Downtime Prediction In CNC Machines

Tulasi Anantharamakrishnan
American International School Chennai (AISC)
ramtulasi15@gmail.com

Advisor: Mauricio Hernández (Duke University)

Abstract – **Machine downtime is a key metric which helps the manufacturers in improving the overall equipment effectiveness of the machine. Predicting downtime plays an important role since the amount of time that machine is not operating due to unplanned failure or planned downtime could act as primary decision making for operational excellence. The objective of our study is to build a predictive model using statistical techniques such as multilinear regression and Auto Regressive Integrated Moving Average (ARIMA) using time series analysis to predict downtime. Shift wise production data of a Computer Numeric Control (CNC) machine has been collected from the digital dashboard in the manufacturing plant "IPRings Ltd". Where, shift value is a continuous period of 8 hours split (A, B, and C) considered according to the operational requirement. A multilinear regression model has been built with the time the machine is in operation (runtime) and shift as independent variables to predict the downtime. After analyzing the residuals of the multilinear regression model, an ARIMA model has been built with weekly downtime data and the predictions has been given for next 10 weeks. The advantage of ARIMA model is that it uses the past observations of the target variable to predict the future values. The downtime losses were also classified to identify the major determinants that affect its behavior. With prediction and classification of downtime the manufacturer can take necessary action to reduce downtime.**

I. Introduction

Downtime in manufacturing refers to the period during which production is stopped. It includes both the planned and unplanned maintenance losses. Over the years, the industries have moved on from reactive maintenance to preventive maintenance processes to reduce downtime and improve effectiveness.

By predicting downtime beforehand, we can update the preventive maintenance schedule, plan beforehand for the availability of engineering spare parts and raw materials. Doing so, results in the reduction in downtime which in turn increases the overall equipment effectiveness.

Overall Equipment Effectiveness:

As introduced by Nakajima S (1982) [5], Overall Equipment Effectiveness (OEE) is used as a method to evaluate the effectiveness of the equipment and it is made of three ratios: Availability, Performance, and Quality. Availability describes the percentage of machine downtime. Performance describes the percentage of the maximum operational speed. Quality refers to the percentage of good parts produced. The mathematical expressions [4] are as follows,

OEE = Availability × Performance × Quality (1)
Availability = Run time / Planned production time (2)
Performance = (Cycle time × Total count) / Run time (3)
Quality = Number of units without defect / Number of units produced (4)

II. METHODOLOGY

Multiple Linear Regression:

As narrated in "Business Analytics, the science of data driven decision making" [1], Multiple Linear regression is a statistical technique that establishes the existence of linear relationship (association) between a dependent variable(Y) and independent variables (X_k).. The regression models do not establish causal relationship, however, can be used to check whether there is an association relationship between dependent variable(Y) and independent variables(X_k). We can only establish that change in value of (Y) is caused due to change in values of (X_k).

The functional form of MLR is given by,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i \qquad (5)$$

In the above equation the variable Y is the dependent variable; $X_1, X_2, ..., X_k$ are independent variables; β_0 is a constant; $\beta_1, \beta_2, ..., \beta_k$ are called the partial regression coefficients corresponding to the explanatory variables $X_1, X_2, ..., X_k$ respectively; and ε_i is the error term.

Ordinary least squares method to estimate the regression parameters:

Consider the Multiple Regression model with n independent variables as given in the above equation. Then, the assumptions of multiple regression model are as follows:

1. The regression model is linear in parameter.

2. The explanatory variable X_i is assumed to be non-stochastic (that is, X_i is deterministic).
3. The conditional expected value of the residuals, E $(\varepsilon_i|X_i)$, is zero.
4. In a time series data, residuals are uncorrelated, that is, $\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.
5. The residuals ε_i, follows normal distribution.
6. The variance of the residuals, $Var\left(X_i\right)$, is constant for all values of X_i. When the variance of the residuals is constant for different values of X_i, it is called homoscedasticity. A non-constant variance of residuals is called heteroscedasticity.
7. There is no high correlation between independent variables in the model (called multi-collinearity). Multicollinearity can destabilize the model and can result in incorrect estimation of the regression parameters.

The method of Ordinary Least Squares (OLS) is used to fit a polygon through a set of data points, such that the sum of the squared distance between the actual observations in the sample and the regression equation is minimized. OLS provide the Best Linear Unbiased Estimate (BLUE) that is, $E\left[\beta - \hat{\beta}\right] = 0$, where β is the population parameter and $\hat{\beta}$ is the estimated parameter value from the sample.

Validation of Multiple Regression Model:
The following measures and tests are carried out to validate a multiple linear regression model:

1. Coefficient of multiple determination (R-square) and Adjusted R-Square, which can be used to judge the overall fitness of the model.
2. T-test to check the existence of statistically significant relationship between the response variable and individual explanatory variable at a given significance level (α) or at $(1 - \alpha)100\%$ confidence level
3. F-test to check the statistical significance of the overall model at a given significance level (α) or at $(1 - \alpha)$ 100% confidence level.
4. Conduct a residual analysis to check whether the normality, homoscedasticity assumptions have been satisfied. Also, check for any pattern in the residual plots to check for correct model specification.
5. Check for presence of multicollinearity (strong correlation between independent variables) that can destabilize the regression model.
6. Check for auto-correlation in case of time series data.

Auto-Regressive Integrated Moving Average (ARIMA) Process:
ARIMA model was proposed by Box and Jenkins (1970) and thus is also known as Box-Jenkins methodology. ARIMA model is an integrated model of Auto Regressive(AR) and moving average (MA) models. Auto Regressive (AR), Moving Average (MA), and Auto Regressive Moving Average (ARMA) models can be used only when the data is stationary. ARIMA models can be used even when the data is non-stationary. The presence of stationarity in the data can be checked using the autocorrelation function plot and Dickey fuller test or the augmented Dickey Fuller test. If a time series data, Y_t, is stationary, then it satisfies the following conditions:

1. The mean values of Y_t at different values of t are constant
2. The variances of Y_t at different time periods are constant (Homoscedasticity)
3. The covariances of Y_t and Y_{t-k} for different lags depend only on k and not on time t.

ARIMA model predicts future value using a linear combination of past observations of a specific value to be predicted by composing autocorrelation using time series modeling. That is, early observations affect later observations. Additionally, MA error will affect the predicted value during the prediction phase. Thus, the AR model uses lags of dependent variables as independent variables. However, the MA model uses past errors that follow a white noise distribution as explanatory variables. If y is denoted as the d-th difference, yt' is the differenced series. The model assumes stationarity and is analyzed after preprocessing, such as log transformation and differencing for forecasting and prediction. Thus, the general forecasting model can be expressed as follows,

$$y'_t = \varphi_0 + \varphi_1 y'_{t-1} + \varphi_2 y'_{t-2} + \dots + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta \qquad (6)$$

where y regressed on itself lagged by the nth period; $\varphi_i(i = 1, \dots, p)$ and $\theta_j(j = 1, \dots., q)$ are defined by the weights for the AR and MA parameters; φ_1 and θ_1 are therefore the coefficients of the first AR and MA terms, respectively; and ε_t is a residual term with mean zero and variance σ_ε^2 . In Equation (1), the error term (ε) reflects the previous state at present. This implies that the MA model estimates the rate of change using auto correlated errors.
ARIMA has the following three components and is represented as ARIMA (p, d, q):

1. Auto-regressive component with p lags AR(p).
2. Integration component or number of non-seasonal differences needed for stationarity(d).
3. Moving average with q lags, MA(q).

For example, an ARIMA (0, 0, 0) model is white noise, which means that the errors are uncorrelated across time. An ARIMA (1, 0, 0) model is a first-order AR model, which is a stationary and auto correlated series. It can be predicted as a multiple of its previous value with a constant $(y'_t = \varphi_0 + \varphi_1 y'_{t-1})$. For the

ARIMA (0, 0, 1), the MA model processes means identically as the infinite sum of exponentially weighted past observations of the process.

For model selection, we used the validation data to select the model with the highest accuracy value. During this evaluation process, the ARIMA model parameters were found in various combinations using p, d, and q.

Model Evaluation using Error Metrics:

After model building both the regression model and the ARIMA model was evaluated using the MSE, RMSE, and MAPE:

$$\text{MSE} \ = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \qquad (7)$$

$$\text{RMSE} \ = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \qquad (8)$$

$$\text{MAPE} \ = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \qquad (9)$$

Where n is the sample data points, y_i denotes the actual values, and $\hat{y}_i$ denotes the prediction values.

III. RESULTS

Classification of Downtime:

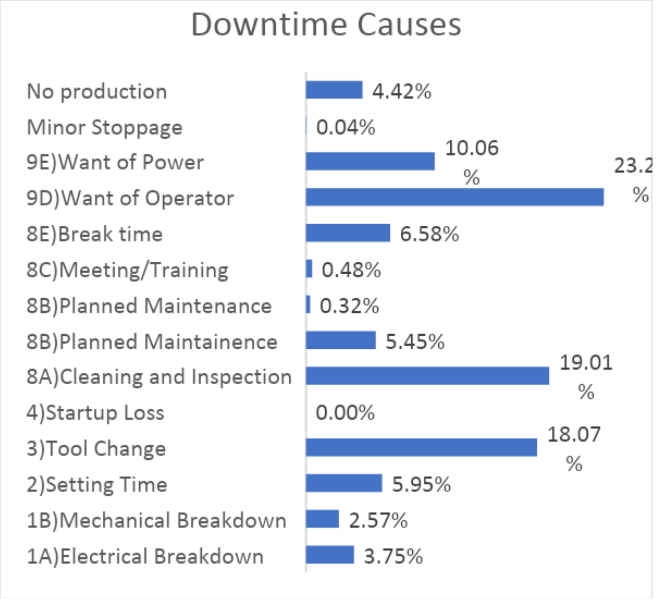


FIGURE 1

PERCENTAGE OF DOWNTIME WITH RESPECT TO DOWNTIME LOSS CATEGORIES

Want of operator occurring due to lack of manpower, cleaning and inspection and tool changes are the main categories out of the 14 loss categories which contributes to 70.42% of the total downtime observed in the data which was collected between the period 01-09-2020 and 30-06-2022.

Outlier Analysis:

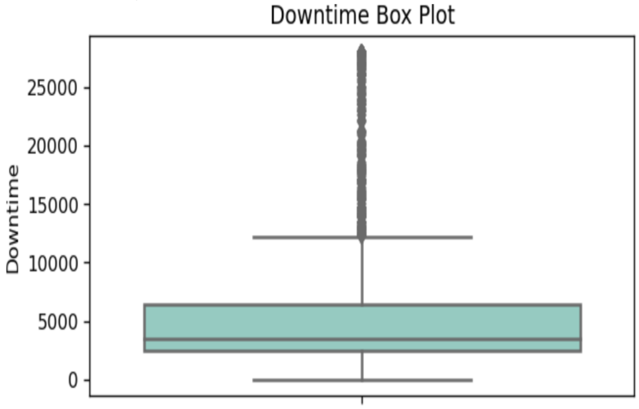
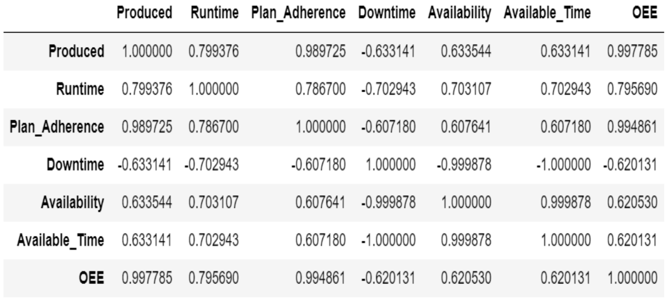


FIGURE 2
BOX PLOT OF DOWNTIME INDICATING THE PRESENCE OF OUTLIERS

Predicting downtime using multiple linear regression:

Before starting with the development of our model, during the exploratory data analysis stage we found outliers as in Fig (2) in our downtime dataset and removed them. In feature selection we shortlisted one categorical column which is the shift (A shift, B shift and C shift) at which the production is happening and one numerical column the time the equipment is in operation (runtime) using correlation analysis and variable inflation factor as in Fig (3 and 4).

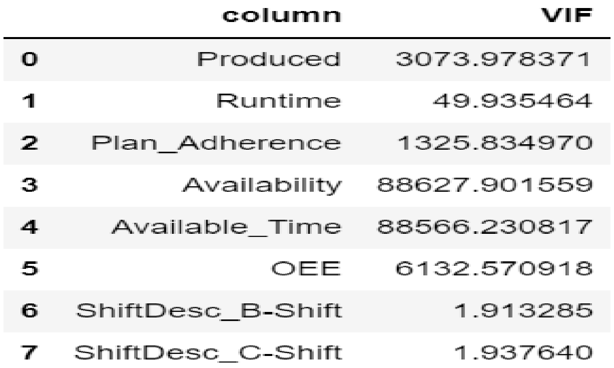
Correlation Analysis:



	Produced	Runtime	Plan_Adherence	Downtime	Availability	Available_Time	OEE
Produced	1.000000	0.799376	0.989725	-0.633141	0.633544	0.633141	0.997785
Runtime	0.799376	1.000000	0.786700	-0.702943	0.703107	0.702943	0.795690
Plan_Adherence	0.989725	0.786700	1.000000	-0.607180	0.607641	0.607180	0.994861
Downtime	-0.633141	-0.702943	-0.607180	1.000000	-0.999878	-1.000000	-0.620131
Availability	0.633544	0.703107	0.607641	-0.999878	1.000000	0.999878	0.620530
Available_Time	0.633141	0.702943	0.607180	-1.000000	0.999878	1.000000	0.620131
OEE	0.997785	0.795690	0.994861	-0.620131	0.620530	0.620131	1.000000

FIGURE 3
CORRELATION OUTPUT INDICATING PRESENCE OF MULTICOLLINEARITY

Handling multicollinearity with variable inflation factor:



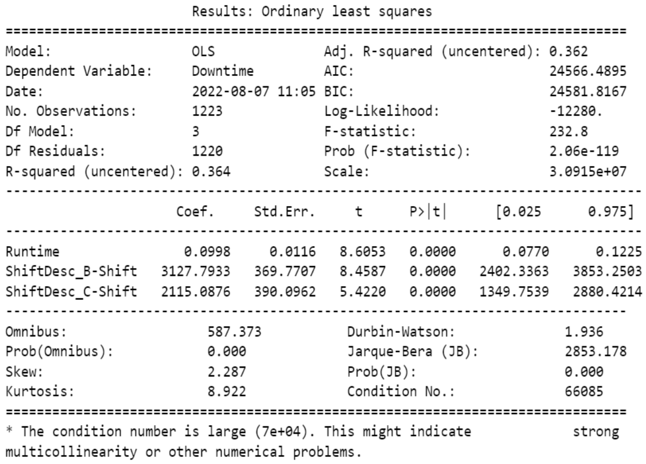
	column	VIF
0	Produced	3073.978371
1	Runtime	49.935464
2	Plan_Adherence	1325.834970
3	Availability	88627.901559
4	Available_Time	88566.230817
5	OEE	6132.570918
6	ShiftDesc_B-Shift	1.913285
7	ShiftDesc_C-Shift	1.937640

FIGURE 4
INDEPENDENT VARIABLES AND THEIR CORRESPONDING VARIABLE INFLATION FACTOR, AS A THUMB RULE VARIABLES HAVING VIF VALUE GREATER THAN 4 ARE REMOVED.

Model Output:

A regression model has been built with features “Runtime” and “ShiftDesc” by splitting the data into train (80 percent) and validation or test set (20 percent).

The linear regression model output is as in Fig (4),



```
                        Results: Ordinary least squares
===============================================================================
Model:                 OLS              Adj. R-squared (uncentered): 0.362
Dependent Variable:    Downtime         AIC:                         24566.4895
Date:                  2022-08-07 11:05 BIC:                         24581.8167
No. Observations:      1223             Log-Likelihood:              -12280.
Df Model:              3                F-statistic:                 232.8
Df Residuals:          1220             Prob (F-statistic):          2.06e-119
R-squared (uncentered): 0.364           Scale:                       3.0915e+07
---------------------------------------------------------------------------------
                      Coef.     Std.Err.     t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Runtime              0.0998      0.0116   8.6053   0.0000      0.0770      0.1225
ShiftDesc_B-Shift 3127.7933    369.7707   8.4587   0.0000   2402.3363   3853.2503
ShiftDesc_C-Shift 2115.0876    390.0962   5.4220   0.0000   1349.7539   2880.4214
--------------------------------------------------------------------------------
Omnibus:                  587.373          Durbin-Watson:             1.936
Prob(Omnibus):            0.000            Jarque-Bera (JB):          2853.178
Skew:                     2.287            Prob(JB):                  0.000
Kurtosis:                 8.922            Condition No.:             66085
===============================================================================
* The condition number is large (7e+04). This might indicate            strong
multicollinearity or other numerical problems.
```

FIGURE 5
LINEAR REGRESSION MODEL SUMMARY

The model R-squared value is 0.364, that is, the model explains 36.4 percent of the variation in Downtime. The p value is 0 which means that there is a statistically significant relationship between features B-shift, C-shift, Runtime and Downtime. But the probability value of f statistic is close to 0 which means that the overall model is statistically significant.

Residual Analysis:

In Fig (5), the diagonal line is the cumulative distribution of a normal distribution, whereas the dots represent the cumulative distribution of the residuals. Since the dots are not close to the diagonal line we can conclude that the residuals do not follow normal distribution, not even approximately.

In Fig (6), there is some sort of pattern in the plot which indicates that the condition of homoscedasticity has failed.

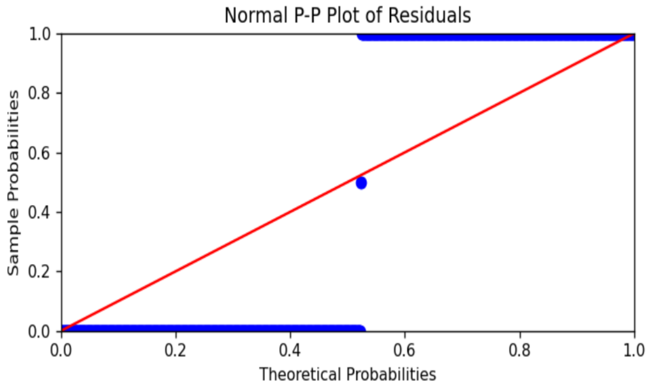


FIGURE 6
NORMAL P-P PLOT OF RESIDUALS TO CHECK WHETHER THE RESIDUALS FOLLOWS NORMAL DISTRIBUTION

Measuring accuracy with error metrics:

Mean square error(MSE) – 32830370, Root mean square error(RMSE) – 5729.77 and Mean absolute percentage error(MAPE) – infinite (Since there are observations in the validation data that have zero downtime for that particular shift)

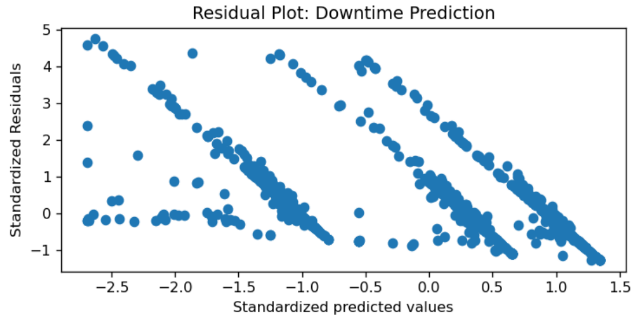
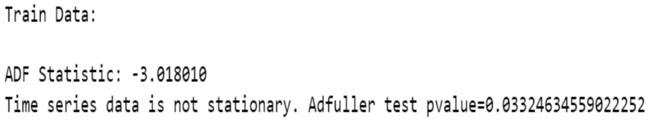


FIGURE 7
STANDARDIZED RESIDUALS PLOTTED AGAINST STANDARDIZED PREDICTED VALUES

Predicting downtime using ARIMA model:

From the downtime data of 96 weeks, we chose the first 89 weeks for our training data and remaining 17 weeks as our test data.

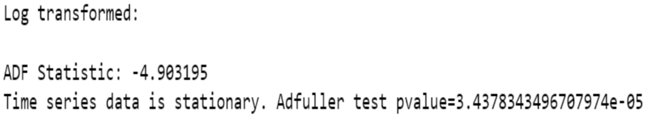
Stationarity Checking:



```
Train Data:

ADF Statistic: -3.018010
Time series data is not stationary. Adfuller test pvalue=0.03324634559022252
```

FIGURE 8
DICKEY FULLER TEST RESULT FOR TRAIN DATA



```
Log transformed:

ADF Statistic: -4.903195
Time series data is stationary. Adfuller test pvalue=3.4378343496707974e-05
```

FIGURE 9
DICKEY FULLER TEST RESULT FOR TEST DATA

From the results of the Augmented Dickey Fuller tests for $\alpha = 0.01$ as in Fig (7), with the p value of 0.032 the downtime was not stationary. So as in Fig (8) the log transformed data was stationary with p value of 0.0000343.

Time series decomposition:

We decomposed the weekly time series data of CNC 43 into several components representing trend, seasonality and residuals as shown in Fig (9). As the trend describes there is a decreasing trend from Nov 2020 to Mar 2021 and there is an increasing trend from May 2021 to Oct 2021 and there is no trend in downtime between Jan 2022 and Jun 2022. As the seasonal chart describes, the seasonality of the time is unaffected by a specific seasonal factor. And the residuals show that the downtime is at its highest of the 3rd week of September 2020 and at its lowest on the 4th week of May 2021.

ARIMA Model Output:

An ARIMA model has been built on the log transformed data with the value (1,1,2) for parameter (p, d, q) which seems to be the model which has the highest accuracy on the validation or test data. As shown in Fig (10) the model output is statistically significant with p values less than 0.05 for all the coefficients.

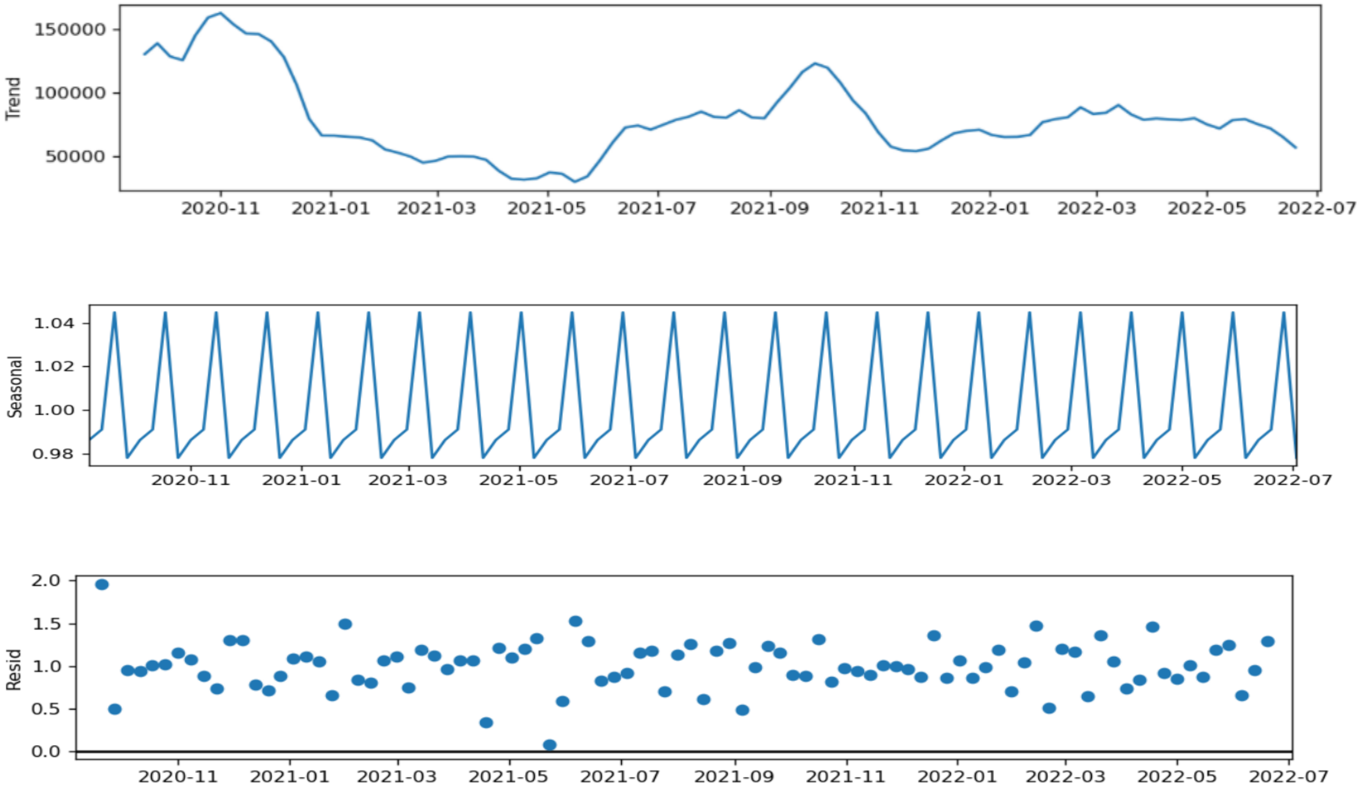
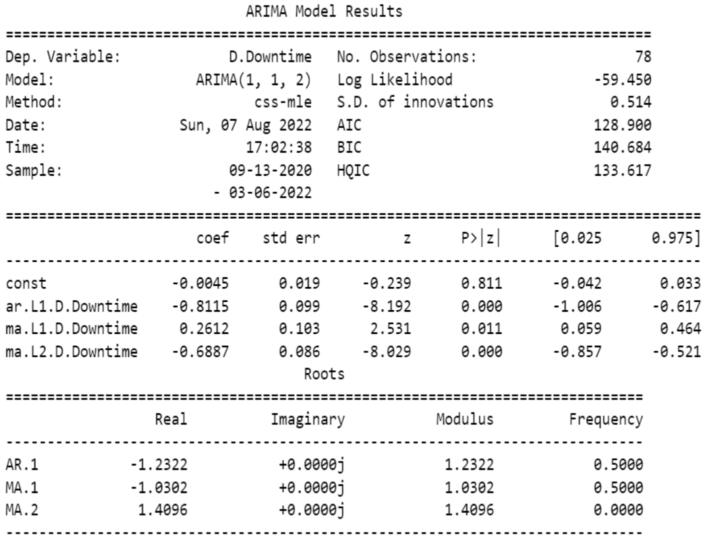


FIGURE 10

TIME SERIES DECOMPOSITION OF CNC 43 WEEKLY DOWNTIME DATASET



```
                              ARIMA Model Results
==============================================================================
Dep. Variable:            D.Downtime   No. Observations:                   78
Model:                 ARIMA(1, 1, 2)   Log Likelihood                -59.450
Method:                      css-mle   S.D. of innovations              0.514
Date:               Sun, 07 Aug 2022   AIC                            128.900
Time:                       17:02:38   BIC                            140.684
Sample:                   09-13-2020   HQIC                           133.617
                        - 03-06-2022
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
const               -0.0045      0.019     -0.239      0.811      -0.042       0.033
ar.L1.D.Downtime    -0.8115      0.099     -8.192      0.000      -1.006      -0.617
ma.L1.D.Downtime     0.2612      0.103      2.531      0.011       0.059       0.464
ma.L2.D.Downtime    -0.6887      0.086     -8.029      0.000      -0.857      -0.521
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -1.2322           +0.0000j            1.2322            0.5000
MA.1           -1.0302           +0.0000j            1.0302            0.5000
MA.2            1.4096           +0.0000j            1.4096            0.0000
-----------------------------------------------------------------------------
```

FIGURE 11

ARIMA MODEL OUTPUT SUMMARY

Residual Analysis:
ARIMA model is a regression model and thus has to satisfy all the assumptions of regression. The residuals should be white noise and not correlated. This can be observed by using ACF and PACF plots of the residuals as shown in Fig (11) and Fig (12).

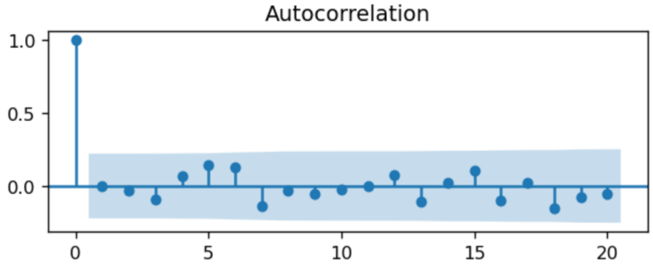


FIGURE 12
ACF PLOT OF THE RESIDUALS

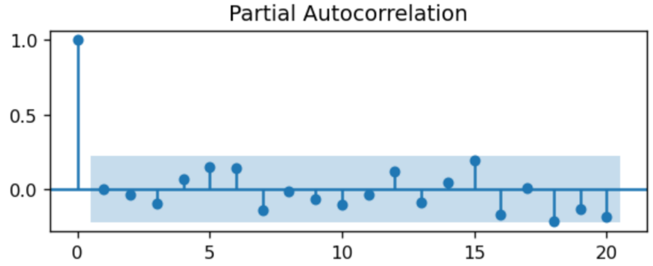


FIGURE 13
PACF PLOT OF THE RESIDUALS

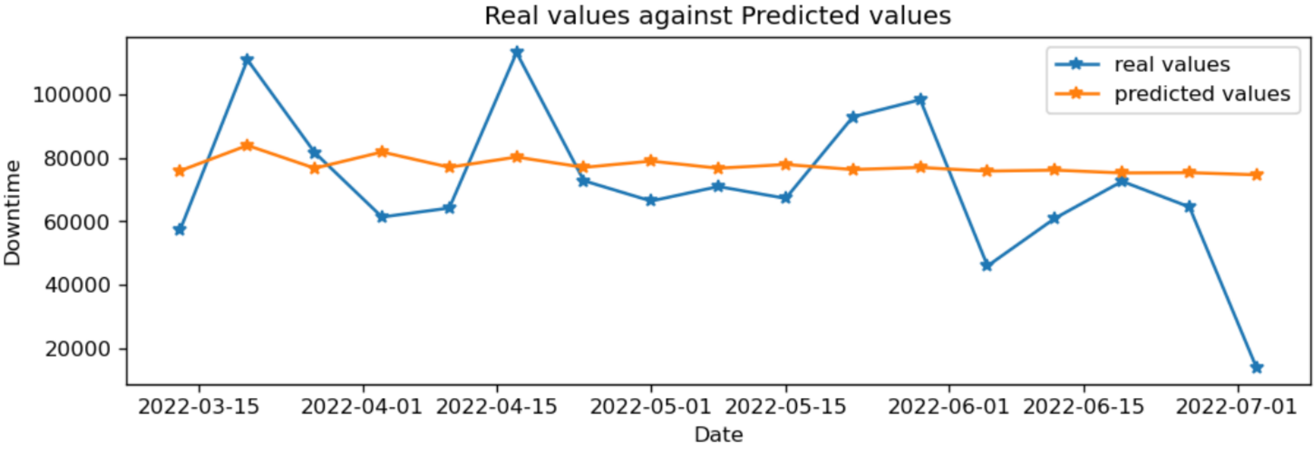


FIGURE 14
PREDICTIONS AGAINST ACTUALS

Measuring accuracy with error metrics:
Mean square error(MSE) – 517508683.94, Root mean square error(RMSE) – 22748.81and Mean absolute percentage error(MAPE) – 0.46

DISCUSSION

Our research question is predicting and reducing downtime. So to predict the downtime we utilized two statistical prediction techniques one is multiple linear regression and the other one is ARIMA modeling using time series analysis.

The Linear regression model achieved an R square value of 0.364 meaning only 36 percent of the variation in the downtime is explained by our model which is built with “Runtime” and “ShiftDesc” as the features or independent variables. Also the multiple linear regression model did not satisfy the assumption of regression that the residuals are normally distributed. During our error metric analysis on the validation dataset the model achieved an RMSE value of 5729 which is not particularly impressive to make any predictions since it failed the assumptions.

The ARIMA model was built on the weekly downtime which is the time series data. The log transformed data with parameters (1,1,2) for (p,d,q) gave the best accuracy during the error metric analysis on validation data where the MAPE value is 0.46,which means our model is giving predictions with 46 percent error as seen in Fig (14). But unlike the multiple linear regression model the ARIMA model did satisfy all the assumptions of regression i.e, no correlation among the residuals. So we made predictions for the next 10 weeks for which the data wasn’t available.

The total downtime for those 10 weeks predicted was 804526 seconds or 223 hours. From our classification we found that 80 percent of the downtime is caused by three losses i.e, 1) want of operator (23.2%), 2) cleaning and inspection (19.01%), 3) tool change (18.07) and 4) want of power (10.06%). By addressing these causes for downtime we can reduce the downtime.

As mentioned in the literature survey [3], there are many different machine learning approaches taken by researchers to study the key performance indicator OEE. Even the case we studied was having OEE as a KPI to measure the effectiveness of the production line. According to the survey most of the research papers published we’re addressing the quality ratio of the OEE. But we have successfully implemented a ARIMA model which uses time series analysis to predict downtime. If we can reduce the predicted downtime as mentioned above, we can increase the availability which in turn improves the overall equipment effectiveness (OEE).

CONCLUSION

The aim of the research was to predict the downtime of a CNC machine by utilizing the downtime data collected from the dashboard. The four major downtime losses are want of operator, cleaning and inspection, tool change and want of power which contributes to 80 percent of the downtime over the 2 year period (approx.) Though we were able to build the statistical models, the performance of the multiple regression model was not good enough which is attributed to data insufficiency where the model lacked independent variables to explain the variance in the downtime data. If we can collect some more data from machine sensors like temperature, tool life etc. we can build a more accurate model to predict downtime. As far as our ARIMA model, though the model was statistically significant and it satisfied all the assumptions of the model, the accuracy on validation data was low. The model can be improved also by collecting data during a longer period of time and by training the model with more data. We can conclude that building predictive models for downtime is possible provided we have sufficient data. With our prediction and by addressing the losses we can improve the Overall Equipment Effectiveness.

REFERENCES

[1] Business Analytics: The Science of Data Driven Decision Making – by U Dinesh Kumar

[2] Descriptive Time Series Analysis for Downtime Prediction Using the Maintenance Data of a Medical Linear Accelerator by KH Kim - https://www.mdpi.com/2076-3417/12/11/5431

[3] Machine learning applications in production lines: A systematic literature review by Ziqiu Kang et al-https://www.sciencedirect.com/science/article/pii/S036083522030485X

[4] Analysis of Overall Equipment Effectiveness by M Sayuti et al-https://www.researchgate.net/publication/333678538_Analysis_of_the_Overall_Equipment_Effectiveness_OEE_to_Minimize_Six_Big_Losses_of_Pulp_Machine_A_Case_Study_in_Pulp_and_Paper_Industries_Analysis_of_the_Overall_Equipment_Effectiveness_OEE_to_Minimize

[5] Nakajima S (1982) TPM tenkai. Japan Institute of Plant Maintenance, Tokyo