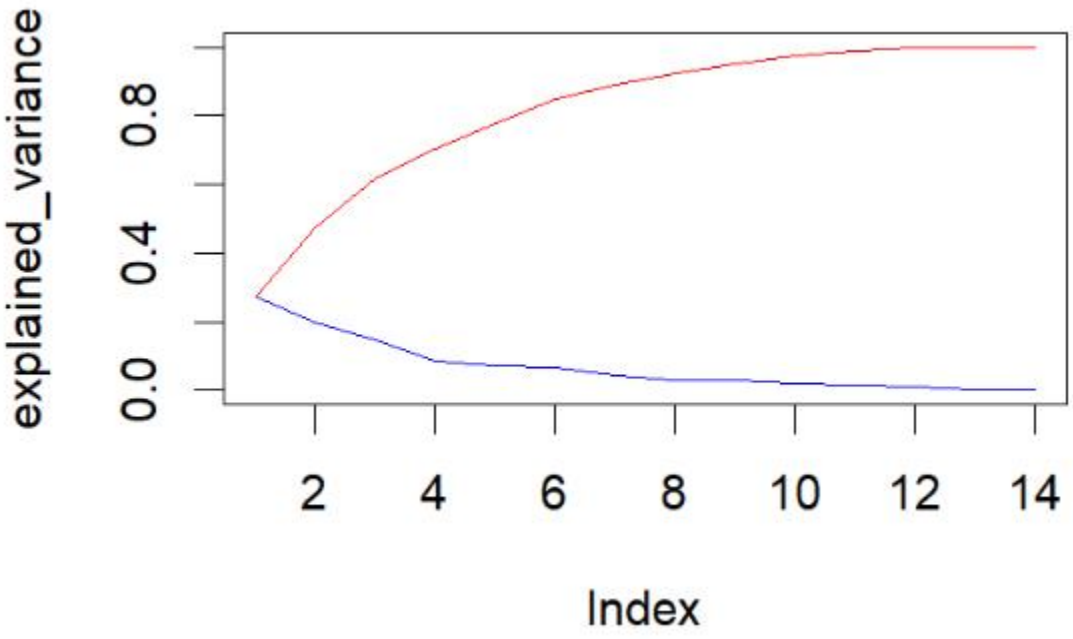


Figure. Scree plot and cumulative explained variance for PCA

4. Boosting

Then, we attempt boosting using all variables.
For IC, the best number of iterations is 4859, and the cross-validation estimated prediction error is

0.4626691. This is better than OLS and regularizaed regression. The variable importance is as follows, which differs significantly from linear regression. Core CPI, Hourly Earnings, △Job Losers, Job Losers, and Adjusted real GDP are the most important variables.

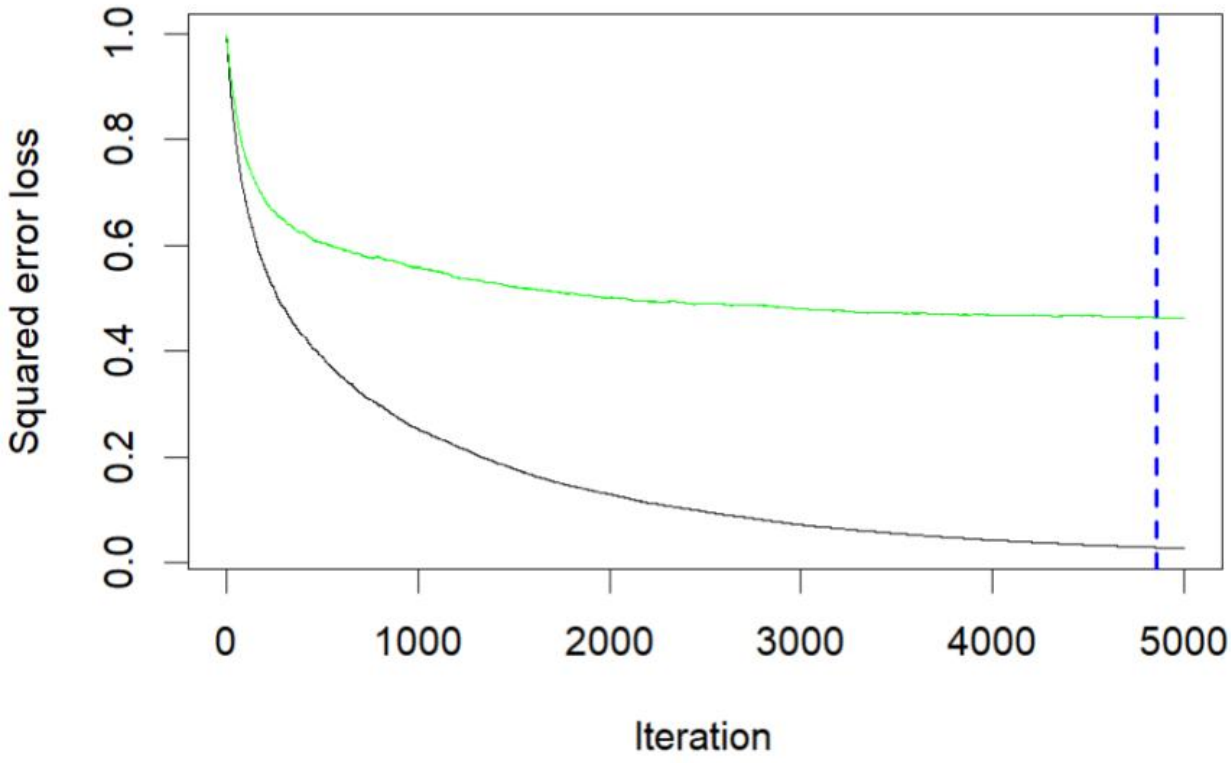


Figure. IC numbers of iterations vs CV MSE

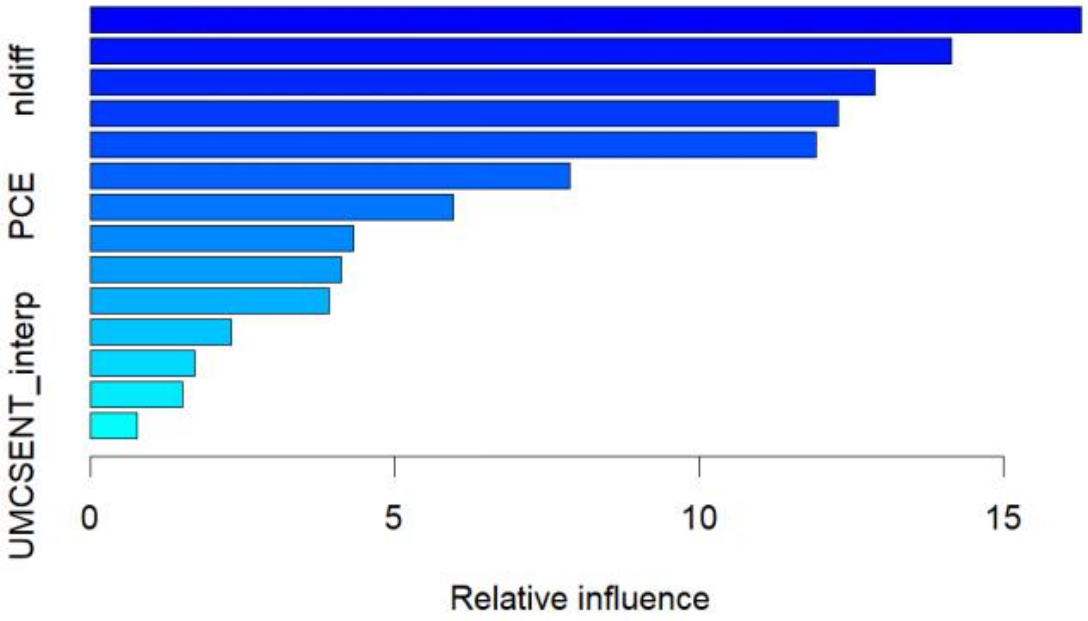


Figure. IC variable importance for boosting

Table. IC variable importance for boosting

variable	importance
△rCPILFESL	16.2731002
Adjusted Hourly Earnings	14.146991
△Job Losers	12.8905347
Job Losers	12.2831192
Adjusted real GDP	11.9168041
△Unemployment Level	7.8793987
△rPCE	5.9589564
△rPCEPI	4.3199819
△Interest Rate	4.1183168
Unemployment Level	3.9245615
△rCPI	2.3071207
Interest Rate	1.7090938
△rNASDAQ Index	1.5081705
Consumer Sentiment	0.7638506

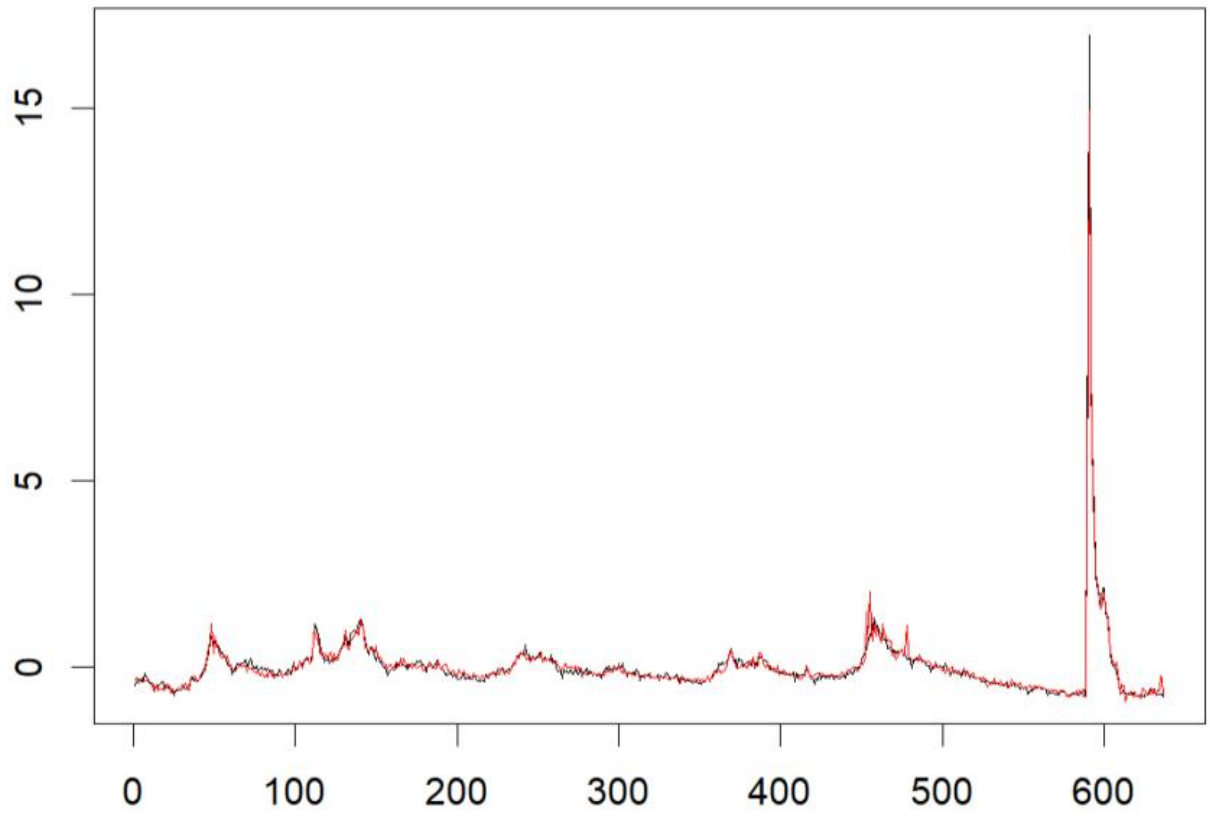


Figure. IC (black line) and IC boosting prediction (red line) on whole data set

For CC, the best number of iterations is 5,000. Although this is the maximum number of trees, further increasing it does not result in a change beyond the standard deviation, and there is a risk of overfitting. The cross-validation estimated prediction error is 0.125189, which performs better than both OLS and regularized regression. The variable importance is as follows, again indicating that Job Losers is the most important variable for CC, followed by Hourly Earnings.

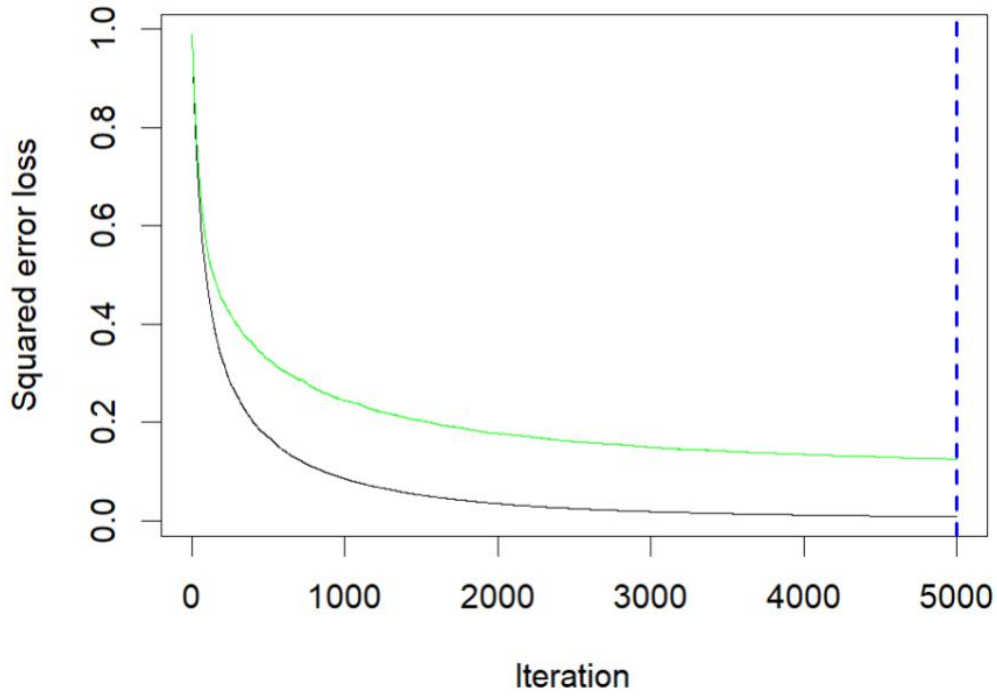


Figure. CC numbers of iterations vs CV MSE

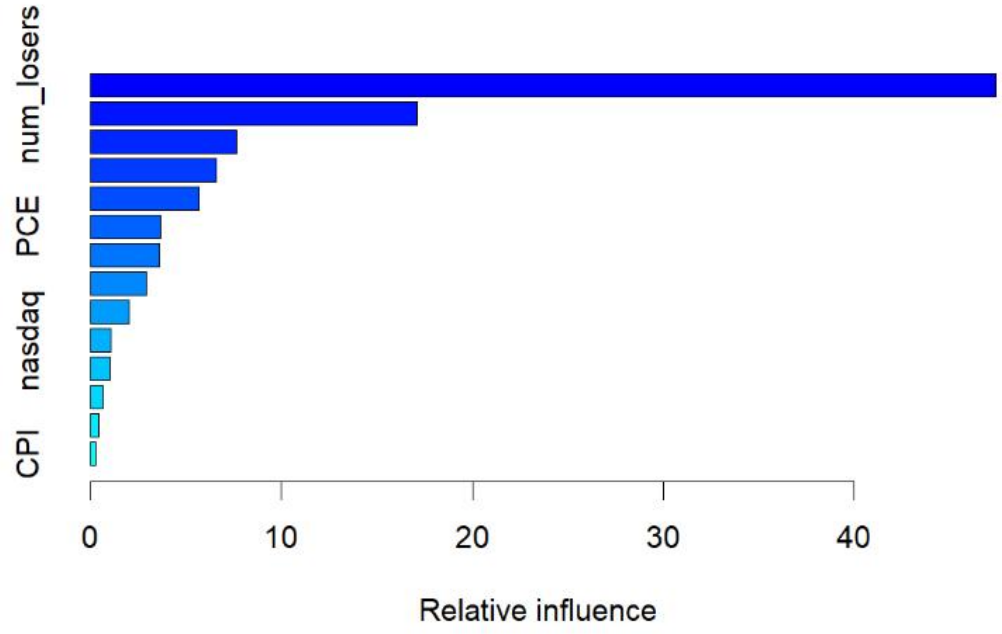


Figure. CC variable importance for boosting

Table. CC variable importance for boosting

variable	importance

Job Losers	**47.423717**
Adjusted Hourly Earnings	**17.098836**
Unemployment Level	**7.623399**
△Job Losers	**6.5741915**
△Unemployment Level	**5.6819568**
△rPCE	**3.6632971**
Adjusted real GDP	**3.6097698**
△rCPILFESL	**2.9410209**
Interest Rate	**2.0047984**
△rNASDAQ Index	**1.0691382**
Consumer Sentiment	**0.9921032**
△rPCEPI	**0.6129388**
△Interest Rate	**0.4172188**
△rCPI	**0.2876153**

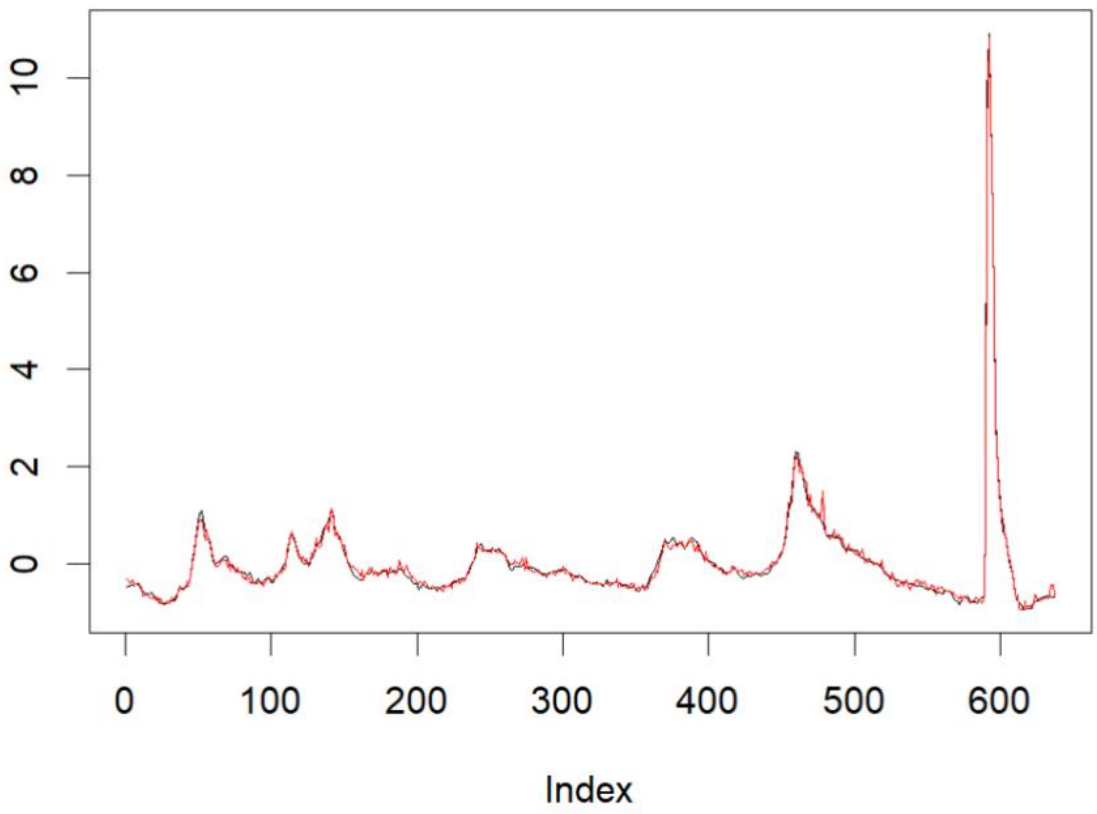


Figure. CC (black line) and CC boosting prediction (red line) on whole data set

5. GLM Model

We developed two types of generalized linear models to forecast unemployment insurance claims: a Gaussian model with an identity link and a Gamma model with a log link. Given the large magnitude and variability of the original data, both Initial Claims (IC) and Continued Claims (CC) were standardized using z-score normalization. For the Gamma-log models, we applied a positive shift to ensure all values remained strictly positive, as required by the Gamma distribution.

Key Findings

The Gamma-log models produced substantially lower AIC values for both IC and CC compared to the Gaussian counterparts, indicating superior goodness-of-fit while penalizing model complexity.

The Gaussian-identity models achieved lower in-sample mean squared errors (MSE) for both IC (0.225) and CC (0.120), compared to the Gamma-log models (IC: 4639.0, CC: 73.7). This is expected since Gaussian models minimize squared error directly.

These results are consistent with the properties of the Gamma family, which naturally models increasing variance with higher mean levels, maintaining the validity of model assumptions.

Model Diagnostics

Gaussian Models (Identity Link):

For Initial Claims, the predicted vs. actual plot (on the scaled scale) shows that the model effectively captures moderate levels of claims but underestimates high spikes. The residuals vs. predicted plot exhibits a funnel-shaped pattern, indicating heteroscedasticity—the variance of errors increases with prediction size.

Similar patterns are observed in the Continued Claims model, with increasing residual spread and underestimation at higher claim levels.

Gamma Models (Log Link):

For Initial Claims, the predicted vs. actual plot demonstrates that the model better tracks extreme values after accounting for the scale shift. Residuals remain relatively homoscedastic, validating the Gamma assumption of variance increasing proportionally with the mean.

The Continued Claims Gamma model also shows improved tracking across a wide prediction range, and residuals appear more stable across fitted values.

These findings are consistent with the theoretical advantages of the Gamma distribution in modeling non-negative, right-skewed data where variance increases with the mean.

Model Selection Considerations

Although the Gaussian model achieves lower MSEs, this performance comes at the cost of violating key assumptions, particularly the assumption of constant variance. The observed residual diagnostics suggest that model errors increase unpredictably with claim size, potentially undermining forecast reliability in volatile periods.

Conversely, the Gamma-log model, while producing higher in-sample MSEs, offers:

- A statistically appropriate variance structure,
- Greater robustness to outliers, and
- Better alignment with the underlying distributional properties of the target variables.

Therefore, despite the MSE tradeoff, we recommend the Gamma-log specification for forecasting unemployment insurance claims, as it better accommodates the data's structural characteristics and is more likely to generalize effectively in practice.

6. KNN

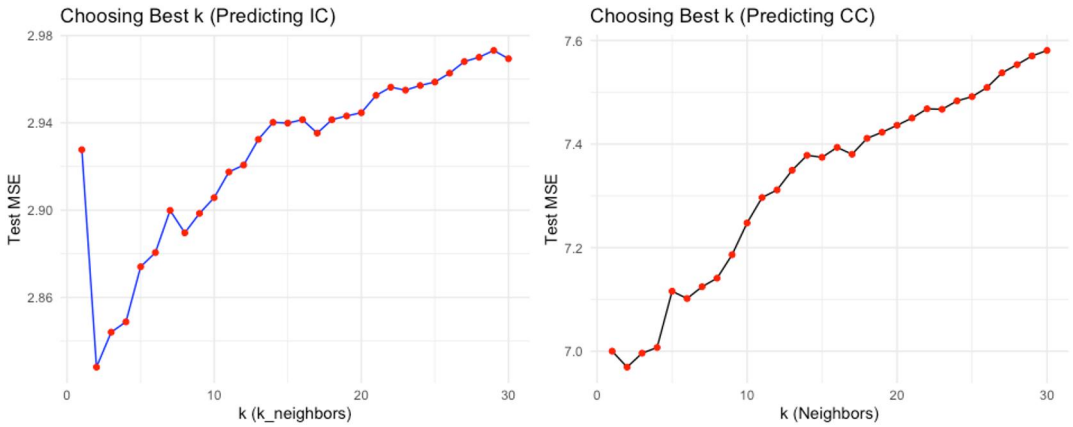
In this project, a K-Nearest Neighbors (KNN) regression model was built to predict two target variables respectively: IC and CC. KNN is an instance-based, non-parametric learning algorithm that makes predictions for a new data point by looking at the most similar observations in the training set. The number of neighbors (k) we choose for the model depends on the bias–variance trade-off. A small k (e.g. 1–3) yields very flexible, low-bias predictions that can overfit noise, while a large k (e.g. 20–30) produces smoother, high-bias forecasts that may miss abrupt spikes in claims. To choose k, we performed a grid search over values from 1 to 30—using Euclidean distance on features first standardized to zero mean and unit variance—and selected the k that minimized test-set mean squared error. In addition, when applied to the time-series forecasting, we maintain chronological order in our train/test split so that each prediction only uses information from earlier dates, emulating a true forecasting scenario. Therefore, the earliest 70 % of observations formed the training set, and the most recent 30 % the test set.

We used Euclidean distance on our 12 standardized predictors including both the stated economic indicators and time features (i.e. year, month, quarter, and time_since_start). We computed time_since_start simply by taking each observation's date, subtracting the very first date in our series, and converting that gap into a continuous count (e.g. number of weeks or months since the start). To raw calendar fields like year, month and quarter, having a monotonic "time since start" variable lets our models learn both long-term trends (e.g. gradual economic shifts) and cyclical patterns (e.g. seasonal hiring freezes or stimulus effects) without having to infer them purely from noisy macro indicators.

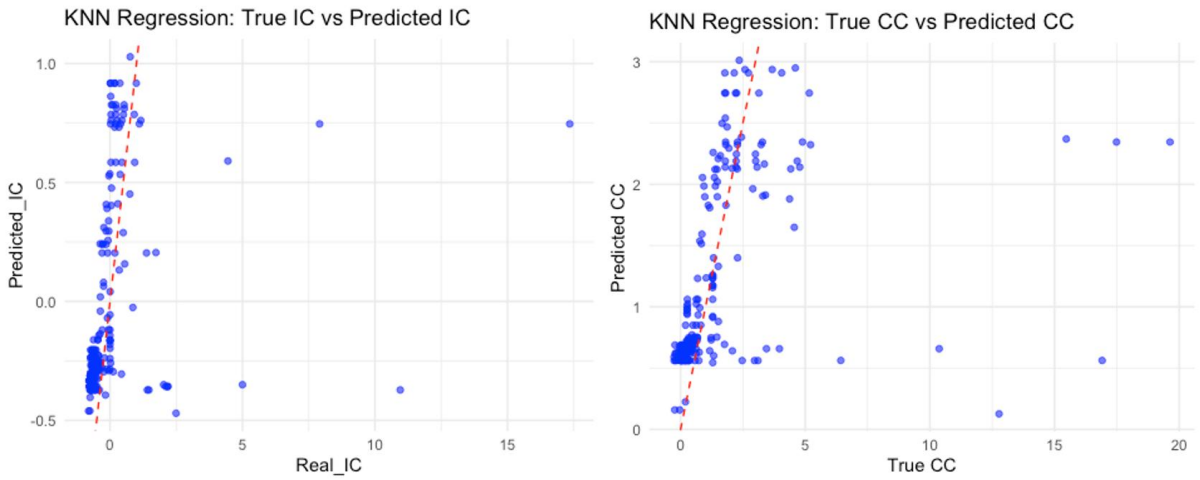
Findings and evaluations:

For both IC and CC targets, the optimal neighborhood size is k = 2, with the test MSE hitting their lowest point at 2.83 and 6.97 respectively, striking the best balance between capturing local claim

dynamics and avoiding excessive smoothing. Meanwhile, larger k neighbors after the best k value lead to an almost consistent increase in test MSE for both models. This indicates that the models are capturing more local patterns in the data. These local dynamics appear to be more informative for predicting unemployment claims than broader trends averaged across many observations. In contrast, increasing k numbers introduces more distant and potentially irrelevant neighbors, which can dilute meaningful signals and reduce prediction accuracy.

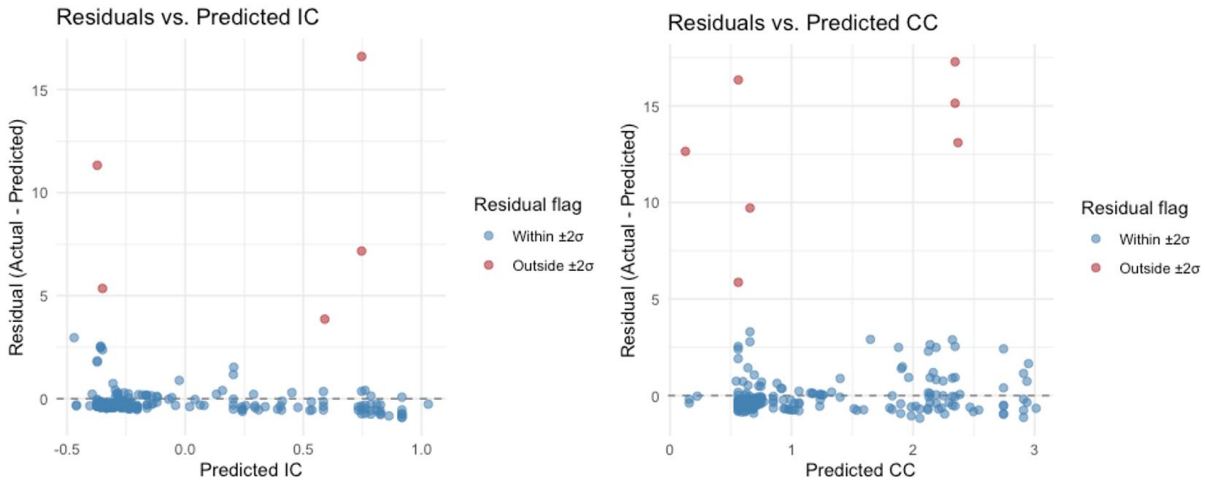


The two scatterplots below reveal important insights into the models' performance on monthly unemployment claim data. Both predictions generally align with the ideal prediction (red dashed) line, indicating reasonable model performance. However, although predicted values tend to cluster tightly near the lower end of the true values, they are often underestimated for higher observations, which are especially visible in the CC plot. This reflects a limitation of KNN when dealing with sparse or extreme cases, as the model relies heavily on local neighborhoods and lacks the flexibility to extrapolate beyond the observed training data.



Utilizing the test sets, the KNN model achieved a mean squared error (MSE) of 2.83 for the IC target and 6.97 for the CC target. When translated into pseudo R-squared , these correspond to values of 0.0716 for IC and 0.1062 for CC. In other words, while the model captures 7.16% of the variance in initial claims, it explains roughly 10.62% of the variance in continuing claims.

The IC residuals vs. predicted plot shows that most residuals are tightly clustered around zero, with only a few outliers falling beyond ±2 standard deviations. This pattern suggests that the model generally performs consistently across different prediction levels. In contrast, the CC plot has slightly more outliers and the other residuals are scattered more loosely around zero. Nonetheless, in both cases, the residuals appear randomly scattered around zero without a clear funnel shape, implying no strong evidence of heteroscedasticity.



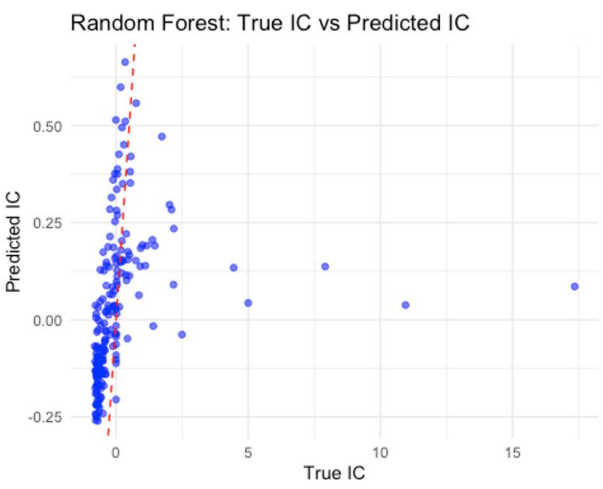
7. Random Forest

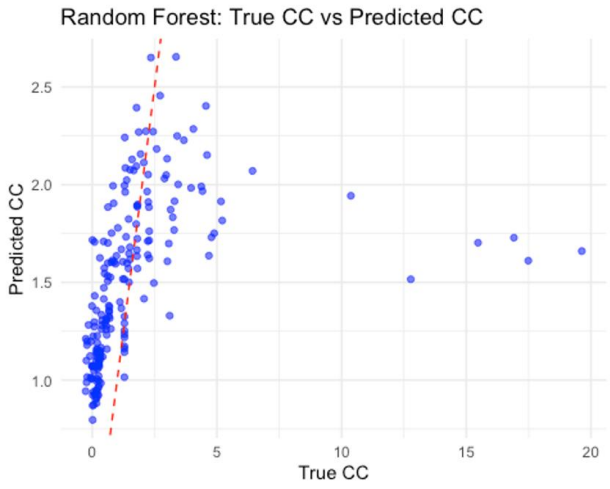
Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions through averaging, leading to more stable and accurate results than any individual tree. Each tree is trained on a bootstrap sample of the training data, and at each split, a random subset of predictors is considered—introducing randomness that reduces overfitting and enhances generalization.

In this project, we applied Random Forest to predict both IC and CC for unemployment insurance. The model was trained using 500 trees (ntree=500) and a feature subset size equal to the square root of the number of predictors. Predictor variables were standardized before training to ensure comparability. The model's ensemble structure helped capture nonlinear relationships and interactions among predictors, offering more robust forecasts than simpler methods like KNN or linear regression.

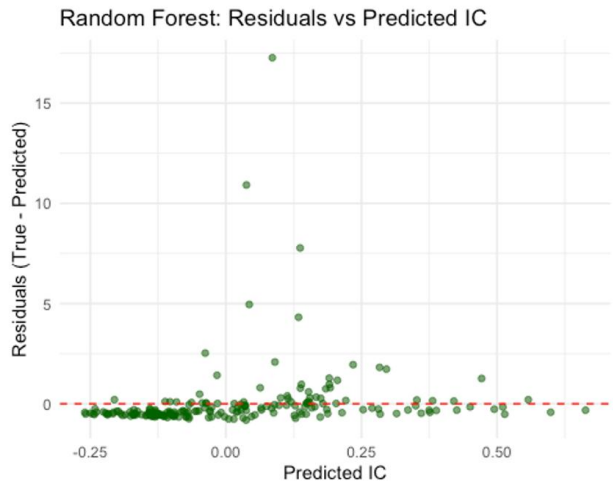
Findings and evaluations:

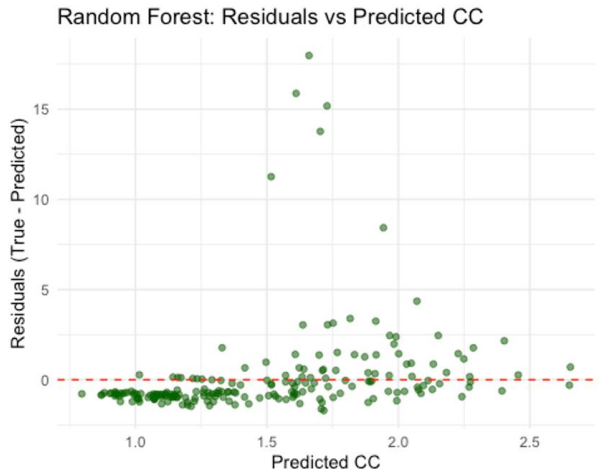
The Random Forest model performs well overall, with predicted values for both CC and IC aligning closely with actual values, especially in the lower range. The red dashed line in both plots indicates perfect prediction; most points cluster around this line, suggesting good model fit. However, again the model underpredicts extreme values as many points are to the far right of the ideal line, especially for values above 5. This systematic underestimation of high values suggests that the Random Forest model is less effective at capturing rare spikes in unemployment claims, possibly due to the limited number of extreme observations in the training data. The overall spread is tighter for IC, suggesting slightly better accuracy, while CC predictions show more dispersion.





For CC, the model achieved a test MSE of 7.09 and a pseudo R^2 of 0.0971, suggesting that it captured around 9.71% of the variation in CC after scaling. The residuals were symmetrically distributed in general, indicating no major signs of bias or heteroscedasticity. Whereas the model performed more effectively on IC, with a higher test MSE of 2.91 but with a lower pseudo R^2 of 0.0454. Visual inspection of residual plots revealed a relatively consistent spread around zero for both targets, though IC residuals showed tighter and better clustering.



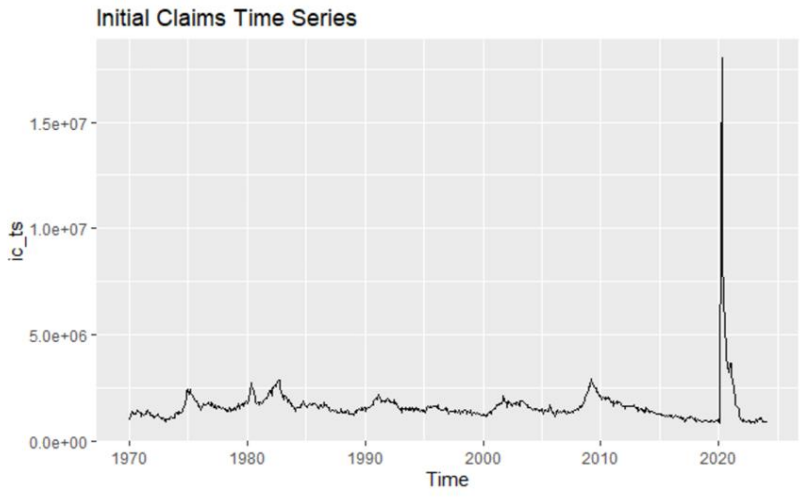


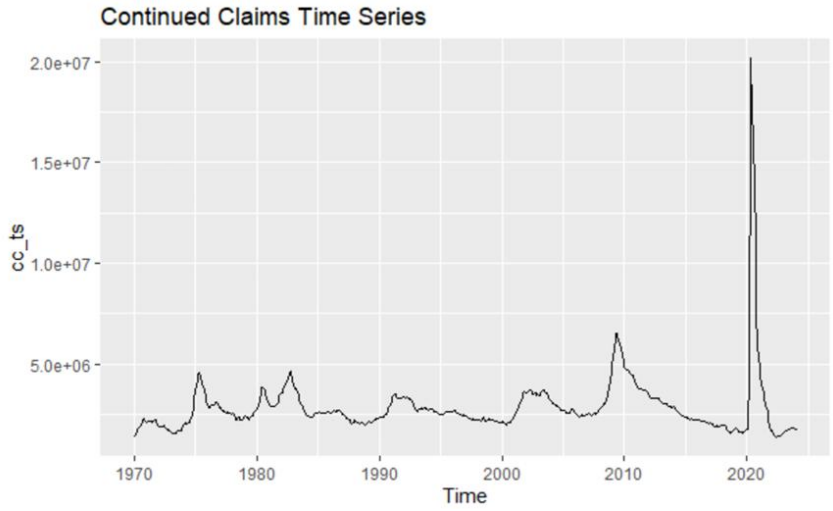
8. Time Series Analysis

Time series is a sequence of data points collected or recorded at regular time intervals—such as daily, monthly, or annually. Each data point represents the value of a variable at a specific point in time. It is especially useful for forecasting the unemployment rate because it helps identify and model patterns over time that are inherent in economic indicators. It can capture temporal trends since the unemployment rate exhibits trends(long-term increases or decreases), seasonally and cyclical behavior. It can make future predictions using historical data without requiring many external assumptions, and does not need to consider some variables of the previous data. .

This part presents a time series analysis and 12-month forecast of Initial Claims (IC) using ARIMA models. The objective is to evaluate future trends and assess the uncertainty of projections using two approaches:

- An Auto ARIMA model selected based on information criteria.
- A manually chosen ARIMA(1,1,1) model, often used for economic data with a clear trend component.



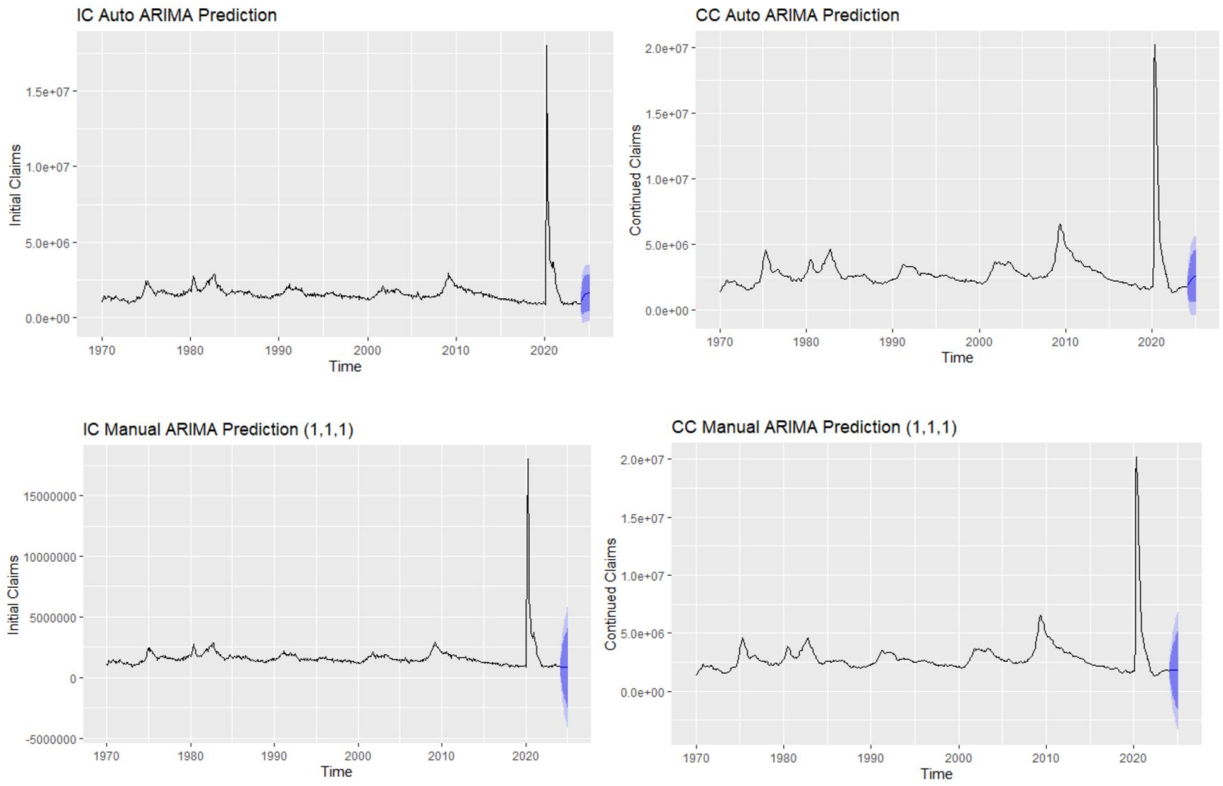


Two ARIMA models were fitted to the IC time series:

- Auto ARIMA (3,0,2): Automatically selected based on the lowest AIC/BIC values. This model assumes the series is stationary and fits three autoregressive terms and two moving average terms.
- Manual ARIMA(1,1,1): A more parsimonious model, using one lag of differencing to address non-stationarity, along with one autoregressive and one moving average term. This model is

commonly used for forecasting in economic and actuarial applications due to its simplicity and reliability.

The models were used to generate 12-step-ahead forecasts with 95% confidence intervals.



Initial Claims	AIC	BIC
Auto ARIMA	18906.03	18937.37
Manual ARIMA	19003.77	19017.19
Continued Claims		
Auto ARIMA	18986.6	19004.51
Manual ARIMA	18997.08	19010.5

Time Series Model Selection

Although the Auto ARIMA performs well, the result is compared with ARIMA(1,1,1). It is still not sure whether the model is the optimal one. Then we use the selected models and forecasted unemployment rates for the next 12 months.

The forecast curves, with confidence intervals, indicate that both Initial Claims and Continued Claims are expected to remain relatively stable in the near future. To further optimize model selection, we conducted a grid search over various (p,d,q) combinations, and found the one that has the lowest AIC/BIC. Based on AIC and BIC criteria, the optimal model of the initial claims based on AIC was identified as ARIMA(4,1,3). The optimal model based on BIC was ARIMA(2,2,4). For the continuous claims, the optimal model based on AIC is ARIMA(4,2,3), and the optimal model based on BIC is ARIMA(3,2,3).

We generally choose AIC and BIC for selecting the best model. Those 2 criteria can provide a quantitative way to balance model fit and complexity, which is crucial when forecasting. ARIMA models can fit the training data very well by increasing the number of parameters (p,d,q) and are also helpful for avoiding overfitting. AIC favors good fit with less aggressive penalty, while BIC has a stronger penalty for complexity and is more conservative.

The range of parameters:

- P: [0,4]
- D: [0,3]
- Q: [0,4]

Initial Claims:

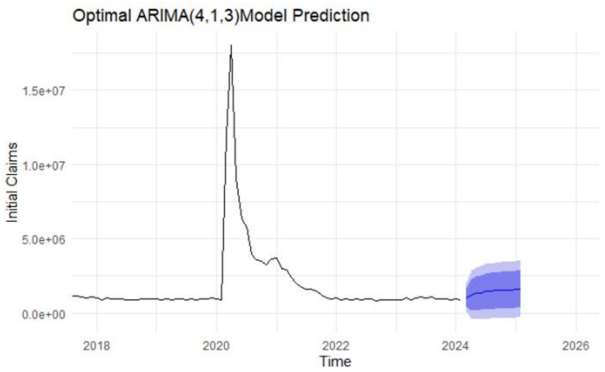
- Optimal Model based on AIC: ARIMA(4,1,3)
- Optimal Model based on BIC: ARIMA(2,2,4)

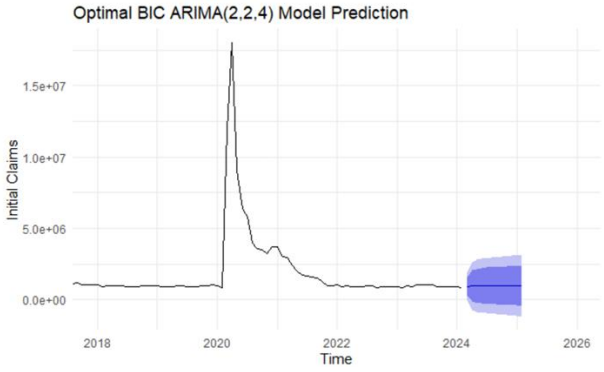
	p <int>	d <int>	q <int>	AIC <dbl>	BIC <dbl>
83	4	1	3	18860.46	18896.27
54	2	2	4	18862.12	18893.44
67	3	1	4	18862.35	18898.15
84	4	1	4	18863.57	18903.85
49	2	1	4	18864.38	18895.70

5 rows

	p <int>	d <int>	q <int>	AIC <dbl>	BIC <dbl>
54	2	2	4	18862.12	18893.44
49	2	1	4	18864.38	18895.70
83	4	1	3	18860.46	18896.27
67	3	1	4	18862.35	18898.15
71	3	2	3	18872.28	18903.59

5 rows





Continued Claims:

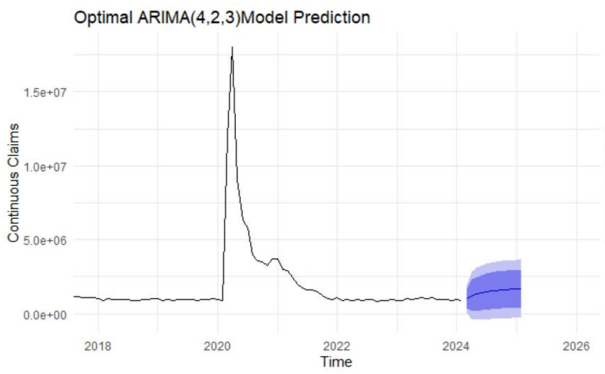
- Optimal Model based on AIC: ARIMA(4,2,3)
- Optimal Model based on BIC: ARIMA(3,2,3)

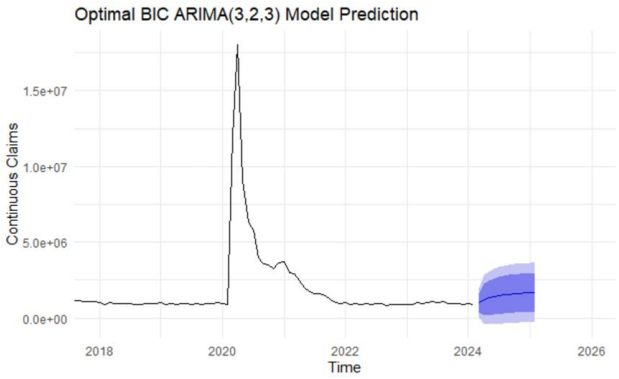
	p <int>	d <int>	q <int>	AIC <dbl>	BIC <dbl>
94	4	2	3	18940.35	18976.15
74	3	2	3	18940.69	18972.00
95	4	2	4	18942.87	18983.14
68	3	1	2	18952.34	18979.19
69	3	1	3	18954.15	18985.48

5 rows

	p <int>	d <int>	q <int>	AIC <dbl>	BIC <dbl>
74	3	2	3	18940.69	18972.00
94	4	2	3	18940.35	18976.15
68	3	1	2	18952.34	18979.19
19	0	3	3	18963.94	18981.83
30	1	1	4	18955.43	18982.28

5 rows





Conclusion and Recommendation

Model	CV MSE (IC)	CV MSE (CC)
OLS Regression	0.5596423	0.7507784
PCA	0.6788507	0.876993
Regularization	0.5274346	0.3288756
Boosting	0.4626691	0.125189
GLM (gamma-log link)	4639	73.7
Time series	N/A	N/A
KNN	2.83	6.97
Random Forest	2.91	7.06

In conclusion, this project explored a range of machine learning and statistical methods to forecast monthly Initial Claims (IC) and Continued Claims (CC) for unemployment insurance in the United States. We implemented and evaluated multiple models, including OLS, Ridge, Lasso, Generalized Linear Models (GLMs), K-Nearest Neighbors (KNN), Random Forest, Gradient Boosting, and ARIMA, using a combination of economic and labor market indicators as predictors. Among the results we

obtained, we can suggest that most models achieved similar test MSE and pseudo R-squared scores for the CC target while there are larger differences in the IC predictions.

Given these findings, we recommend using Gradient Boosting models for deployment, particularly when prediction accuracy is the primary goal. This is because it consistently delivered the best overall performance, achieving the lowest test MSE for both IC and CC. GLMs are recommended in settings where interpretability and formal statistical properties are important. Models should be retrained frequently to incorporate the most recent economic data and to adapt to changing labor market conditions. Moreover, incorporating regional and industry-level variables, or applying hybrid approaches that combine time-series and machine learning methods, could further enhance forecasting performance.

Future Work

Due to the non-normality of the data, we attempted simple skewness correction methods such as logarithmic transformation, but these did not improve the regression performance. Therefore, more advanced normalization techniques, such as the Box-Cox transformation, may be required.

One of our goals is to select the optimal model. Given that tree-based methods like boosting perform well in the presence of those outliers, and considering that policymakers may also be particularly interested in model performance during such exceptional periods, we chose not to remove high-leverage outliers in order to maintain model comparability. To further improve model performance during normal periods, future work may explore removing these outliers when fitting the models.

Finally, to better address the underprediction of rare but critical spikes in claims (which we saw surge during the 2008 financial crisis and the COVID-19 pandemic), future research should explore methods designed for imbalanced or heavy-tailed data. These may include anomaly detection frameworks, extreme value models, or synthetic oversampling techniques. By refining models to better capture volatility, this forecasting system could serve as a valuable tool for timely policy planning and the efficient allocation of unemployment resources.