Time Series Analysis of Employees in Motor Vehicles and Parts

*Abstract*— The prediction of Employees in Motor Vehicles and Parts is presented in this study using various forecasting models such as ARIMA, ETS, TBATS, NN, and PROPHET. This research makes use of R-Studio. Data that has been preprocessed is used to create forecasts. Outlier analysis, data cleansing, and unit root checking are all carried out. The forecast models are fitted after dealing with noisy data, and their performances on both train and test data are compared using numerous comparison criteria.

Keywords—Forecast, employees, nnetar, ets

# INTRODUCTION

The major goal of this research is to figure out how many people work in motor vehicle and parts departments, according to data from the US Bureau of Labor Statistics. A monthly number of employees dataset from 1990 to 2020 is researched and explored for this aim.

In addition to exploring the series, ARIMA, ETS, TBATS, Neural Network, and Prophet models are used to anticipate the number of employees who will be hired in the future. Then, using comparison criteria such as root mean squared error (RMSE), mean absolute error (MAE), and so on, their results were compared (MAE).

# DATA DESCRITPION AND PREPROCESSING

The data collection comes from https://fred.stlouisfed.org/, the website of the Federal Reserve Bank of St. Louis, one of the country's 12 regional reserve banks. The data set contains 360 monthly observations from 1990 to 2020.

Chart

Description automatically generated

***Graph 1****: Time Series Plot of Data Set*

Chart, bar chart, histogram

Description automatically generatedThe plot appears to be non-stationary, as can be seen. The mean term is not constant. Time has an effect on it. Also, some time intervals appear to have a growing trend, but the line is not straight, which makes us skeptical of a stochastic trend. However, we can't draw any conclusions from that plot.

***Graph 2****: ACF Plot of Data Set*

Chart

Description automatically generatedIn the ACF plot, it can be observed that there is a steady linear decay, which supports the judgment made in the first figure. As a result, we are considered to be in the midst of a nonstationary process.

***Graph 3****: PACF Plot of Data Set*

PACF appears to cut off after the first and second lags. There is no need to analyze the PACF plot of the data set because we have already determined that the process is non-stationary by looking at the ACF and time series plots of the data.

The data set is separated into test and train sets at the start of the analysis. The last year's (12 observations) are used as a test set during this process.

Chart

Description automatically generatedThen, using stl decomposition, we check for anomalies in the data and prove that the series has abnormalities.

***Graph 4****: Anomaly Detection Plot*

The tsclean function removes anomalies in the train data and replaces them with interpolated values.

Applied model-based statistical analysis frequently necessitates the fulfillment of some assumptions by the data under consideration. Covariance-stationarity is frequently necessary to begin the modelling process when working with time series. As a result, it seems sense to look for a variance stabilizing transformation that will help the data approach this assumption. Because the interval for lambda value comprises zero when we construct a lambda value for our dataset, the log transformation is the most suitable transformation for the data set to stabilize the variance.

The tests are used to obtain stationary in mean and make the dataset stationary after reaching stationary in variance.

Because this study uses the monthly dataset, the non-stationary is confirmed and assessed using both the KPSS and ADF tests.

The data is not stationary at the first level of the KPSS test. (p<0.05) The second level of this test indicates that the data is harmed by the presence of a unit root. (p<0.05) The ADF test also represents this outcome. (p>0.05)

To solve this problem, consider the first order difference of the series and examine if it is stationary. While ADF indicates that the data does not have a unit root (p<0.05), the KPSS test indicates that, despite the first order difference, the data still has a unit root. (p<0.05).

A picture containing text, antenna

Description automatically generatedBecause the KPSS test is more powerful than the ADF in this case, the result of the KPSS test is taken into account, and the second order difference of the series is computed and proven to be stationary using the KPSS test. (p>0.05).

***Graph 5****: Time Series Plot of Differenced Data Set*

# MODEL SUGGESTION

We'll utilize ACF and PACF plots of the series to suggest a model after we've obtained stationary series.

Chart, box and whisker chart

Description automatically generated

***Graph 6****: ACF Plot of Stationary Data Set*

ACF of the process cuts off after lag 1. By looking this plot, we identify the MA order of process.

Chart

Description automatically generated

***Graph 7****: PACF Plot of Stationary Data Set*

PACF exhibits exponential behavior, which is a feature of the MA process. Furthermore, it may be claimed that it stops after lags 1 and 2, which reveals the AR model's order. As a result of looking at both the ACF and PACF plots, ARIMA(1,2,3)(1,0,1) is the suggested model.

# MODELLING AND DIAGNOSTIC CHECKING

Following the selection of a model that may fit the data, its relevance and performance are tested against a set of criteria. As a result, ARIMA(1,2,3)(1,0,1) is identified as a significant model. There is no other model to compare because it is the only one with significant values. Following the selection of the best model, we will conduct a diagnostic examination of the model.

The goodness of fit of the model and the validity of the assumptions should be evaluated after identifying and estimating the time series model. The ARIMA forecast is produced if we have a perfect fit model. In time series analysis, model verification is comparable to traditional regression analysis and is based on residual analysis.

The visual inspection tool Q-Q plot, as well as the formal tests Shapiro-Wilk and Jarque-Bera, are used to verify normality assumptions.

Chart, line chart

Description automatically generated

***Graph 7****: QQ plot of the standard residuals*

The residuals may not follow a normal distribution because the Q-Q plot reveals S shapes. Shapiro-Wilk and Jarque-Bera tests should be used to confirm non-normality. Fortunately, the findings of both tests are identical. Jarque-Bera proposes that errors have a normal distribution (p>0.05), and Shapiro-Wilk shows that errors have a normal distribution. (p<0.05).

The serial autocorrelation is now examined. We have numerous ways for identifying serial correlation, including the residual plot as visual, the Durbin Watson test (only for the AR(1) process), the Breusch Godfrey test, and the Ljung-Box test modified version of the Box-Pierce test.

Chart

Description automatically generatedThe ACF plot of residuals is the initial method of determining serial correlation. All spikes should be in the White Noise band in order to state we don't have a correlation problem.

***Graph 8****: ACF Plot of the standard residuals*

As can be observed, the White Noise does not contain all of the spikes. As a result, we can deduce that our residuals are perhaps correlated. Following the ACF plot of residuals, we should do the formal tests described above on our best model, which we selected in the previous steps. Both Box-Ljung and Box-Pierce are used to accomplish this, and both indicate that residuals from the best model are uncorrelated. (p>0.05).

The residuals' heteroscedasticity is the last assumption to be tested. We can use the ACF and PACF plots of squared residuals, as well as White's and Bresuch-Pagan tests, to examine this assumption.

Chart

Description automatically generated

***Graph 9****: ACF and PACF of the squared residuals*

The squared residuals are in the 95 percent White Noise Band, as can be observed. As a result, the errors are said to be homoscedastic. However, the formal test should be used to verify this conclusion. According to the Studentized Breusch-Pagan test (p>0.05), the errors are homoscedastic. As a result, we have constant variance across time and do not require the usage of a GARCH model.

We'll use the ets function in R's forecast package to try to determine the best exponential smoothing model after the ARIMA model. Below is the best exponential smoothing model for the series.

***Table 3:*** *Summary of ETS Model*

|  |
| --- |
| ETS(M,Ad,N) |
| Smoothing parameters: |
| alpha = 0.5315 |
| beta = 0.2484 |
| phi = 0.8 |
|  |
| Initial states: |
| l = 960.686 |
| b = 69.0753 |
|  |
| sigma: 15.2923 |
|  |
| AIC AICc BIC |
| 3941.765 3942.011 3964.878 |

As can be seen, the exponential smoothing model has a multiplicative error and an additive trend. The residuals of the ETS model are examined using the Shapiro-Wilk test after fitting the model, and it is determined that they do not follow a normal distribution. (p<0.05).

After exponential smoothing model, TBATS model is fitted to the series. The model details are given below.

***Table 4:*** *Summary of TBATS Model*

|  |
| --- |
| TBATS(1, {3,2}, 1, {-,-}) |
|  |
| Call: tbats(y = train) |
|  |
| Parameters |
| Lambda: 0 |
| Alpha: 0.2925506 |

The residuals of the TBATS model are evaluated using the Shapiro-Wilk test after fitting the model, and it is determined that they do not follow a normal distribution. (p<0.05)

We'll fit a Neural Network model with historical data as input variables. The model's specifications are as follows:

***Table 5:*** *Summary of NNETAR Model*

|  |
| --- |
| model |
| ## Series: data |
| ## Model: NNAR(13,1,7) [12] |
| ## Call: nnetar(y = data) |
| ## Average of 20 networks, each of which is |
| ## a 13-7-1 network with 106 weights |
| ## options were - linear output units |

The model is called NNAR (13,1,7). It's a neural network with one neuron in the hidden layer with the last one observation Y (t-1) as input for forecasting output Y (t). According to the Shapiro-Wilk test, we have nonnormal residuals when looking at NN residuals. (p<0.05)

Finally, we'll apply the prophet model. The model is described in full in the appendix. We can see that the residuals of the models are not normally distributed when we use the Shapiro-Wilk test. (p<0.05).

After training the models, we use the forecast function to get forecast values for each technique and calculate their accuracy. The models' accuracy is listed below.

***Table 6:*** *The train accucary of models*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ME | RMSE | MAE | MAPE | MASE | ACF1 |
| ETS | -0.41 | 15.182 | 8.157 | 0.819 | 0.163 | 0.0109 |
| TBATS | -0.02 | 15.668 | 8.213 | 0.827 | 0.164 | -0.003 |
| NNETAR | **0.004** | **9.87** | **5.459** | **0.535** | **0.108** | **-0.034** |

***Table 7:*** *The forecasting performance of models*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ME | RMSE | MAE | MAPE | MASE | ACF1 |
| ETS | -22.87 | 29.11 | 22.87 | 2.334 | 0.456 | 0.432 |
| TBATS | -26.02 | 32.56 | 26.02 | 2.65 | 0.519 | 0.493 |
| NNETAR | **-18.98** | **25.076** | **18.98** | **1.94** | **0.378** | **0.370** |

In comparison of both the train and test sets, both tables reveal that the Neural Network model outperforms the other approaches on all measures.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

# DISCUSSION AND CONCLUSION

The stationary check of the series was the first stage in this study after data cleaning and splitting; however, after looking at the time series, ACF&PACF plots, and results of the KPSS and ADF tests, it was clear that this condition was not met because the series had a stochastic trend. The use of differencing methods was used to tackle this problem. Using particular methodologies such as ACF&PACF plots, some tentative models were offered once the process became stationary. The best model, which is also the sole significant model, is then chosen.

Diagnostic checks on residuals are performed after the best model has been fitted. Because the series has S shape residuals, the problem of non-normality of errors arises at this point. Using visual inspection tools and formal testing, it is confirmed that the mistakes are uncorrelated and homoscedastic. Different forecasting methods were investigated in addition to the best ARIMA models, and forecasts were generated using them. Finally, when compared to other models, NN has the best performance in both modeling series and predicting future values. Overall, during the analytical process, several problems were encountered, and some of them were solved, while others were not. Despite these issues, the best model was able to be obtained.

|  |
| --- |
| H. Kim, S. W. Park, S. Lee ve X. Xue, «Determinants of House Prices in Seoul: A quantile Regression Approach,» *Pacific Rim Property Research Journal,* Vol. 21, Issue 2, pp. 91-113, 2015. |
| H. Selim, «Determinants of House Prices in Turkey: Hedonic Regression Versus Artificial Neural Network,» *Expert Systems with Applications,* Vol. 36, Issue 2, pp. 2843-2852, 2009. |
| S. C. Bourassa, M. Hoesli ve J. Sun, «A Simple Alternative House Price Index Method,» *Journal of Housing Economics,* Vol. 15, Issue 1, pp. 80-97, 2006. |
| S. Lu, Z. Li, Z. Qin, X. Yang ve R. Goh, «A Hybrid RegressionTechnique for House Prices Prediction,» %1 içinde *IEEE International International Conference on Industrial Engineering and Engineering Management (IEEM)*, Singapore, 2017. |