REVA UNIVERSITY
Bengaluru, India

A Project Report on

Pattern Discovery and Forecasting of Attrition

Submitted in Partial Fulfilment for Award of Degree of
Master of Business Administration
In Business Analytics

Submitted By
Saumyadip Sarkar
R19MBA07

Under the Guidance of
Phaneendra Akula
Senior Data Science Manager

REVA Academy for Corporate Excellence - RACE
REVA University
Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064
race.reva.edu.in

August, 2022

REVA
UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, Saumyadip Sarkar hereby declare that I have completed the project work towards the Master of Business Administration in Business Analytics at REVA University on the topic entitled '**Pattern Discovery and Forecasting of Attrition**' under the supervision of Phaneendra Akula.

This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year 2022.

Saumyadip Sarkar

Place: Bengaluru
Date: 27-Aug-2022

Saumyadip Sarkar
Signature of Student

REVA
UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the project work entitled ‘**Pattern Discovery and Forecasting of Attrition**’ carried out by Saumyadip Sarkar with SRN R19MBA07, is a bonafide student of REVA University, is submitting the project report in fulfilment for the award of Master of Business Administration in Business Analytics during the academic year 2022. The Project report has been tested for plagiarism and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said degree.

A. Phaneend

Phaneendra Akula
Guide

Dr. Shinu Abhi
Director

External Viva
Names of the Examiners

1. Vaibhav Sahu, Strategic Cloud Engineer, Google
2. Abhishek Sinha, Data Science Manager, Capgemini

Place: Bengaluru
Date: 27-Aug-2022

REVA
UNIVERSITY
Bengaluru, India

Acknowledgement

I would like to express my sincere gratitude to our Chancellor **Dr. P. Shyama Raju**, **Dr. S.Y. Kulkarni**, Chancellor, **Dr. M Dhananjaya**, Vice-Chancellor **and Dr. N Ramesh**, Registrar for supporting the RACE program, specifically designed for working professionals, and providing necessary facilities and infrastructure required for the best learning experience. I am immensely proud of being part of this program and REVA University.

I would like to thank **Dr. Shinu Abhi**, Director, and **Dr. Rashmi Agarwal**, Assistant Professor, RACE (REVA Academics for Corporate Excellence) for their guidance and constant supervision in providing necessary information regarding the project thus ensuring complete adherence to project guidelines.

A special thanks to my project guide **Mr. Phaneendra Akula** for his valuable guidance provided in understanding the concept and executing the project. This list cannot be complete with mentioning about **Dr. J. B. Simha**, without whom the project could not have started. He showed us our starting point.

I am also thankful to my fellow classmates for their constructive criticism and friendly advice during the project work.

Last but not the least, I would like to thank my family and colleagues who have extended their invaluable co-operation and encouragement throughout this project lifecycle.

Place: Bengaluru	Name of the Student:
Date:27-Aug-2022	Saumyadip Sarkar

REVA UNIVERSITY
Bengaluru, India

Similarity Index Report

This is to certify that this project report titled **Pattern Discovery and Forecasting of Attrition** was scanned for similarity detection. Process and outcome are given below.

Software Used: Turnintn
Date of Report Generation: 28-Aug-2022
Similarity Index in %: 6%
Total word count: 6364
Name of the Guide: Phaneendra Akula

Saumyadip Sarkar

Place: Bengaluru
Date: 27-Aug-2022

Saumyadip Sarkar
Signature of Student

Verified by: M N Dincy Dechamma

Signature
Dr. Shinu Abhi,
Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	ETS	Exponential Triple Smoothing
2	ARIMA	Auto Regressive Integrated Moving Average
3	SARIMA	Seasonal Auto Regressive Integrated Moving Average
4	TS	Time Series
5	LOB	Line of Business
6	SLA	Service Level Agreement
7	HWS	Holt-Winters Smoothing
8	MAD	Mean Absolute Deviation
9	RMSE	Root Mean Square Error
10	MAPE	Mean Absolute Percentage Error

List of Figures

List of Tables

Abstract

A specific organization’s Transportation line of business has been seeing a lot of attrition in the last couple of years. The average attrition is around 34% the for last 3 years which is way above the industry average. Time and again, it has struggled with managing a healthy attrition rate. As a result, there has been a constant occurrence of missed SLAs resulting in huge penalties. This has adversely affected customer feedback and employee morale as well. In extreme cases, the contracts were not renewed. High attrition is also a very expensive affair for a company.

For managers, managing workload has become extremely tedious in the current context. This has been always highlighted in every Quarterly Business Review. Management has also accepted this fact, but the challenge still remains as there have been no improvements noticed in last 12 months. With the constant change in management, the focus kept shifting from one approach to another.

The aim of this study is to forecast attrition using time series analysis at various levels based on only the attrition data available for the last 14 months.

The proposed solution is to compare various Time Series Forecasting techniques like ARIMA, Seasonal ARIMA, Exponential Smoothing, Holt-Winters, Moving Average, Ratio to Moving Average, based on Attrition Date of last 12 months at Enterprise level and drill-down by top contracts with further drill-down by employee levels, pay range & locations. The forecasted data will then be compared with the following 2 months attrition data to arrive at the best possible solution.

What’s unique about this study is the use of timeseries forecasting technique to identify future attrition trends based on only attrition data. This forecasted data can be used to better workload management which will in turn reduce missed SLAs and penalties.

Keywords: Forecasting, Time series, ARIMA, Seasonal ARIMA, Exponential Smoothing, Holt-Winters, Data Discovery, Pareto, Trend Analysis, Regression, Moving Average, LSTM

Contents

Chapter 1: Introduction

For any organization, finding the right candidate is of paramount importance. However, this is just the first step. Then there starts a long journey of fitting the candidate into the right job. This is not only time intensive, but also involves a huge cost of orientation and training (Ner, 2020). However, when the same candidate decides to leave the organization within a year, it creates a huge void jeopardizing the entire setup with a cascading effect going down till customer feedback, and what if 34 % of regular employees keep moving out every year, what if there are frequent reshuffles at the top level, and ironically if there is lack of support from human resource owing to high attrition at HR department itself, the problem gets extrapolated many folds.

That's precisely the condition the transportation Line of Business (LOB) is going through. With attrition going up the roof, lack of management & HR support, and lack of availability of proper HR data, the transportation LOB is struggling to meet their SLAs resulting in a huge outflow of money in form of penalties. Customers are not satisfied and so do employees.

The hypothesis here is, if probable attrition is forecasted well in advance at various levels, a plan can be put in place to hire employees and make them available whenever there is demand from contract managers.

Taking this concept forward, top contracts with high attrition rate can be equipped with the forecast model which can be used periodically to predict attrition trends and give them enough time to create provisions for probable attrition.

In recent times, human resource departments of various organizations have been trying to map employee life cycle which involves a lot of phases (Smither, 2003). Using machine learning algorithms, predictions are being made about employee tenure within organizations (Singh Sisodia et al., 2017). However, when there is a lack of support from human resource and the availability of good HR data, it comes down to individual contracts to manage their own human resources in an efficient way and especially managing the menace of high rate of attrition.

Is there a way to come out with a solution that is easy for end users like the tried and tested timeseries forecasting model? The study aims to achieve this and in turn help individual contract managers manage their workload efficiently, and reduce their missed SLAs, thereby reduce penalties which sometimes run into thousands of dollars.

A step-by-step approach starting from exploratory data analysis to using the most accurate time series forecasting technique will form the core area of this study. To have a better understanding of the attrition problem and various timeseries techniques, a detailed literature review has been conducted as discussed in the next section.

Chapter 2: Literature Review

Attrition is the term used to describe when employees leave a company (either voluntarily or involuntarily), for any cause (including retirement, termination, death, or resignation). When employees leave, they take with them priceless tacit knowledge. At the same time, employee attrition is a very costly affair for any industry. The direct costs of workforce turnover include the cost of hiring new employees, the cost of training new employees, the time it takes to transition, the cost of temporary employees, the cost of lost expertise, and the cost of the job itself (Chakraborty et al., 2021).

Several studies have been conducted on employee attritions. However, most of the employee attrition studies have concentrated on using various Machine Learning Algorithms using several factors. In the study (Kumar Jain et al., 123 C.E.) , several machine learning algorithms like Decision Tree, SVM, Random Forests have been used to estimate if an employee will leave or not.

In another study (Fallucchi et al., 2020), Gaussian Naïve Bayes classifier has been used to classify if an employee will attrit or not. XGBoost classifier has also been used to classify employee attrition (Jain & Nayyar, 2018).

However, there is one study that stands out from the rest is the use of Ensemble Model Based on Machine Learning Algorithms for automated employee attrition prediction (Alsheref et al., 2022).

As seen in most of the studies related to attrition predictions, classification is the go-to approach. However, using time-series techniques to forecast future attrition has not been explored enough based on the observations during the literature review. *This establishes a unique opportunity for this study.*

Generating scientific projections based on data with historical time stamps is known as time series forecasting. It entails creating models through historical study, using them to draw conclusions and guide strategic decision-making in the future (*Time Series Forecasting: Definition & Examples | Table No.au*, 2020).

Time series analysis and forecasting are important for a variety of applications, including business, the stock market and exchange, the weather, electricity demand, cost, and usages of goods like fuels and electricity, etc., and in any setting where there are periodic, seasonal variations seen (Mahalakshmi et al., 2016) .

There are several timeseries techniques available, notably Moving Average, Exponential Smoothing, Holt-Winters Smoothing method, ARIMA, Seasonal ARIMA, LSTM, etc.

One well-known technical indicator, the moving average, is employed in time series analysis to forecast future data. Researchers have produced numerous variations and implementations of it during its evolution (Hansun, 2013). Another variation of moving average is Ratio to Moving Average, which is superior to the simple average method, is predicated on the idea that seasonal variance for any given month is a continuous component of the trend. Moving average methods reduce periodic movements if any (Sailaja & Prasad, 2019).

Another immensely popular timeseries technique is Exponential Smoothing. Its popularity is based on the fact that surprisingly accurate forecasts can be obtained with minimal effort. This has been proved in this study as well where Time-series forecasting methods is used via an Excel (FORECAST.ETS) function. The superior efficacy of this model has been nicely illustrated in the paper by Dewi Rahardja. With this Excel function, forecasting is simple and quick while considering the model's level (intercept), trend (slope), and seasonality (Rahardja, 2021).

There is another variation of Exponential Smoothing technique popularly known as Holt-Winters after the name of inventors. When there is both a trend and seasonality in the data, the Holt-Winters (HW) exponential smoothing gives a great result. The two primary HW models are multiplicative for time series exhibiting multiplicative seasonality and additive for time series exhibiting additive seasonality (Kalekar, 2004).

However, there are times when, for the same number of data, a Long Short-Term Memory (LSTM) multivariate machine learning model outperforms a Holt-Winters univariate statistical model (Ueno et al., 2020). By utilizing the nonlinearities of a particular dataset, LSTM networks can overcome the constraints of conventional time series forecasting methods and produce state-of-the-art outcomes on temporal data (Chimmula & Zhang, 2020).

From a statistical modelling perspective, another timeseries technique that produces robust result in short-term prediction is ARIMA, first introduced by Box and Jenkins in 1970. It is also known as the Box-Jenkins methodology and consists of a series of steps for locating, calculating, and diagnosing ARIMA models using time series data. A very well-researched paper available in this context is (Adebiyi et al., 2014), which shows ARIMA's strength of predicting future stock prices. The limitation of ARIMA model is however the number of periods. It is recommended to employ at least 50 and ideally 100 observations (Box et al., 1976).

While ARIMA has its own strength, when it comes to seasonal data, there is a variation of ARIMA available commonly known as SARIMA or seasonal ARIMA. For climate related data SARIMA has been a valuable tool (Dimri et al., 2020) Then there is FSARIMA, which combines the version of SARIMA and the fuzzy regression model (Tseng & Tzeng, 2002).

With a lot of statistical techniques already being widely used in forecasting techniques, recent studies have now been conducted using Deep Learning (DL) models showing outstanding results when compared to traditional forecasting techniques. A comparative study (Sezer et al., 2019) has shown that the DL method significantly improves upon machine learning (ML) models.

Then there is Facebook's Prophet Forecasting model. It is a method for predicting time series data that uses an additive model to suit non-linear trends with seasonality that occurs annually, monthly, daily, and on weekends as well as during holidays. Strongly seasonal time series and multiple seasons of historical data are ideal for it. Prophet typically manages outliers well and is robust to missing data and changes in the trend (Taylor & Letham, 2017).

Chapter 3: Problem Statement

It has been observed that, in the last 12 months, close to $2,082k has been paid in terms of penalties (*data collected from the organization*). Penalties are incurred when SLAs are not being met. This also leads to negative customer feedback and in extreme cases, non-renewal of contracts or even termination.

One of the main reasons identified is “attrition” which presently stands at 34%. This has almost remained constant for the last three years.

With constant churn at top management and lack of HR support, there seems almost no headway in managing attritions. Ironically, the lack of HR support is due to the fact that HR department has failed to manage its own attrition rate.

The problem is further exacerbated due to the lack of proper data collection at the HR end. With a lack of manpower in HR, the onus has fallen on individual contracts to manage their own attrition. The data shared was only attrition data for the last 14 months. Request for other data points like current headcount data and salary information of the current employees were turned down citing confidentiality of information.

With the limited data, will it be possible to come out with a solution that will aid contract/account managers to forecast probable attrition and create provisions accordingly? Will it help reduce penalties on missed SLAs and improve customer sentiments? These are a few answers this study aims to find.

Chapter 4: Objectives of the Study

Considering the present problem area, this study aims to do the following:

1. Identify the required attrition dataset. In this case, only the attrition data of last 14 months are available.

2. Study data at various levels and categories to identify trends, patterns, and top contributors which then will be used to create subsets of the main dataset for modelling.

3. Explore various timeseries techniques to identify the best timeseries forecasting model which can be used to forecast future attrition.

4. Propose future actions which will help managers manage workload better to reduce missed SLAs and penalties.

Chapter 5: Project Methodology

This project uses CRSIP-DM framework which begins with understanding the business as a whole and then narrowing it down to a specific area. Figure No. 5.1 shows a typical life cycle of the process.

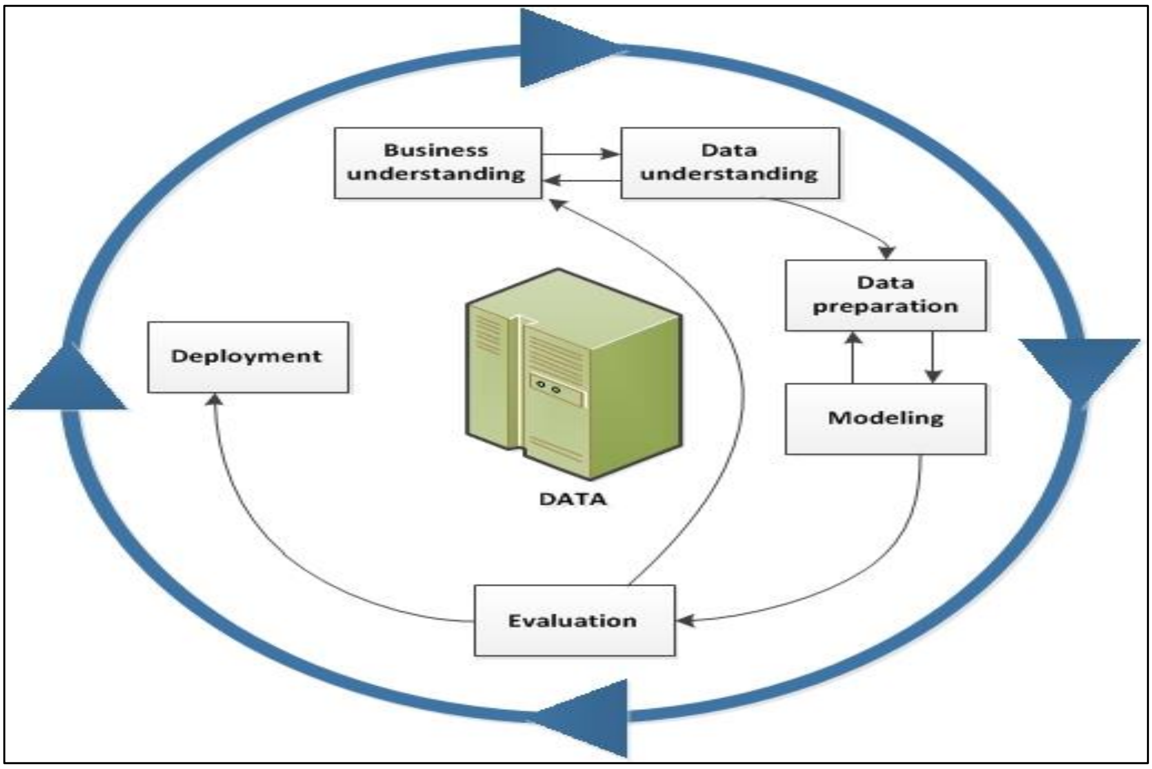


Figure No. 5.1 CRISP-DM Framework (*CRISP-DM Help Overview - IBM Documentation*, 2021)

In this case, it is about the transportation line of business which is reeling under huge attrition for the last 3 years and there seems no solution to bring it down to a desired level. Transportation LOB mainly deals with automated tolling, automated parking, and public safety. Apart from providing technological solutions, a substantial number of employees are deployed to resolve disputes. Dispute resolutions form a critical aspect of SLA, as missing on agreed TAT for dispute resolutions leads to missed SLAs which is turn incur penalties.

The next step in CRISP-DM process is data understanding. Here, the data is the employee attrition record captured for the last 14 months. The other data points are missed SLA numbers at contract levels and penalties paid at contract levels for last 14 months.

In this next step of data preparation, the main goal is to identify if the data is fit for timeseries analysis. This involves looking at trends to see if the trend is strong. The next step is to look for any seasonal trends. Based on the findings, data transformation can be done to make the data stationary.

The main approach in the modeling phase is to select the best timeseries forecasting technique like Moving Average, ARIMA, Seasonal ARIMA, Exponential Smoothing, and Holt Winters. This is done at various levels and categories to see how the timeseries model is performing across various subsets of the main dataset.

Post modeling technique, the evaluation phase starts. In this phase, the efficacy of the various timeseries forecasting techniques is assessed. The one which most accurately mimics the test data would be finalized.

In Deployment phase, the forecasted numbers can be used to hire a pool of employees. These employees will then be suitably used in various contracts based on the current need.

Chapter 6: Business Understanding

The transportation line of business is one of the most profitable units in this organization. It provides the following solutions as depicted in Table No. 6.1 to its clients.

Solutions	Description
Automated Tolling	Captures vehicle details when a tolling both is crossed and bills customer accordingly. A team also works on dispute resolutions pertaining to technical failure, failed auto-debit attempts, customer complaints, etc.
Automated Parking	It provided intelligent parking solutions mainly for governments. The solution involves fee collections, dynamic pricing, enforcement solutions, etc.
Public Safety	It provides automated photo enforcement, traffic violation solutions, etc.

Table No. 6.1 Transportation Solutions

However, it's not rosy all throughout. One of the biggest headaches the transportation line of business is dealing with is abnormal attrition rate for the last 3 years which on average is hovering around 34 % and there seems no trick working for them.

Along with high attrition, constant change in management and lack of HR support have made life difficult for contract managers to manage their workload efficiently.

Apart from providing technological solutions, there are dedicated teams working on resolving disputes pertaining to technological failures, missed billing, payment disputes, failed payments, etc. There are agreed turnaround time for resolving disputes which is part of service level agreement with individual contracts depending upon the nature of support.

There are also several SLAs linked with penalties. A missed SLA will incur penalties to be paid to the client. It's been observed that almost $ 2,082k penalties were paid in the last 14 months by accounts reeling under huge attritions.

Since dispute resolution is manually intensive work, more employees need to be allotted during peak times apart from ensuring regular availability. But with a lack of available manpower due to high attrition, dispute resolutions often lead to missed TAT, and missed TAT directly influences SLAs which in turn leads to penalties being levied at contract levels.

This study aims to forecast probable attrition, which in turn will help plan healthy workload management. This will in turn help control missed SLAs and bring penalties to an acceptable limit.

Chapter 7: Data Understanding

Data in this study is only attrition data collected for the last 14 months.

The following are some of the important parameters present in the attrition data.

1. Employee details – ID, Employee Name, Salary, Last Performance Rating
2. Employment Details - Employee Type (Regular or Contract), Joining Date, Termination Date, Employee Level, Type of Termination, Termination Code, Cost Centre, Job Name
3. Contract Details – Contract Name, Sector, Business Category, Location City, Country

For this analysis, the focus is on the contract level. The following diagram as appears in Figure No. 7.1 shows a Pareto Chart of Attrition vs Contracts. It clearly shows top 6 which represents 15% of overall contracts are contributing to more than 80% of attrition.

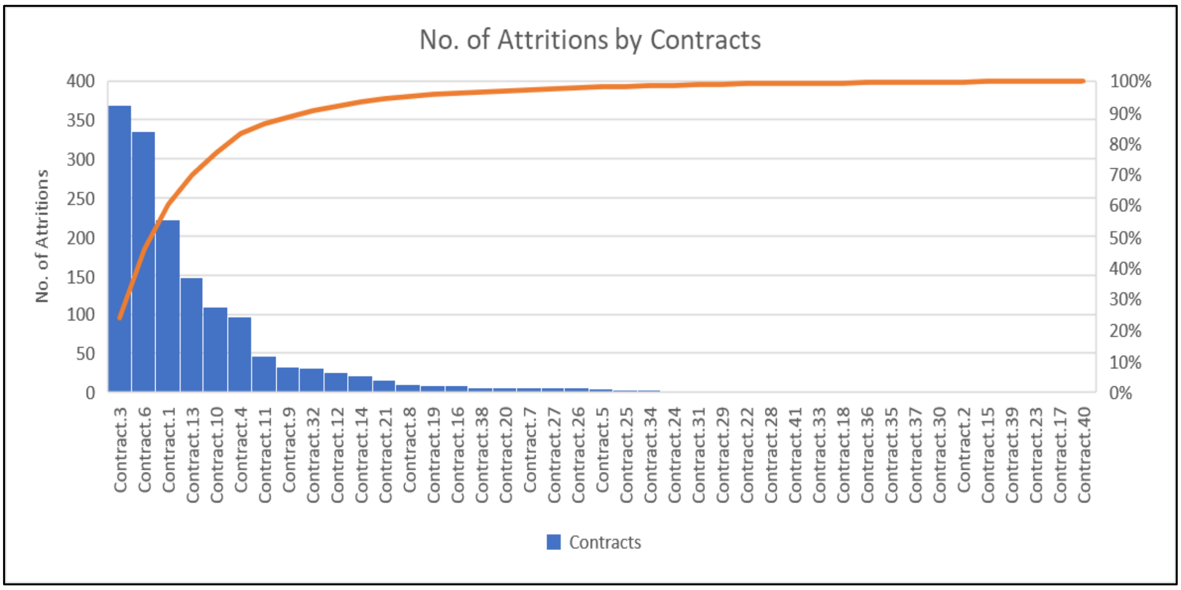


Figure No. 7.1 Pareto – No. of Attrition by Contracts

The next check that's done is to see the penalty contributions by the top 6 contracts.

The following diagram as shown in Figure No. 7.2 shows the penalties paid by the top 6 contracts which total to $ 1,051k representing ~50% of overall penalties.

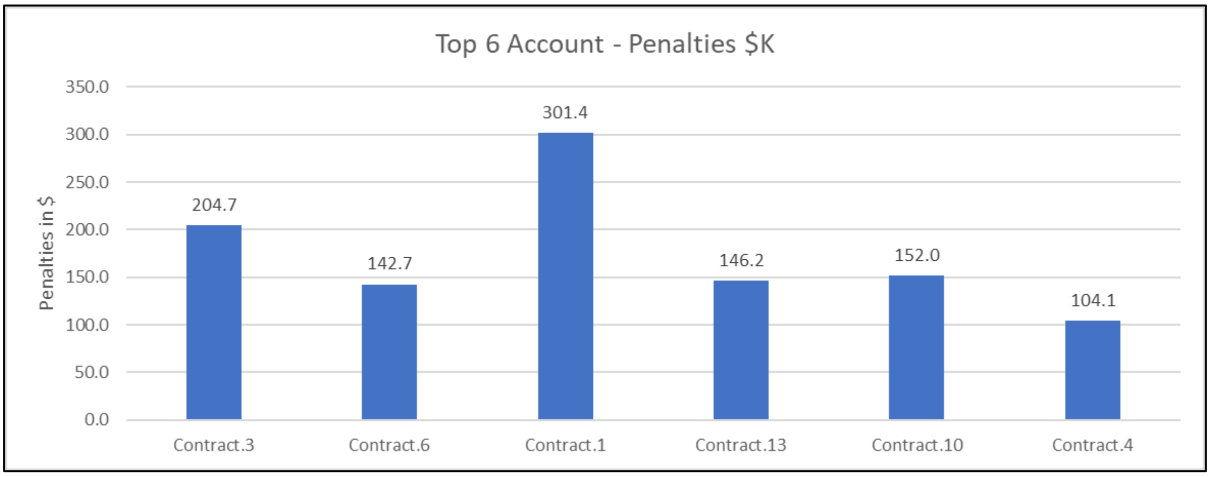


Figure No. 7.2 Penalties by top 6 Contracts

Similarly, Figure No. 4 shows the number of missed SLAs by the top 6 contracts.

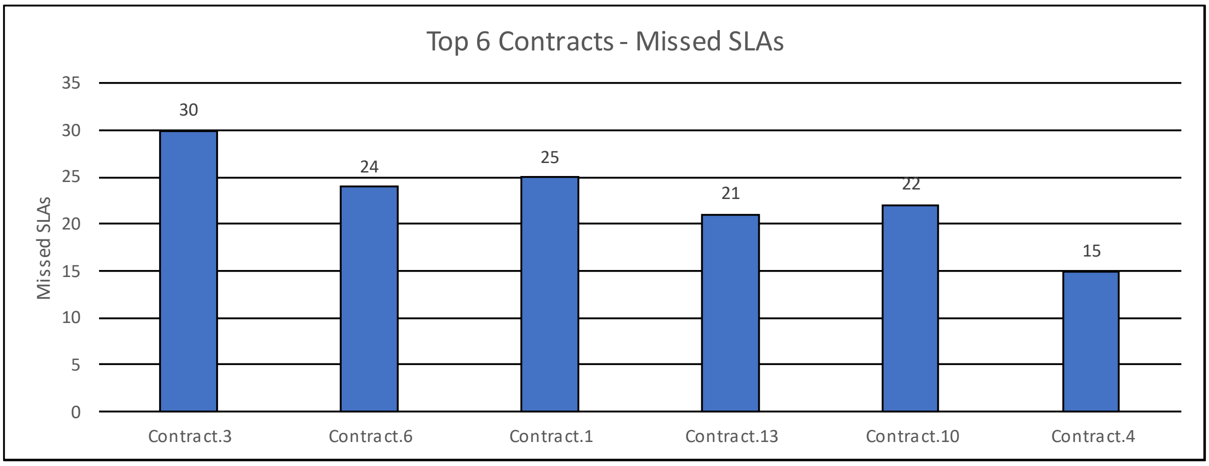


Figure No. 7.3 Missed SLAs by top 6 Contracts

It's also observed that there is also a strong correlation of 0.93 exists between contract level attrition numbers and missed SLAs. Figure No. 7.4 shows a scatter plot relation between Attrition numbers and Missed SLAs at contract levels.

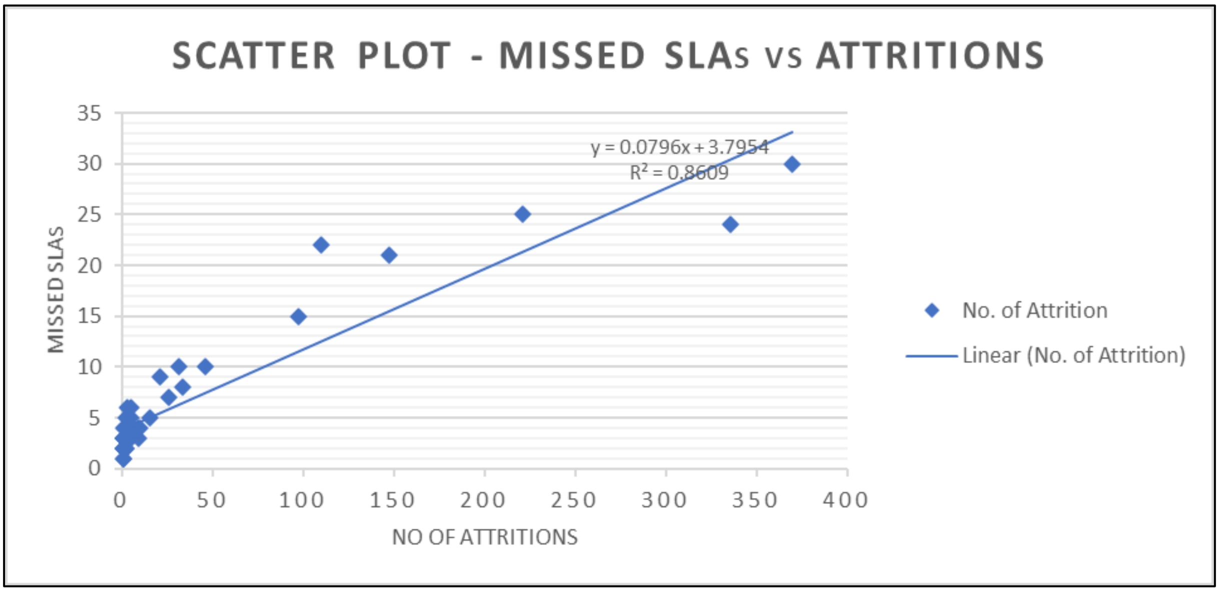


Figure No. 7.4 Scatter Plot - Missed SLAs vs Attrition

An almost similar strong correlation is observed between Missed SLAs and Penalties paid at contract levels as shown in Figure No. 7.5.

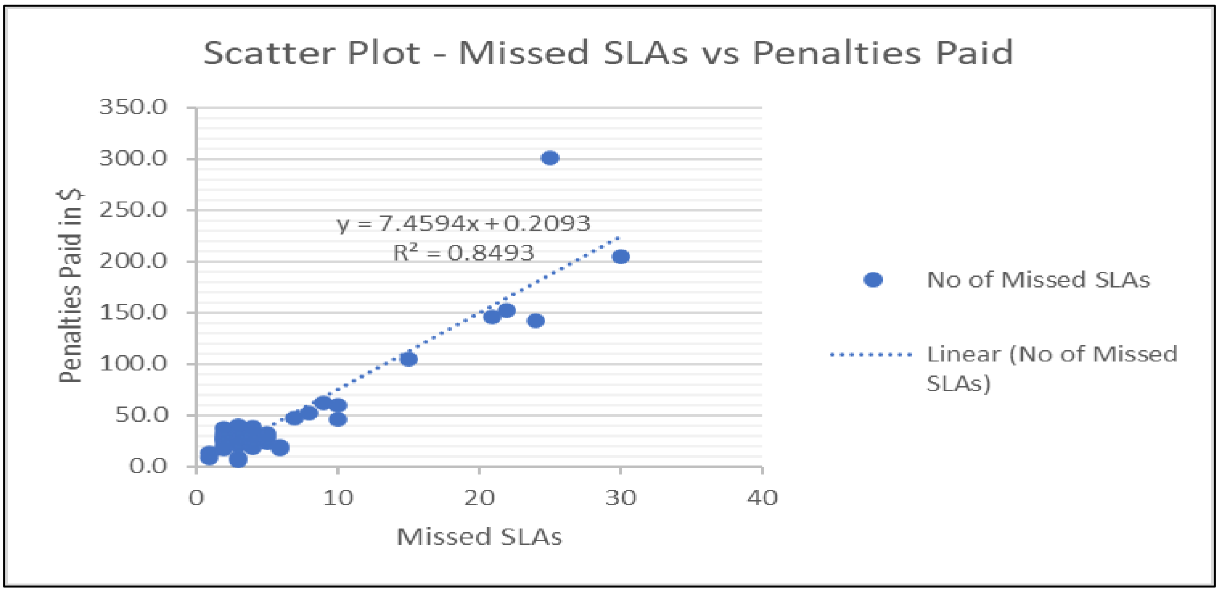


Figure No. 7.5 Scatter Plot - Missed SLAs vs Attrition

Another key point that came out prominently is that close to 90% of attrition is happening at the junior most level (C01). Figure No. 7.6 highlights this fact.

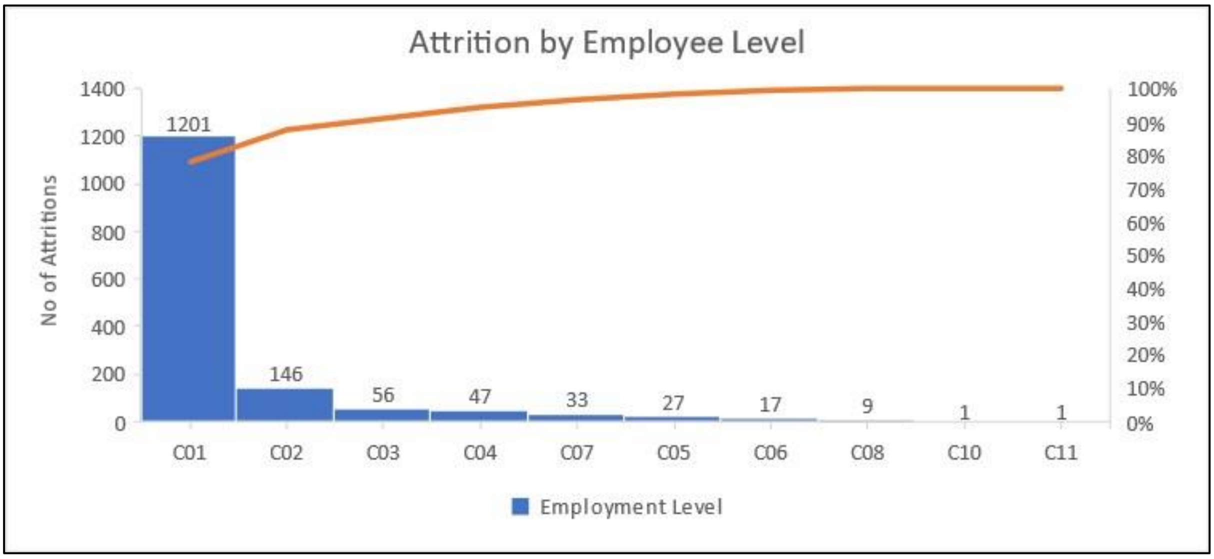


Figure No. 7.6 Pareto Graph - Attrition contribution by Employee Level

This finding remains consistent with our overall contract level attrition trend as shown in Figure No. 7.7. The top 6 contracts remained constant.

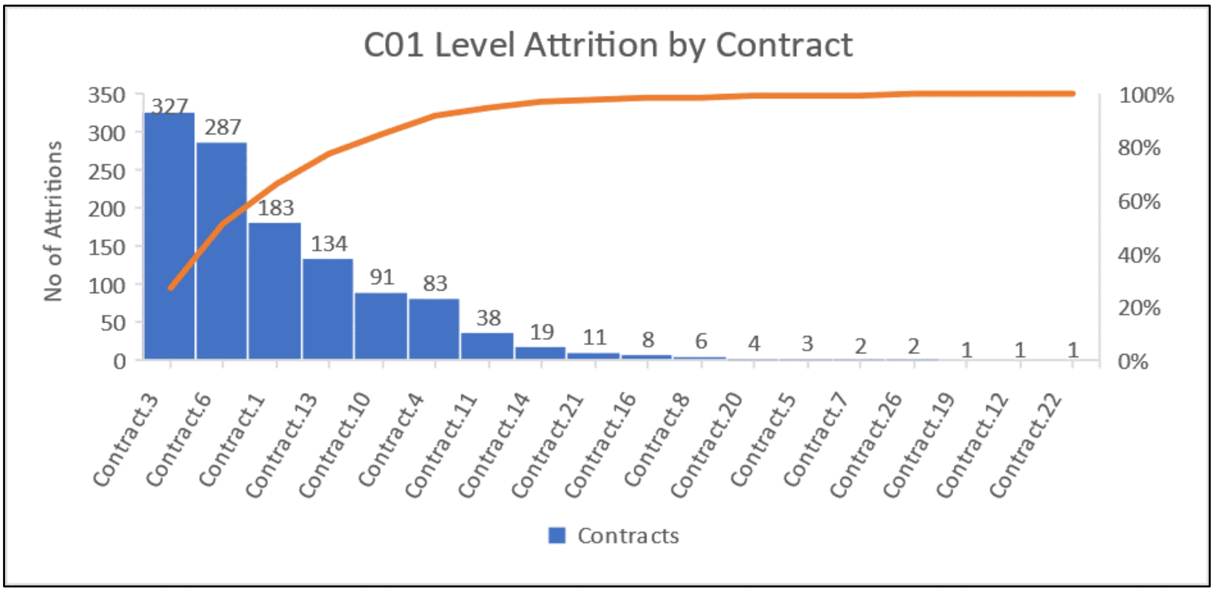


Figure No. 7.7 Pareto Graph - Attrition contribution by C01 Level at contract level

A few other levels that are considered for this study are location-wise attrition and salary range. At various locations, Figure No. 7.8 shows that the top 11 cities contributed to 80% of attritions in the last 12 months.

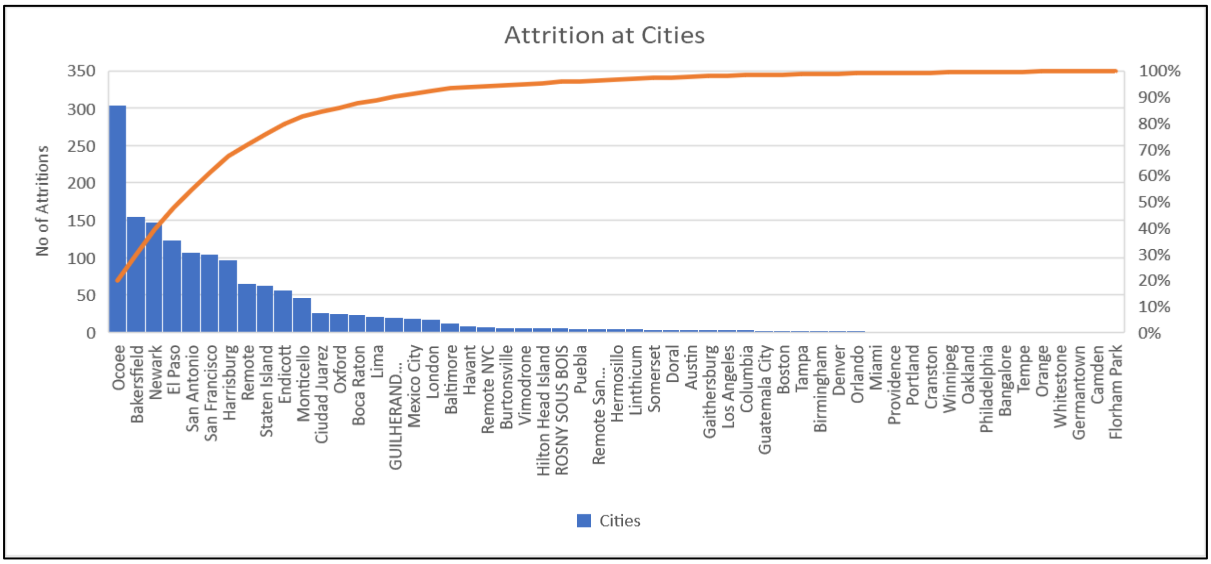


Figure No. 7.8 Pareto Graph - Attrition contribution by cities

The final data, that's investigated is attrition at salary level. To find a particular range contributing to attrition, the salary bucket is created as shown in Figure No. 7.9. Salary ranging from $20k to $40k contribute to maximum attrition.

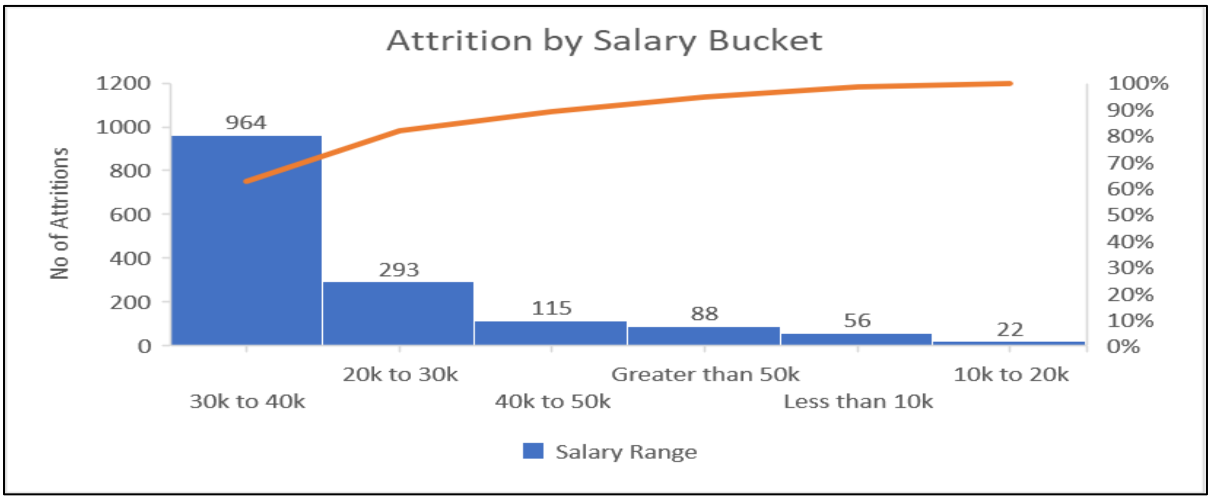


Figure No. 7.9 Pareto Graph - Attrition contribution salary range

To summarize, if the attrition forecast is concentrated on the top 6 contracts which are leading to 80% of overall attrition and contributing to 50% of overall penalties, the majority of the issues related to missed SLAs and penalties can be addressed. The final expected outcome is to enable contract managers manage their workload efficiently by deploying the right number of employees at the right time for the right job. This will in turn ensure fewer missed SLAs and eventually less penalties.

Chapter 8: Data Preparation

Since the data for this study is only attrition data, the first step of data preparation involves creating various levels which can later be used in the modeling phase.

The first level that's created is quarters based on the termination date. The data contains four full quarters of data starting from Apr-21. Since the attrition data has only 14 months of data, the first 12 months of data have been used for modeling and the last 2 months of data have been used for testing forecast accuracy.

Again, based on the termination date, the month with MMM-YY format was created. The quarter and the month form the basis of forecasting data.

The salary bucket is created to understand if there is any particular salary range that is contributing to high attrition. The buckets created here are "Less than 10k", "10k to 20k","20k to 30k", "30k to 40k", "40k to 50k" and "Greater than 50k".

To summarize, the data is divided into the following six categories for our forecasting purpose. This has been done to take into consideration all the levels which might have a significant impact on the forecasting results.

1. Overall Attrition by quarter and month-wise
2. Attrition by top contracts, quarter and month-wise
3. Attrition by top employment levels, quarter and month-wise
4. Attrition by top contracts and top employment level, quarter and month-wise
5. Attrition by top Cities, quarter and month-wise
6. Attrition by top salary ranges, quarter, and month-wise

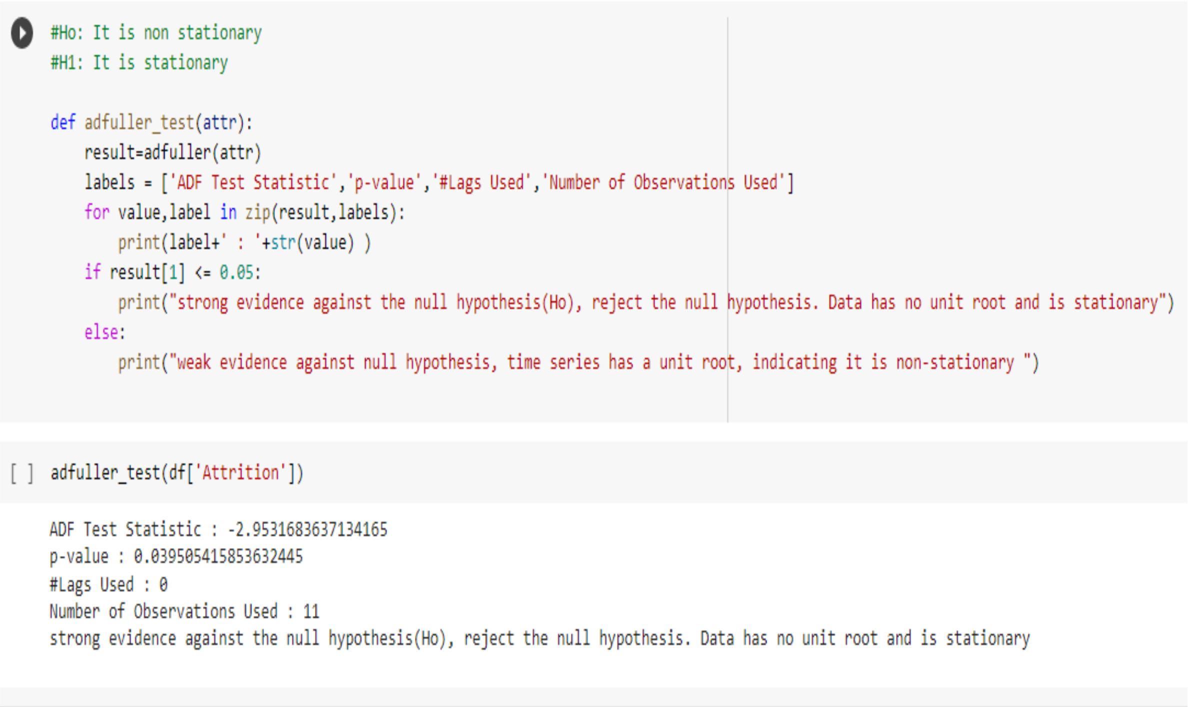
The above categorizations form the basis of the forecast modeling.

Chapter 9: Modeling

Having categorized the data as discussed in the last chapter, now is the time to investigate the modeling techniques used for this study.

Considering the limitation of the available information in the dataset, various timeseries forecasting techniques are considered on each of the categorized levels of data and compared. The top technique is chosen to predict future attrition. The forecasted attrition result can then be further used as the input of regression to predict missed SLAs numbers and the predicted missed SLAs to predict probable penalties. These approaches are discussed in detail in the following sections.

The modeling approach started first by taking the overall data and checked for stationarity of the data using Dickey-Fuller test. The low p-value of 0.0395 which is less than 0.05 indicates that the data has no unit root and is stationary. This result remained constant for all other 5 datasets. The result of Dickey-Fuller test on overall data is shown in Figure No. 9.1.



```
#Ho: It is non stationary
#H1: It is stationary

def adfuller_test(attr):
    result=adfuller(attr)
    labels = ['ADF Test Statistic','p-value','#Lags Used','Number of Observations Used']
    for value,label in zip(result,labels):
        print(label+' : '+str(value) )
    if result[1] <= 0.05:
        print("strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data has no unit root and is stationary")
    else:
        print("weak evidence against null hypothesis, time series has a unit root, indicating it is non-stationary ")
```

```
[ ] adfuller_test(df['Attrition'])
```

```
ADF Test Statistic : -2.9531683637134165
p-value : 0.039505415853632445
#Lags Used : 0
Number of Observations Used : 11
strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data has no unit root and is stationary
```

Figure No. 9.1 Dickey-Fuller test

Then a quick check on the trend graph shows an overall upward trend as appears in the Figure No. 9.2 below. This is a bit contradictory to the findings in Dickey-Fuller test. Hence, several timeseries forecasting techniques are tried to find the most suitable and robust model.

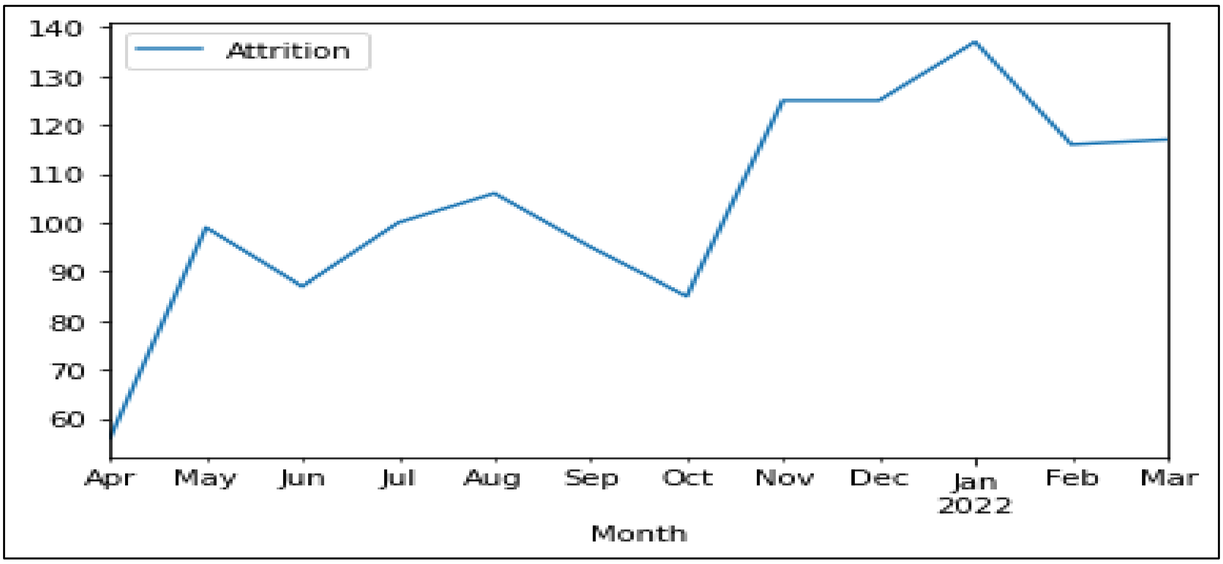


Figure No. 9.2 Overall Attrition trend

This trend is almost similar to all other 5 subsets of the data used here in the technique. Figure No. 9.3 shows the trend for the top 6 contracts, Figure No. 9.4 for the C01 employee level, and Figure No. 9.5 for the top 6 contracts at the C01 level.

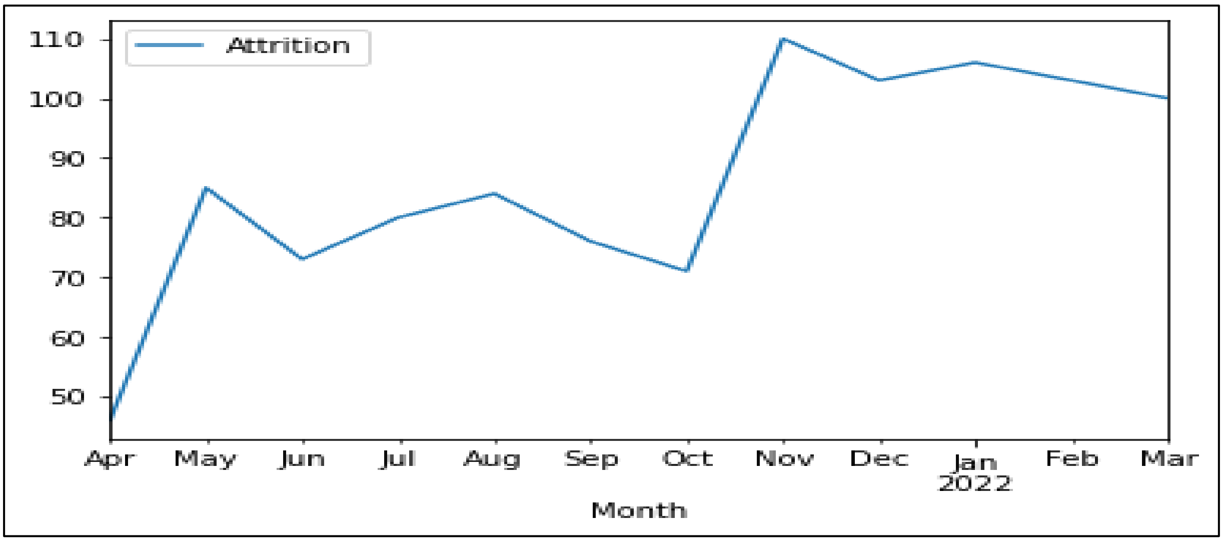


Figure No. 9.3 Attrition trend - top 6 contracts

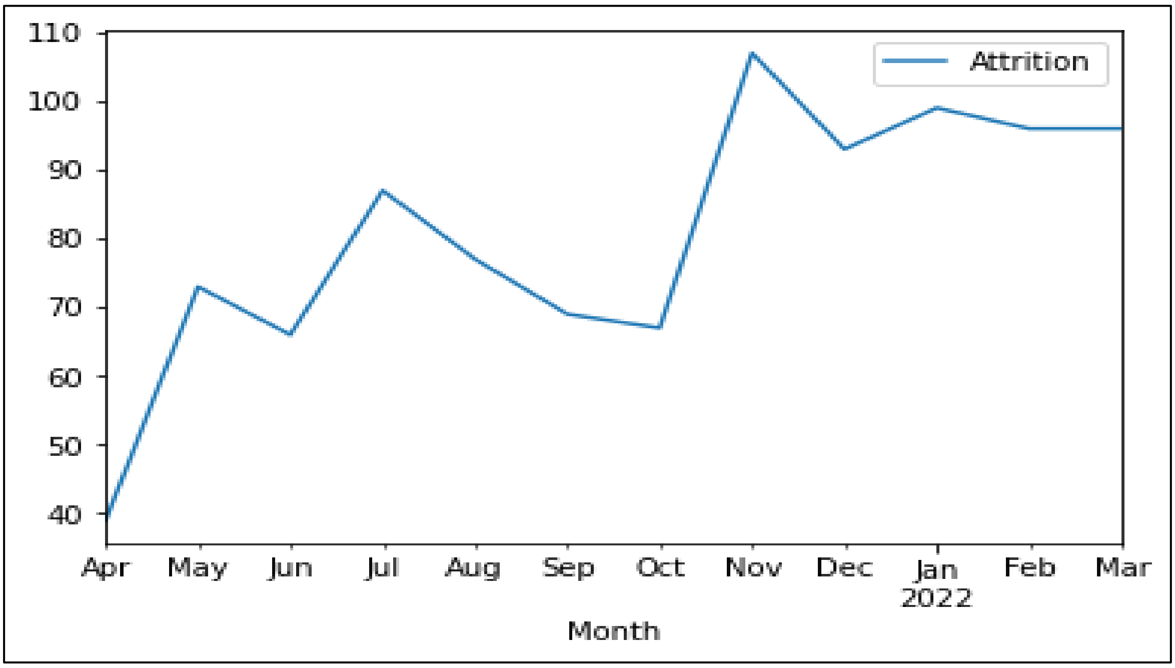


Figure No. 9.4 Attrition trend at C01

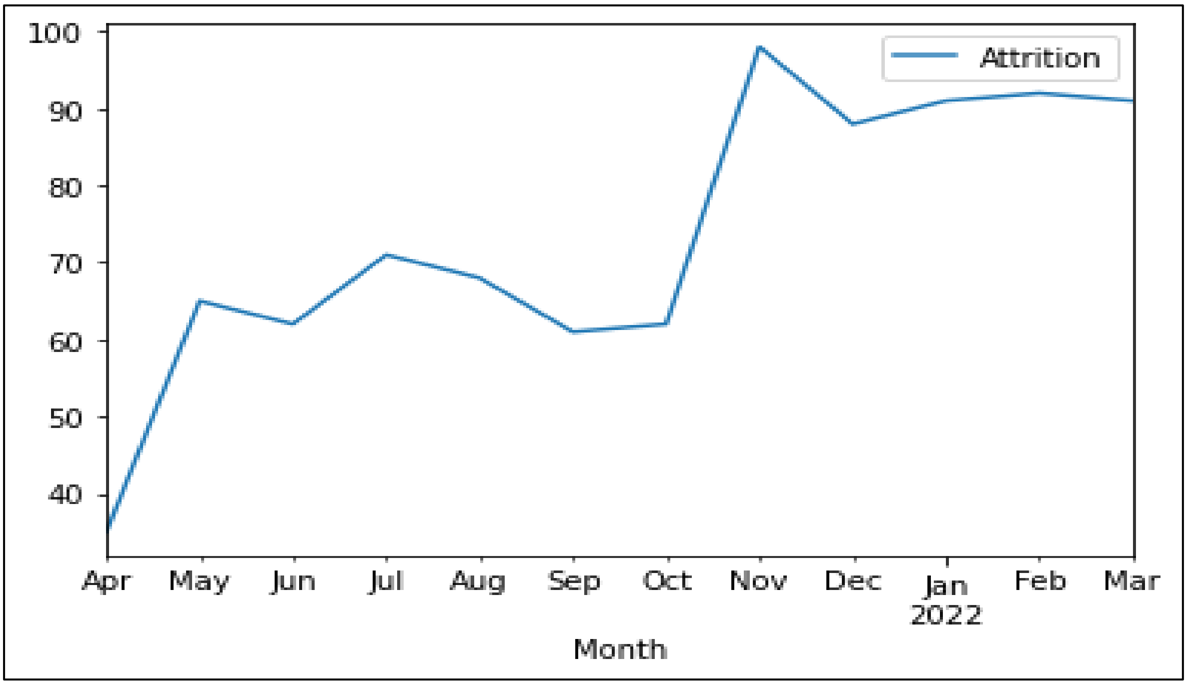


Figure No. 9.5 Attrition trend at Top 6 Contracts (C01 level)

Figure No. 9.6 and Figure No. 9.7 show the trend for top cities and top salary buckets, respectively.

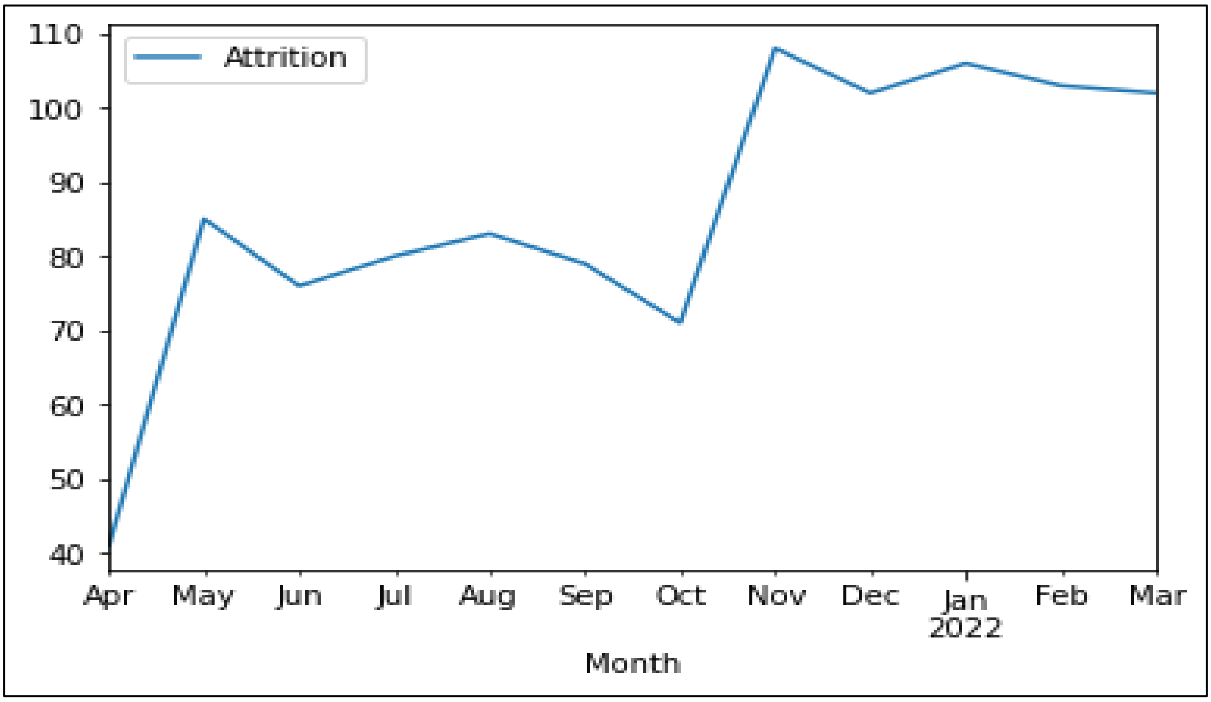


Figure No. 9.6 Attrition trend at Top Cities

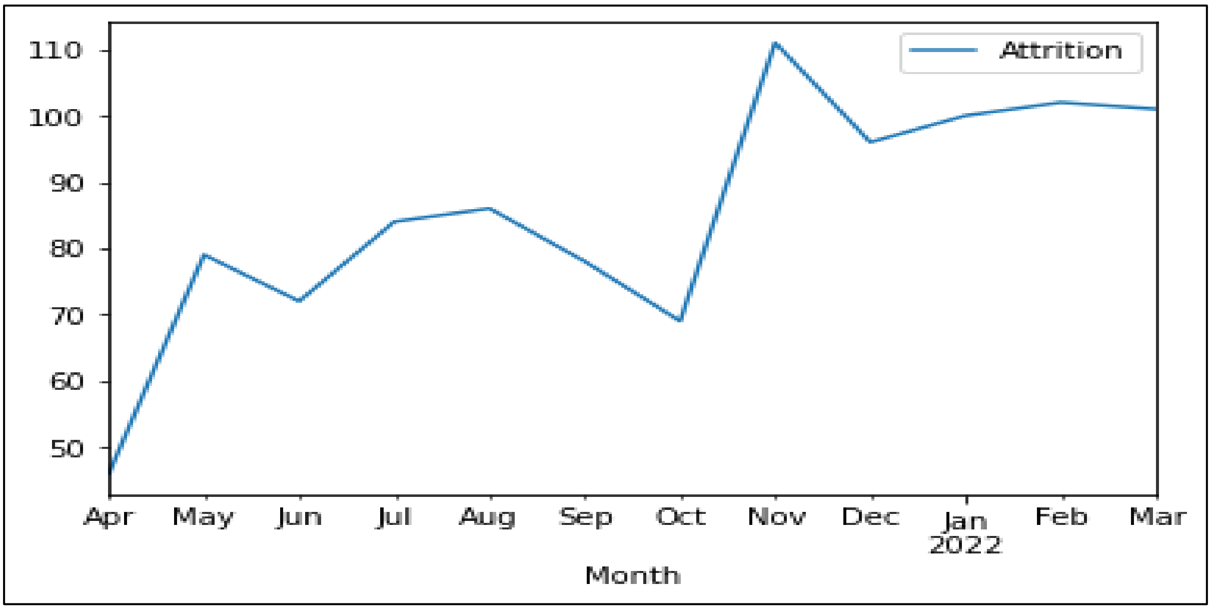


Figure No. 9.7 Attrition trend at Top Salary Bucket

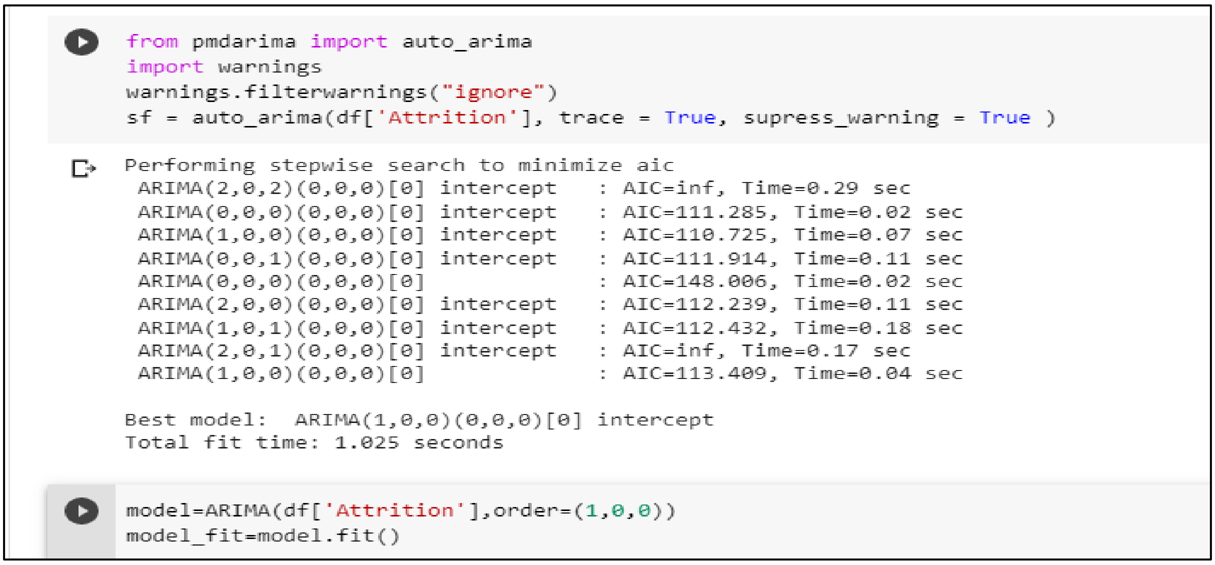
The first technique, that's used is "moving average", which is followed by "ratio to moving average" and "exponential smoothing". Moving Average, Ratio to moving average, and exponential smoothing are done using the most widely used analytical tool Excel.

In the moving average, the moving average of 3 months is considered. This is done since the months are divided into quarters with each quarter consisting of 3 months.

Building on moving average, the next technique that is used in ratio to moving average. This contains some additional steps like deseasonalizing the data and building a regression model on the deseasonalized data to forecast. This method is frequently used to show the data's overall movement without taking seasonal effects into account.

The next forecasting technique considered is Microsoft Excel’s in-built forecast algorithm. It uses FORECAST.ETS function and allows for auto-detection of seasonality. The other in-built function that is used to show forecast statistics is FORECAST.ETS. STAT. As seen in the next section, it is this technique that has given the best overall forecasted result.

The ARIMA, Holt-Winters (Smoothing 1, Smoothing 2 Additive & Multiplicative, Smoothing 3 Additive & Multiplicative), and LSTM techniques are also explored using python. For ARIMA, the auto ARIMA is used to find the best combination of the order (p,d,q). As shown in Figure No. 9.8, the best order found was (1,0,0). This order is used in ARIMA Model to forecast attrition as shown in Figure No. 9.9.



```
from pmdarima import auto_arima
import warnings
warnings.filterwarnings("ignore")
sf = auto_arima(df['Attrition'], trace = True, supress_warning = True )
```

```
Performing stepwise search to minimize aic
 ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=inf, Time=0.29 sec
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=111.285, Time=0.02 sec
 ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=110.725, Time=0.07 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=111.914, Time=0.11 sec
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=148.006, Time=0.02 sec
 ARIMA(2,0,0)(0,0,0)[0] intercept   : AIC=112.239, Time=0.11 sec
 ARIMA(1,0,1)(0,0,0)[0] intercept   : AIC=112.432, Time=0.18 sec
 ARIMA(2,0,1)(0,0,0)[0] intercept   : AIC=inf, Time=0.17 sec
 ARIMA(1,0,0)(0,0,0)[0]             : AIC=113.409, Time=0.04 sec

Best model:  ARIMA(1,0,0)(0,0,0)[0] intercept
Total fit time: 1.025 seconds
```

```
model=ARIMA(df['Attrition'],order=(1,0,0))
model_fit=model.fit()
```

Figure No. 9.8 Auto ARIMA for best model

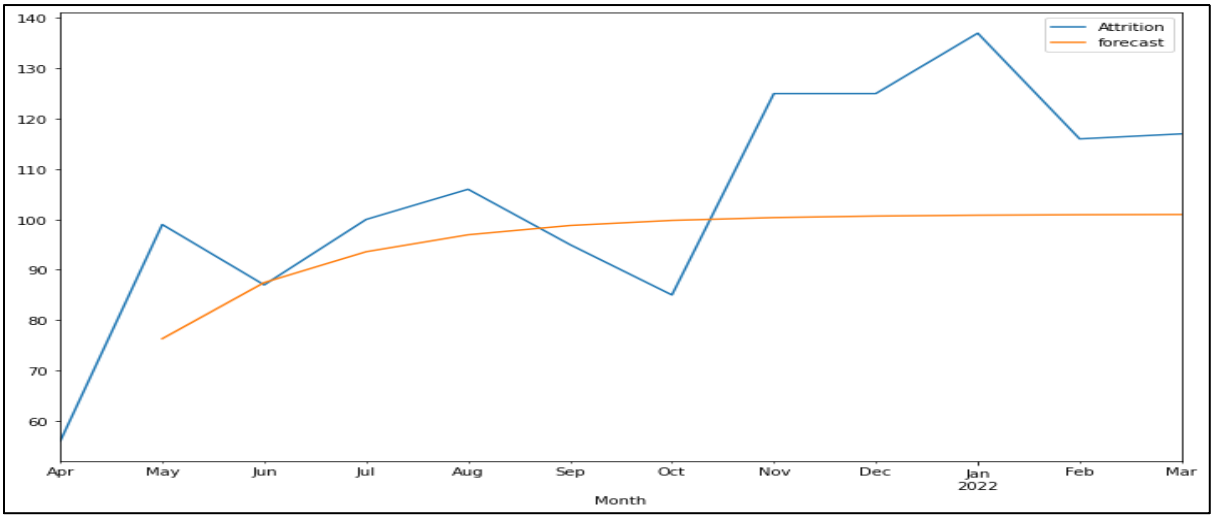
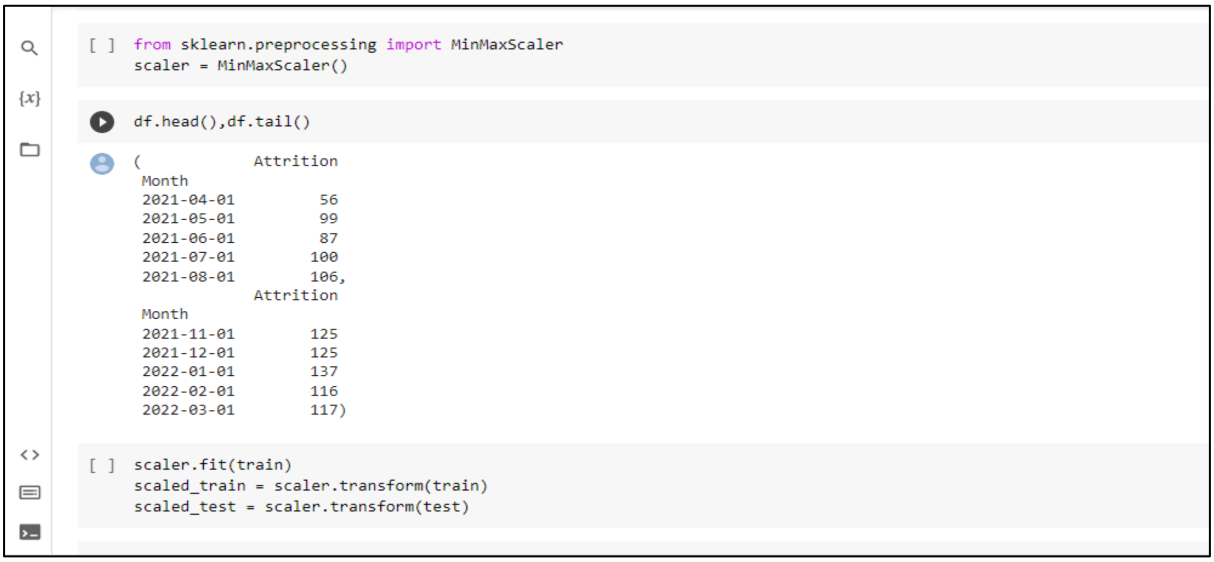


Figure No. 9.9 ARIMA Prediction Result against actual attrition

As seen clearly in Figure No 9.9 that ARIMA couldn’t perform as expected. This led to another technique called LSTM. It is a kind of recurrent neural network that can pick up order dependence in situations involving sequence prediction. The data is divided into train and test data and MinMax preprocessing technique is used on both the datasets as shown below in Figure No. 9.10.



```
[ ]  from sklearn.preprocessing import MinMaxScaler
     scaler = MinMaxScaler()
```

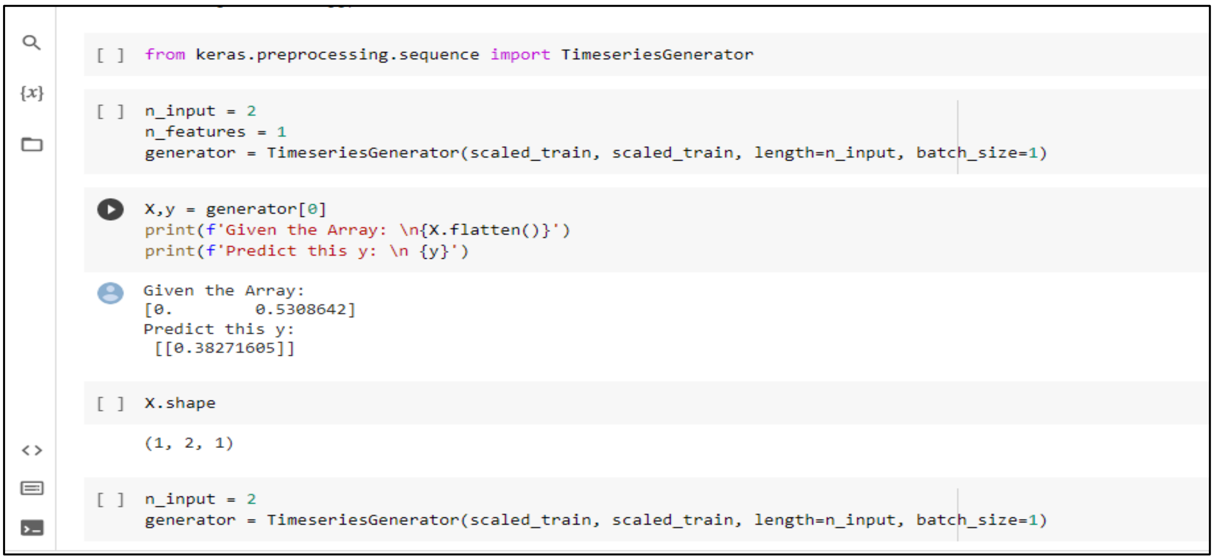
```
df.head(),df.tail()
```

```
(              Attrition
 Month
 2021-04-01           56
 2021-05-01           99
 2021-06-01           87
 2021-07-01          100
 2021-08-01          106,
              Attrition
 Month
 2021-11-01          125
 2021-12-01          125
 2022-01-01          137
 2022-02-01          116
 2022-03-01          117)
```

```
[ ]  scaler.fit(train)
     scaled_train = scaler.transform(train)
     scaled_test = scaler.transform(test)
```

Figure No. 9.10 MinMax Scaler in LSTM

LSTM learns using “TimeSeriesGenerator” function with input as 2, meaning it will count the rows in which the input number appears, study the pattern, and then forecast the following number in the series. This step is shown in Figure No. 9.11



```
from keras.preprocessing.sequence import TimeseriesGenerator
```

```
n_input = 2
n_features = 1
generator = TimeseriesGenerator(scaled_train, scaled_train, length=n_input, batch_size=1)
```

```
X,y = generator[0]
print(f'Given the Array: \n{X.flatten()}')
print(f'Predict this y: \n {y}')
```

```
Given the Array:
[0.        0.5308642]
Predict this y:
 [[0.38271605]]
```

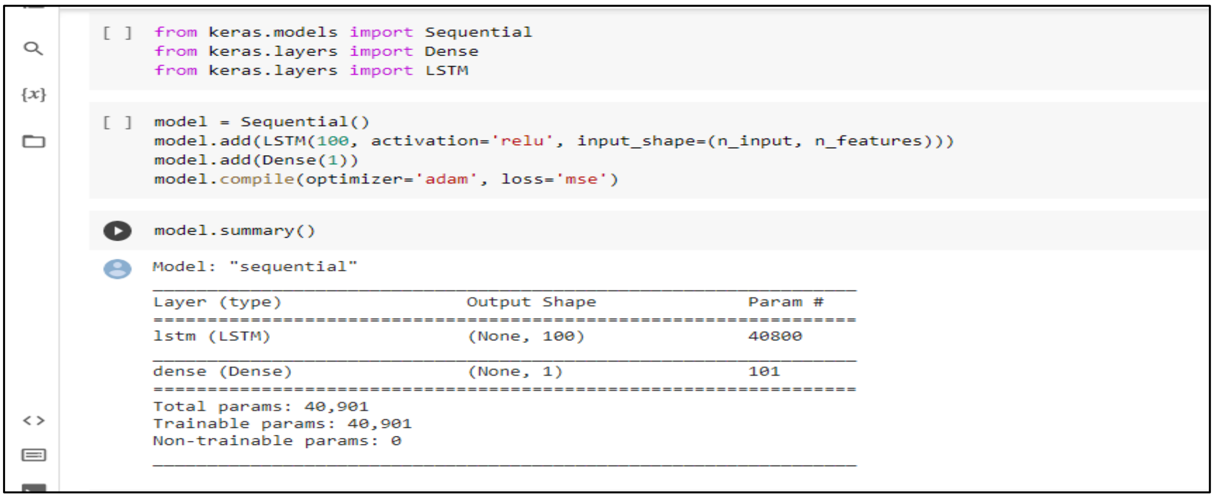
```
X.shape
```

```
(1, 2, 1)
```

```
n_input = 2
generator = TimeseriesGenerator(scaled_train, scaled_train, length=n_input, batch_size=1)
```

Figure No. 9.11 – TimeSeriesGenerator with input as 2

Following that, the Sequential, Dense, and LSTM classes are called from the Keras library. The Sequential class facilitates the layers' sequential addition. With activation function "relu,”, the optimizer is set to "adam" and the loss function "mse" and the model is compiled as shown in Figure No. 9.12.



```
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM
```

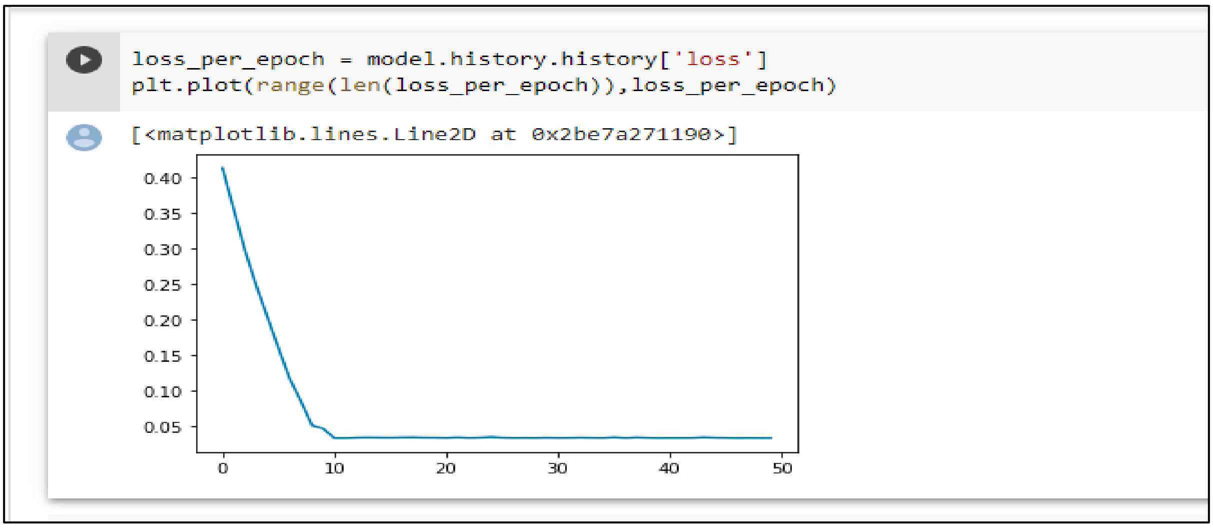
```
model = Sequential()
model.add(LSTM(100, activation='relu', input_shape=(n_input, n_features)))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')
```

```
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 100)               40800
_________________________________________________________________
dense (Dense)                (None, 1)                 101
=================================================================
Total params: 40,901
Trainable params: 40,901
Non-trainable params: 0
_________________________________________________________________
```

Figure No. 9.12 – Model Summary with activation = relu and optimizer = adam

The model is the run 50 epoch to understand the stage at which the loss is lowest for it to learn on its own. Figure No. 9.13 shows that post epoch 10, there is no further change.

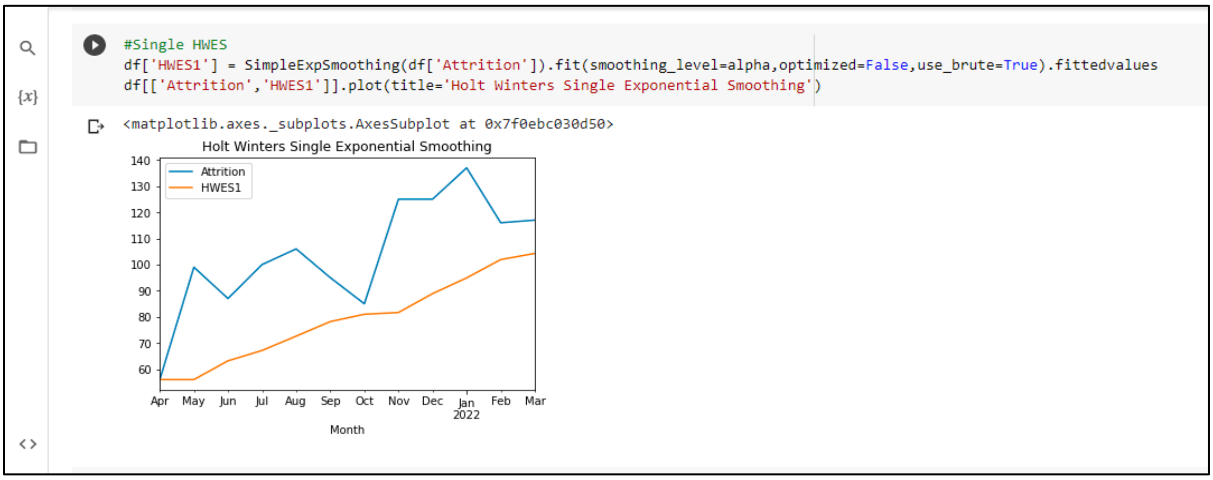


```
loss_per_epoch = model.history.history['loss']
plt.plot(range(len(loss_per_epoch)),loss_per_epoch)
```

```
[<matplotlib.lines.Line2D at 0x2be7a271190>]
```

Figure No. 9.13 – Loss Per Epoch

Finally, Holt-Winters smoothing technique is used to see if the forecast can be improved. It uses modified version of exponential smoothing to account for a linear trend. Exponential smoothing is the process of "smoothing" a time series using an exponentially weighted moving average (EWMA). Simple smoothing used as shown in Figure No. 9.14. Clearly, the result is not as expected.

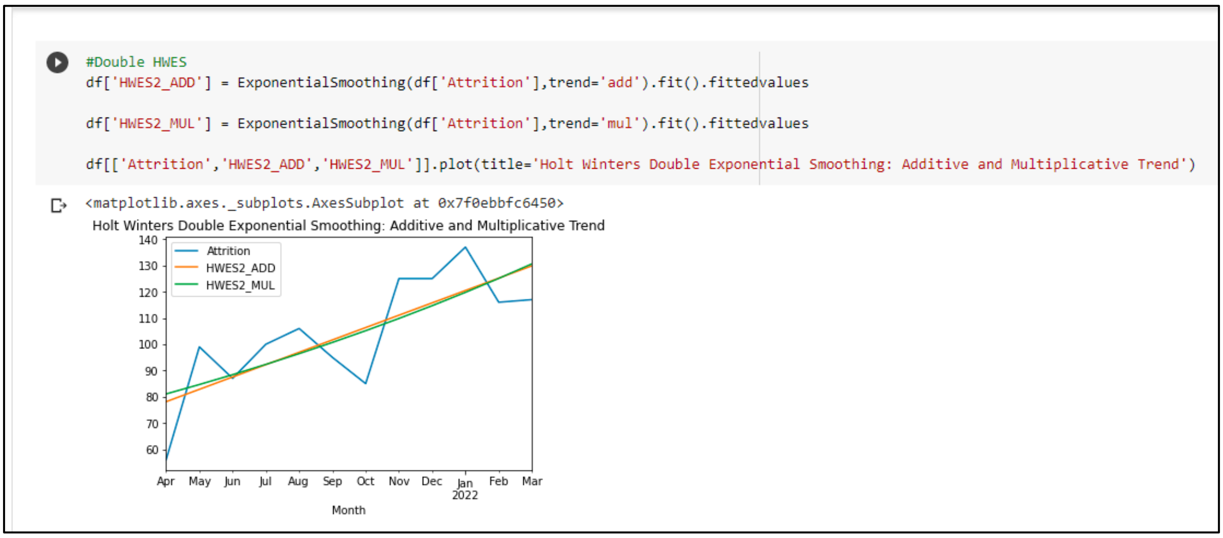


```
#Single HWES
df['HWES1'] = SimpleExpSmoothing(df['Attrition']).fit(smoothing_level=alpha,optimized=False,use_brute=True).fittedvalues
df[['Attrition','HWES1']].plot(title='Holt Winters Single Exponential Smoothing')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0ebc030d50>
```

Figure No. 9.14 – Holt-Winters Smoothing 1

The Holt-Winter Exponential Double Smoothing is also explored to see if the forecast can be improved further. This is depicted in Figure No. 9.15. Again, the results improved but not as expected.

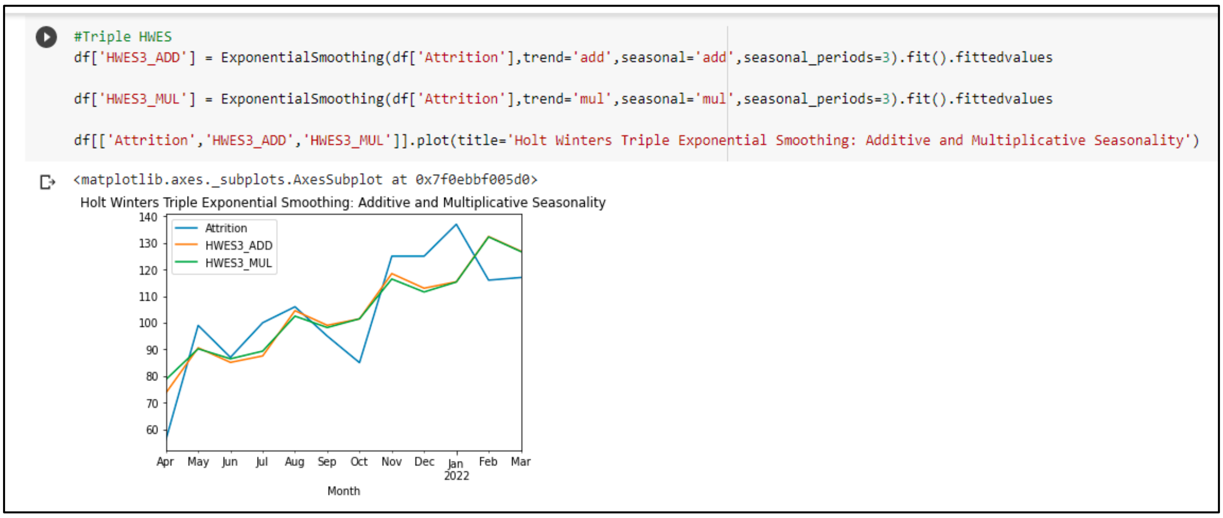


```
#Double HWES
df['HWES2_ADD'] = ExponentialSmoothing(df['Attrition'],trend='add').fit().fittedvalues

df['HWES2_MUL'] = ExponentialSmoothing(df['Attrition'],trend='mul').fit().fittedvalues

df[['Attrition','HWES2_ADD','HWES2_MUL']].plot(title='Holt Winters Double Exponential Smoothing: Additive and Multiplicative Trend')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0ebbfc6450>
```

Figure No. 9.15 – Holt-Winters Smoothing 2

The last Holt-Winter Exponential Smoothing technique used is Triple Smoothing to see if the forecast can be improved upon Excel's FORECAST.ETS function as shown in Figure No. 9.16. As seen later in the analysis section, the best results came from either a simple moving average or Excel's ETS function.



```
#Triple HWES
df['HWES3_ADD'] = ExponentialSmoothing(df['Attrition'],trend='add',seasonal='add',seasonal_periods=3).fit().fittedvalues

df['HWES3_MUL'] = ExponentialSmoothing(df['Attrition'],trend='mul',seasonal='mul',seasonal_periods=3).fit().fittedvalues

df[['Attrition','HWES3_ADD','HWES3_MUL']].plot(title='Holt Winters Triple Exponential Smoothing: Additive and Multiplicative Seasonality')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0ebbf005d0>
```

Figure No. 9.16 – Holt-Winters Smoothing 3

One final modeling technique that's explored is the regression technique to predict the probable missed SLAs based on the attrition data at the contract level, the result of which is shown in Table No. 9.1, and penalties based on the missed SLAs, refer Table No. 9.2. This is conducted based on the fact that attrition & missed SLAs have a high correlation value 0.93. This is consistent when a correlation is considered between missed SLAs and penalties paid which is 0.92. Again, these findings are obvious based on the nature of business.

Regression Statistics (Y = Missed SLAs, X = Attrition)	
Multiple R	0.93
R Square	0.86
Adjusted R Square	0.86
Standard Error	2.75

Table No. 9.1 Regression Statistics (Y = Missed SLAs, X = Attrition)

Regression Statistics (Y = Penalties, X = Missed SLAs)	
Multiple R	0.92
R Square	0.85
Adjusted R Square	0.85
Standard Error	23.18

Table No. 9.2 Regression Statistics (Y = Penalties, X = Missed SLAs)

To summarize, all the above forecasting techniques are explored to ensure that the most accurate forecasting model is found and deployed. This approach also helped in ascertaining the fact that even on a simple dataset, a simple timeseries forecasting model can worm wonders.

Chapter 10: Model Evaluation

Once all the forecast techniques are used, a summary showing the model performance is presented below. When overall data attrition is considered, the best forecast technique is Moving Average for which the MAPE is 9 % as shown in Table No. 10.1

Since the MAPE is on a bit higher side, LSTM technique is tried to see if the performance can be bettered. However, LSTM gave an accuracy of 86% which is lower than the Moving Average technique.

TS Models	MAD	RMSE	MAPE
Moving Average (3)	9.6	10.9	9%
Ratio to Moving Average	11.0	12.9	12%
Exponential Triple Smoothing (ETS)	13.2	14.0	10%
ARIMA	15.8	18.7	14%
Holt Winters ES1	27.5	30.5	25%
Holt Winters ES2_ADD	12.1	13.5	13%
Holt Winters ES2_MUL	18.9	25.0	18%
Holt Winters ES3_ADD	16.9	19.1	16%
Holt Winters ES3_MUL	21.3	24.3	20%

Table No. 10.1 Model Performance on overall attrition data

The graph in Figure No. 10.1 shows how well the forecasted values have performed against the original values.

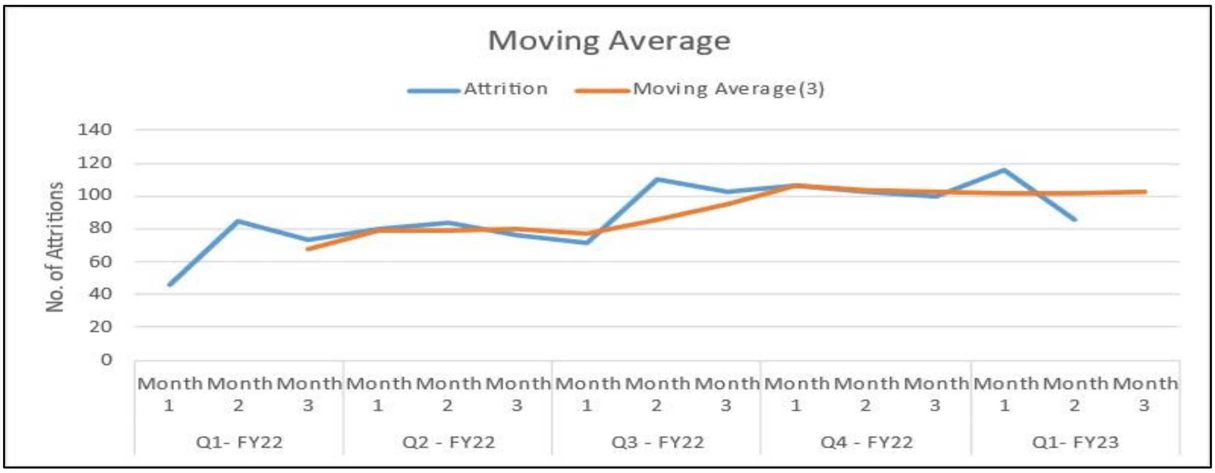


Figure No. 10.1 Moving Average performance on overall attrition

The same approach is used for the other data sets, the performance of which is given below in tables. The performances of the best models are then graphically represented to show forecasted values versus original values.

For the top six contracts data, the results are similar between Moving Average and ETS as appearing in the Table No. 10.2. Hence both the graphs, one for moving average, refer Figure No. 10.2, and the other for ETS shown in Figure No. 10.3 are presented below for a quick comparison.

TS Models	MAD	RMSE	MAPE
Moving Average (3)	5.8	8.8	6%
Ratio to Moving Average	8.5	9.5	11%
Exponential Triple Smoothing (ETS)	6.4	6.9	6%
ARIMA	13.3	15.9	14%
Holt Winters ES1	23.4	25.7	26%
Holt Winters ES2_ADD	9.1	11.1	12%
Holt Winters ES2_MUL	19.0	24.4	25%
Holt Winters ES3_ADD	12.1	13.2	14%
Holt Winters ES3_MUL	14.3	16.2	17%

Table No. 10.2 Model Performance on Top 6 contracts

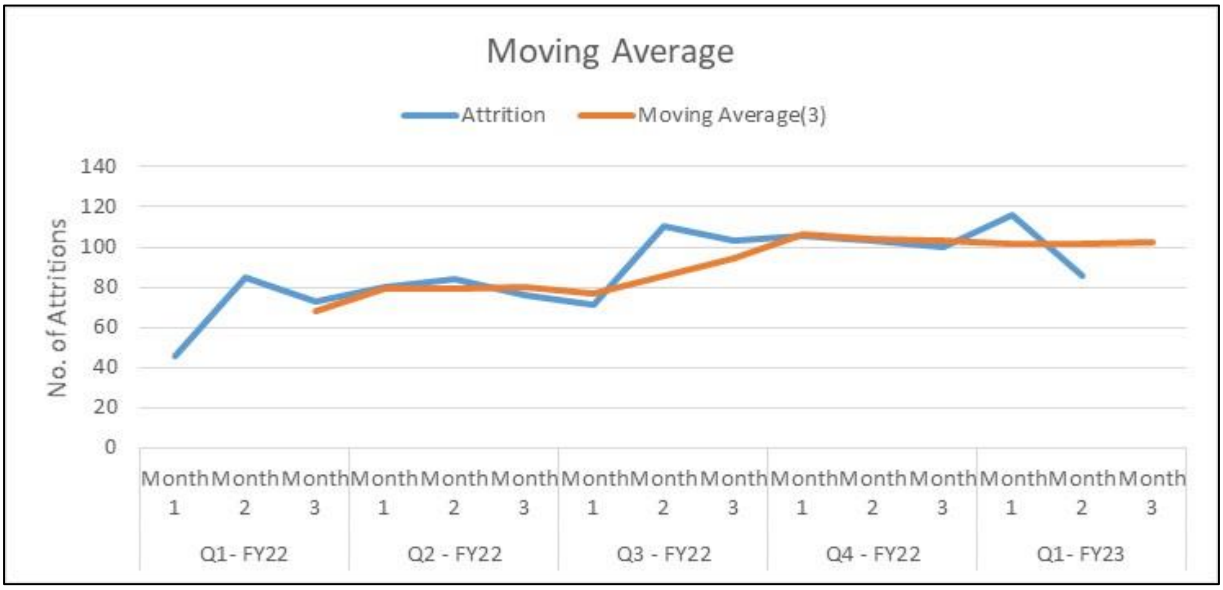


Figure No. 10.2 Moving Average performance on Top 6 contracts

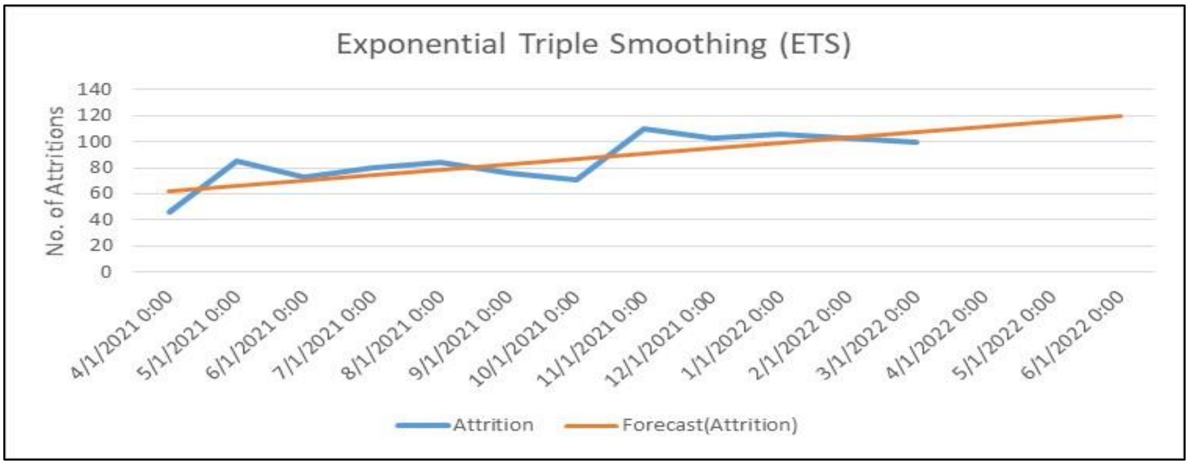


Figure No. 10.3 ETS performance on Top 6 contracts

For C01 employee-level data, ETS is the choice with a MAPE score of 4% appearing in Table No. 10.3. Figure No. 10.4 shows the graphical representation of the ETS model.

TS Models	MAD	RMSE	MAPE
Moving Average (3)	6.3	9.8	7%
Ratio to Moving Average	9.0	10.6	13%
Exponential Triple Smoothing (ETS)	4.4	5.0	4%
ARIMA	13.5	15.4	15%
Holt Winters ES1	24.0	26.5	28%
Holt Winters ES2_ADD	9.3	11.5	13%
Holt Winters ES2_MUL	10.1	12.9	15%
Holt Winters ES3_ADD	14.7	16.8	19%
Holt Winters ES3_MUL	18.9	22.9	24%

Table No. 10.3 Model Performance on C01 employment level data

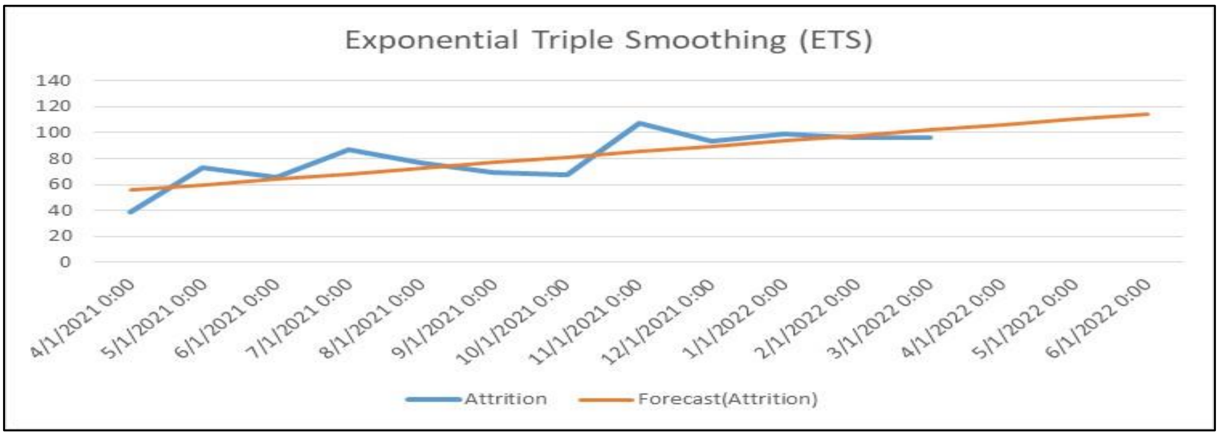


Figure No. 10.4 ETS performance on C01 Employment level

A similar result is achieved for the Top 6 contracts at the C01 employee level with ETS having the least MAPE score of 4% as shown in Table No. 10.4. This is as expected from the earlier results. ETS performance on the Top 6 contracts at C01 level data is shown in Figure No. 10.5.

TS Models	MAD	RMSE	MAPE
Moving Average (3)	5.4	8.7	7%
Ratio to Moving Average	7.9	8.5	12%
Exponential Triple Smoothing (ETS)	3.6	4.2	4%
ARIMA	15.4	17.5	19%
Holt Winters ES1	22.8	24.6	29%
Holt Winters ES2_ADD	7.8	9.6	12%
Holt Winters ES2_MUL	16.7	20.4	25%
Holt Winters ES3_ADD	11.2	12.6	16%
Holt Winters ES3_MUL	14.0	16.2	19%

Table No. 10.4 Model Performance on Top 6 contracts at C01 employment level

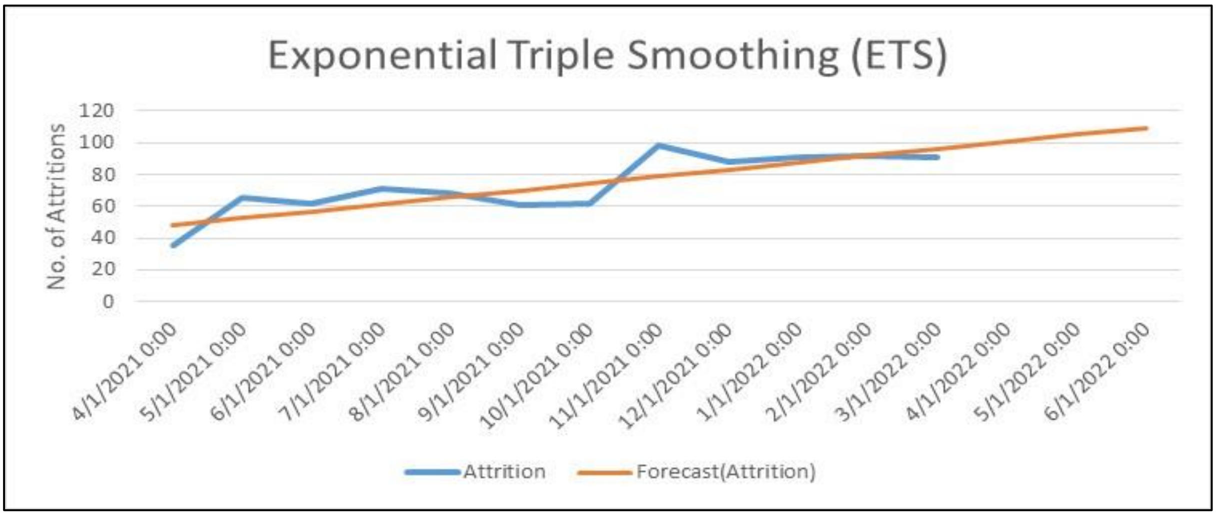


Figure No. 10.5 ETS performance on Top 6 contracts at C01 employee level

For the Top Salary bucket, once again ETS gave an outstanding result with a MAPE score of 2% as appears in Table No. 10.5. Figure No. 10.6 shows the model performance on the actual data.

TS Models	MAD	RMSE	MAPE
Moving Average (3)	6.5	9.2	7%
Ratio to Moving Average	7.2	8.7	9%
Exponential Triple Smoothing (ETS)	2.2	3.4	2%
ARIMA	12.3	14.2	13%
Holt Winters ES1	22.7	24.9	25%
Holt Winters ES2_ADD	14.4	19.2	19%
Holt Winters ES2_MUL	17.5	22.0	23%
Holt Winters ES3_ADD	12.3	13.6	15%
Holt Winters ES3_MUL	15.6	18.0	18%

Table No. 10.5 Model Performance on Top Salary Bucket

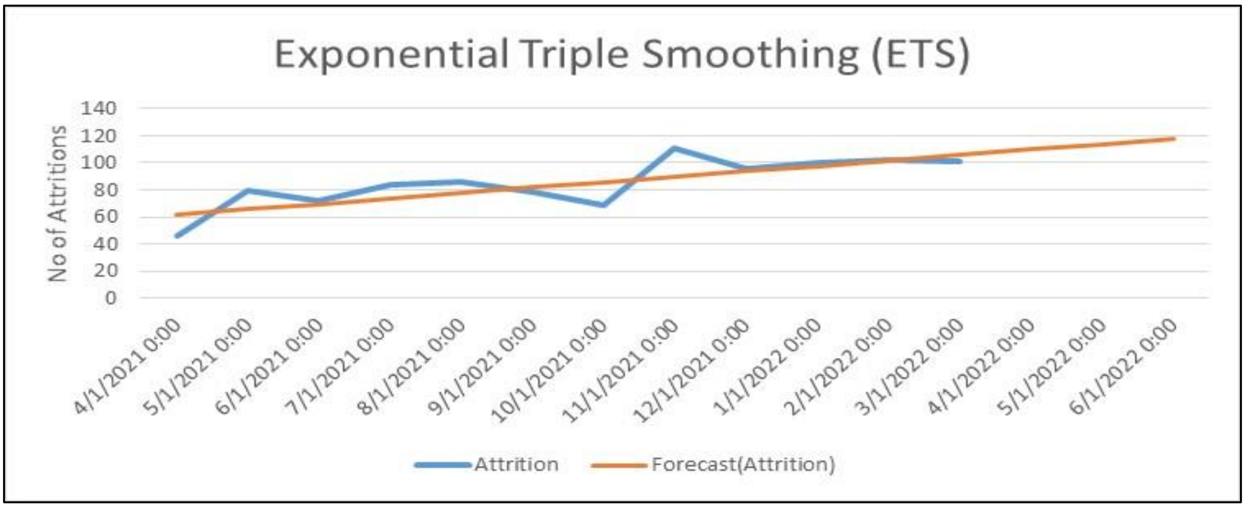


Figure No. 10.6 ETS performance on Top Salary Bucket

For Top cities, ETS is once again the choice with a MAPE score of 4% closely followed by Moving Average MAPE score of 6% appearing in Table No. 10.6.

TS Models	MAD	RMSE	MAPE
Moving Average (3)	5.4	8.3	6%
Ratio to Moving Average	8.9	9.9	12%
Exponential Triple Smoothing (ETS)	4.6	5.7	4%
ARIMA	14.0	16.4	15%
Holt Winters ES1	25.9	27.9	29%
Holt Winters ES2_ADD	9.1	11.3	13%
Holt Winters ES2_MUL	19.7	24.5	27%
Holt Winters ES3_ADD	12.5	14.1	15%
Holt Winters ES3_MUL	15.2	17.8	18%

Table No. 10.6 Model Performance on Top Cities

ETS result versus actual forecast result is shown in Figure No. 10.7.

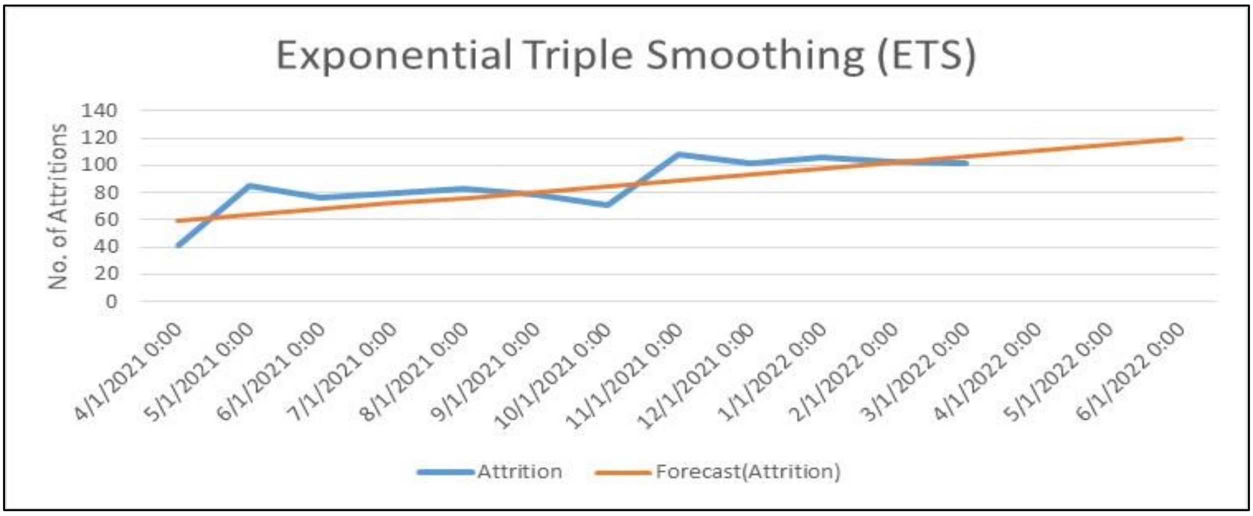


Figure No. 10.7 ETS on Top Cities

Chapter 11: Deployment

The models used here need to be tested with future attrition data to establish the consistency of results. Since other external factors are not considered which can affect attrition, this study would be an ongoing activity. However, the final findings will be shared to gain overall feedback from the management. Based on the feedback a deployment process can be decided.

Chapter 12: Analysis and Results

The forecasting techniques were tested on the actual attrition data of the following two months. When the overall data is used, the Moving Average model is giving MAPE as 17% (Table No. 12.1) compared with the result received at the time of modeling which is 9%.

Overall Data						
Quarter	**Month**	**Attrition - Actual**	**Moving Average (Forecast)**	**MAD**	**RMSE**	**MAPE**
Q1- FY23	Month 1	132	118.8	13.2	174.8	10%
	Month 2	158	119.7	38.3	1466.6	24%
	Month 3		120.6			
				25.8	**28.6**	**17%**

Table No. 12.1 Moving Average Model Outcome for Overall dataset

Since there is a difference of 8% on actual versus model, the ETS is used to compare the results with the Moving Average. The ETS shows better performance on the actual numbers as shown in Table No. 12.2.

Overall Data						
Quarter	**Month**	**Attrition - Actual**	**ETS (Forecast)**	**MAD**	**RMSE**	**MAPE**
Q1- FY23	Month 1	132	128.7	3.3	11.0	3%
	Month 2	158	133.4	24.6	604.5	16%
	Month 3		138.2			
				14.0	**17.5**	**9%**

Table No. 12.2 ETS Model Outcome for Overall dataset

For top contracts, both Moving Average & ETS can be used as their results are almost similar with MAPE for ETS is 6% whereas for Moving Average it is 5%. The results are shown in Table No. 12.3 & Table No. 12.4 below.

Top 6 Contracts						
Quarter	**Month**	**Attrition - Actual**	**MA (Forecast)**	**MAD**	**RMSE**	**MAPE**
Q1- FY23	Month 1	116	103.8	12.2	147.7	10%
	Month 2	126	125.6	0.4	0.2	0%
	Month 3		111.4			
				6.3	**8.6**	**5%**

Table No. 12.3 Moving Average (MA) Model Outcome for Top 6 Contracts

Top 6 Contracts						
Quarter	**Month**	**Attrition - Actual**	**ETS (Forecast)**	**MAD**	**RMSE**	**MAPE**
Q1- FY23	Month 1	116	111.9	4.1	17.1	4%
	Month 2	126	116.0	10.0	99.4	8%
	Month 3		120.2			
				7.1	**7.6**	**6%**

Table No. 12.4 ETS Model Outcome for Top 6 Contracts

For the rest of the datasets, the ETS model result has been considered since it’s able to give the best & consistent results across all datasets. The below tables depict this result. Table No. 12.5 shows the ETS performance on the C01 employment level.

C01 Employee Level						
Quarter	**Month**	**Attrition - Actual**	**ETS (Forecast)**	**MAD**	**RMSE**	**MAPE**
Q1- FY23	Month 1	109	106.2	2.8	7.9	3%
	Month 2	123	110.4	12.6	158.9	10%
	Month 3		114.6			
				7.7	**9.1**	**6%**

Table No. 12.5 ETS Model Outcome for C01 Employee Level

Table No. 12.6 shows the ETS result for Top Contracts.

Top 6 Contracts with C01						
Quarter	**Month**	**Attrition - Actual**	**ETS (Forecast)**	**MAD**	**RMSE**	**MAPE**
Q1- FY23	Month 1	108	100.8	7.2	51.8	7%
	Month 2	113	105.2	7.8	61.1	7%
	Month 3		109.6			
				7.5	**7.5**	**7%**

Table No. 12.6 ETS Model Outcome for Top 6 Contracts at C01 Employee Level

For the top salary bucket, the ETS results are shown in Table No. 12.7.

Top Salary Bucket						
Quarter	**Month**	**Attrition - Actual**	**ETS (Forecast)**	**MAD**	**MSE**	**MAPE**
Q1- FY23	Month 1	113	109.9	3.1	9.7	3%
	Month 2	120	113.9	6.1	37.0	5%
	Month 3		117.9			
				4.6	**23.3**	**4%**

Table No. 12.7 ETS Model Outcome for Top Salary Bucket

Table No. 12.8 show the ETS results for Top cities.

Top Cities						
Quarter	**Month**	**Attrition - Actual**	**ETS (Forecast)**	**MAD**	**MSE**	**MAPE**
Q1- FY23	Month 1	112	110.7	1.3	1.7	1%
	Month 2	122	115.0	7.0	48.9	6%
	Month 3		119.3			
				4.1	**25.3**	**3%**

Table No. 12.8 ETS Model Outcome for Top Cities

From the above findings, Exponential Triple Smoothing (ETS) is the model that can be confidently used for all the various datasets considered for this study. Based on the ETS forecasted value, the number of missed SLAs and subsequent penalties can be predicted using the formula as shown below in Table No. 12.9.

Regression Formulae
Predicted Missed SLAs = (0.07958 * Forecasted Attrition) + 3.7954
Predicted Penalties = (7.45942 * Predicted Missed SLAs) + 0.20933

Table No. 12.9 Regression Formulae

Using the regression model, predicted missed SLAs and penalties on overall attrition data are calculated as shown in Table No. 12.10. The same concept can be used for all other 5 datasets.

Overall Data			
Month	**ETS (Forecast)**	**Predicted Missed SLAs**	**Predicted Penalties in ($ k)**
Month 1	128.7	14	104.9
Month 2	133.4	14	107.7
Month 3	138.2	15	110.6
Total		**43**	**323.2**

Table No. 12.10 Predicted Missed SLAs and Penalties

To conclude, if attrition is forecasted well in advance, proper workload management can be put in place which in turn can help understand the approximate amount that can be saved in form of penalties.

Chapter 13: Conclusions and Recommendations for future work

Attrition remains a burning problem in any sector, and it can have profound consequences when it is way above a tolerable limit.

A lot of studies has been conducted to tackle this problem, however, sometimes a simple solution can be effective when there is limited information available.As seen from this study, timeseries forecasting technique is used to predict future attritions across several datasets. Since this is one of its kind approaches in terms of forecasting attrition, several forecasting techniques have been used on several datasets to ascertain the efficacy of this approach.

From this study, it has been established that given only attrition data, future attrition can be predicted with greater accuracy. “FORECAST.ETS”, an inbuilt function in Microsoft Excel has been the star forecasting technique across all datasets.

To conclude, sometimes a seemingly tough problem can be tackled through simple approaches, in this case, attrition forecasting using timeseries techniques. This will help plan future workload effectively, reduce missed SLAs and penalties.

However, this approach is suitable only when the data dimension is less or there are no specialized teams performing the task. In an ideal scenario, there can be several factors that may affect a company’s attrition, but with limited data, this approach is a way out as it uses monthly attrition data to forecast probable attritions for the next 3 months.

For the future, the scope is tremendous. As with any data modeling technique, this study must be carried out on future datasets to ascertain its validity. External factors, which are not part of this study, can be considered in future works.

This study would pave way for a novel approach in attrition forecasting as this technique is rarely being used in today’s AI / ML age.

Bibliography

Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Stock Price Prediction Using the ARIMA Model. *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*. https://doi.org/10.1109/UKSim.2014.67

Alsheref, F. K., Fattoh, I. E., & M.Ead, W. (2022). Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms. *Computational Intelligence and Neuroscience*, *2022*, 1–9. https://doi.org/10.1155/2022/7728668

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1976). *Time Series Analysis Forecasting and Gontrol FOURTH EDITION*.

Chakraborty, R., Mridha, K., Nath Shaw, R., & Ghosh, A. (2021). Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches; Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches. *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*. https://doi.org/10.1109/GUCON50781.2021.9573759

Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons and Fractals*, *135*. https://doi.org/10.1016/j.chaos.2020.109864

CRISP-DM Help Overview - IBM Documentation. (2021). https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

Dimri, T., Ahmad, S., & Sharif, M. (2020). Time series analysis of climate variables using seasonal ARIMA approach. *Journal of Earth System Science*, *129*(1). https://doi.org/10.1007/s12040-020-01408-x

Fallucchi, F., Coladangelo, M., Giuliano, R., & de Luca, E. W. (2020). *Predicting Employee Attrition Using Machine Learning Techniques*. https://doi.org/10.3390/computers9040086

Hansun, S. (2013). A New Approach of Moving Average Method in Time Series Analysis. In *2013 Conference on New Media Studies (CoNMedia)*. https://doi.org/10.1109/CoNMedia.2013.6708545

Jain, R., & Nayyar, A. (2018). Predicting Employee Attrition using XGBoost Machine Learning Approach; Predicting Employee Attrition using XGBoost Machine Learning Approach. In *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*.

Kalekar, P. S. (2004). *Time series Forecasting using Holt-Winters Exponential Smoothing*.

Kumar Jain, P., Jain, M., & Pamula, R. (123 C.E.). *Explaining and predicting employees' attrition: a machine learning approach*. https://doi.org/10.1007/s42452-020-2519-4

Mahalakshmi, G., Sridevi, S., & Rajaram, S. (2016). *A survey on forecasting of time series data; A survey on forecasting of time series data*. https://doi.org/10.1109/ICCTIDE.2016.7725358

Ner, W. (2020). *THE OFFICIAL PUBLICATION OF TRAINING MAGAZINE NETWORK Training Temperature Check*. www.trainingmag.com

Rahardja, D. (2021). Statistical Time-Series Forecast via Microsoft Excel (FORECAST.ETS) Built-In Function. In *Quest Journals Journal of Research in Applied Mathematics* (Vol. 7, Issue 11). www.questjournals.org

Sailaja, M., & Prasad, A. R. (2019). Identification of Seasonal Effects through Ratio to Moving Average Method for the Number of Train Passengers and Income of South Central Railway Zone. *International Journal of Mathematics Trends and Technology*, *65*, 11. http://www.ijmttjournal.org

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2019). *Financial Time Series Forecasting with Deep Learning : A Systematic Literature Review: 2005-2019*. http://arxiv.org/abs/1911.13288

Singh Sisodia, D., Vishwakarma, S., & Pujahari, A. (2017). *Evaluation of Machine Learning Models for Employee Churn Prediction*.

Smither, L. (2003). *Managing Employee Life Cycles To Improve Labor Retention*. www.thomas-staffing.com/survey99/retention_Table No.2.htm

Taylor, S. J., & Letham, B. (2017). *Forecasting at Scale*. https://doi.org/10.7287/peerj.preprints.3190v2

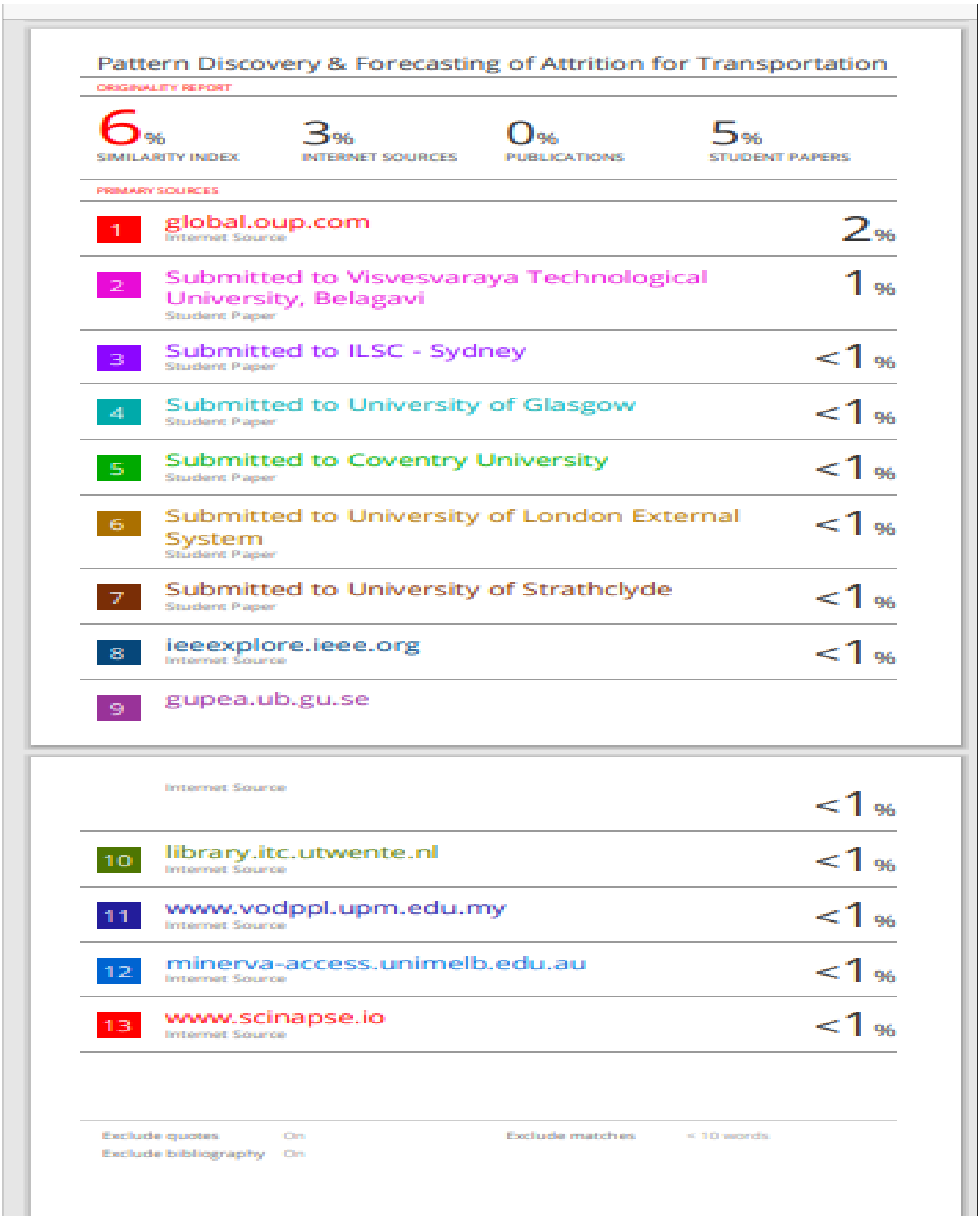
Time Series Forecasting: Definition & Examples | Table No.au. (2020). https://www.Table No.au.com/learn/articles/time-series-forecasting

Tseng, F.-M., & Tzeng, G.-H. (2002). A fuzzy seasonal ARIMA model for forecasting. In *Fuzzy Sets and Systems* (Vol. 126). www.elsevier.com/locate/fss

Ueno, R., Calitoiu, D., & Calitoiu@forces, D. (2020). *Forecasting Attrition from the Canadian Armed Forces using Multivariate LSTM;* https://doi.org/10.1109/ICMLA51294.2020.00123

Appendix

Plagiarism Report[1]



Pattern Discovery & Forecasting of Attrition for Transportation

ORIGINALITY REPORT

6%	3%	0%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	global.oup.com Internet Source	2%
2	Submitted to Visvesvaraya Technological University, Belagavi Student Paper	1%
3	Submitted to ILSC - Sydney Student Paper	<1%
4	Submitted to University of Glasgow Student Paper	<1%
5	Submitted to Coventry University Student Paper	<1%
6	Submitted to University of London External System Student Paper	<1%
7	Submitted to University of Strathclyde Student Paper	<1%
8	ieeexplore.ieee.org Internet Source	<1%
9	gupea.ub.gu.se Internet Source	<1%
10	library.itc.utwente.nl Internet Source	<1%
11	www.vodppl.upm.edu.my Internet Source	<1%
12	minerva-access.unimelb.edu.au Internet Source	<1%
13	www.scinapse.io Internet Source	<1%

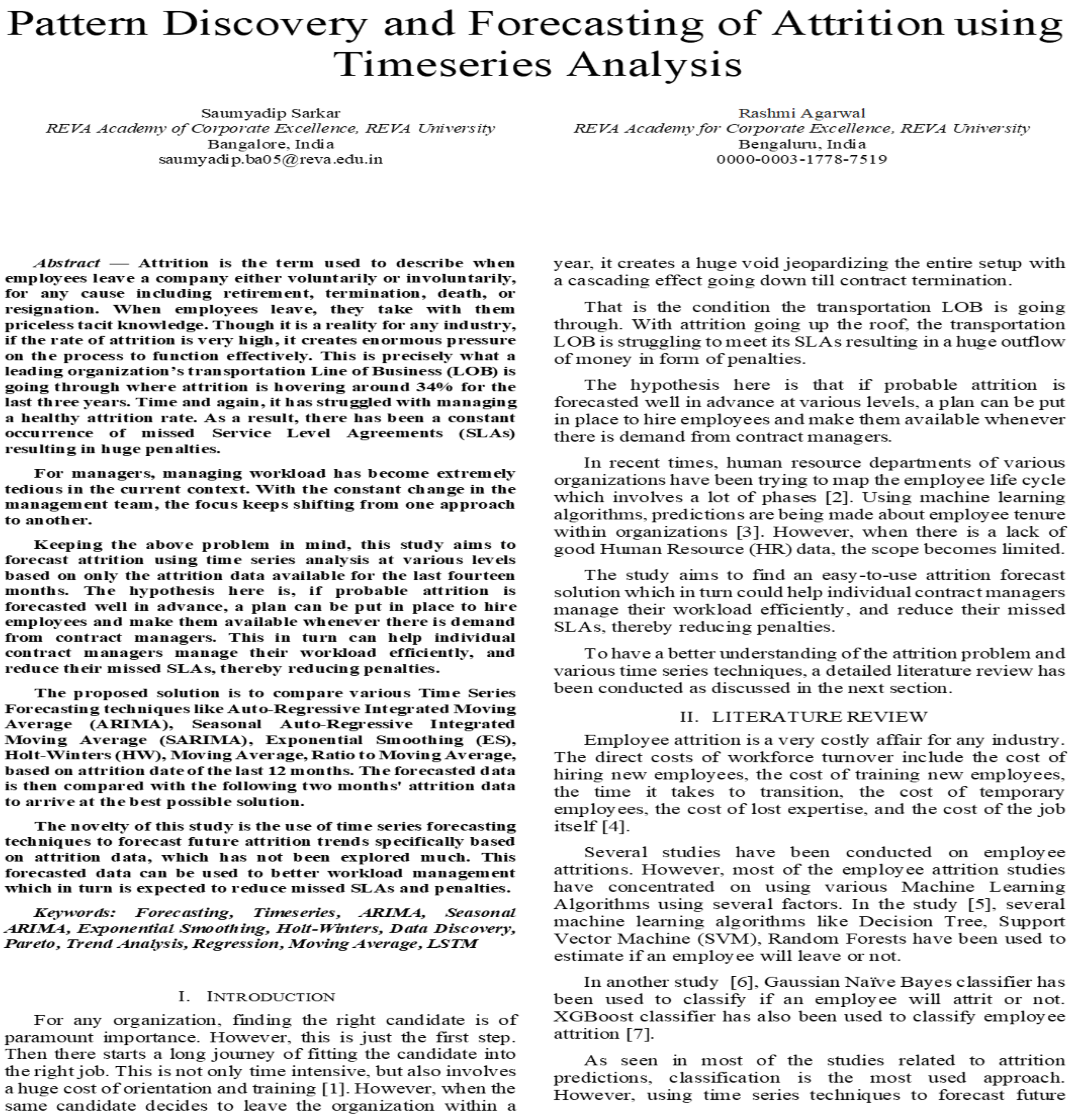
Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

[1] Turnitn report to be attached from the University.

Publications in a Journal/Conference Presented/White Paper[2]

Paper Sumbitted:

Saumyadip Sarkar, *Rashmi Agarwal*, "**Pattern Discovery and Forecasting of Attrition using Timeseries Analysis**", i4c2022 (4th Internal Conference on Circuits, Control, Communication and Computing (I4C) – 2022, Submisson Date: 13th October, 2022.



Pattern Discovery and Forecasting of Attrition using Timeseries Analysis

Saumyadip Sarkar
REVA Academy of Corporate Excellence, REVA University
Bangalore, India
saumyadip.ba05@reva.edu.in

Rashmi Agarwal
REVA Academy for Corporate Excellence, REVA University
Bengaluru, India
0000-0003-1778-7519

***Abstract* — Attrition is the term used to describe when employees leave a company either voluntarily or involuntarily, for any cause including retirement, termination, death, or resignation. When employees leave, they take with them priceless tacit knowledge. Though it is a reality for any industry, if the rate of attrition is very high, it creates enormous pressure on the process to function effectively. This is precisely what a leading organization's transportation Line of Business (LOB) is going through where attrition is hovering around 34% for the last three years. Time and again, it has struggled with managing a healthy attrition rate. As a result, there has been a constant occurrence of missed Service Level Agreements (SLAs) resulting in huge penalties.**

For managers, managing workload has become extremely tedious in the current context. With the constant change in the management team, the focus keeps shifting from one approach to another.

Keeping the above problem in mind, this study aims to forecast attrition using time series analysis at various levels based on only the attrition data available for the last fourteen months. The hypothesis here is, if probable attrition is forecasted well in advance, a plan can be put in place to hire employees and make them available whenever there is demand from contract managers. This in turn can help individual contract managers manage their workload efficiently, and reduce their missed SLAs, thereby reducing penalties.

The proposed solution is to compare various Time Series Forecasting techniques like Auto-Regressive Integrated Moving Average (ARIMA), Seasonal Auto-Regressive Integrated Moving Average (SARIMA), Exponential Smoothing (ES), Holt-Winters (HW), Moving Average, Ratio to Moving Average, based on attrition date of the last 12 months. The forecasted data is then compared with the following two months' attrition data to arrive at the best possible solution.

The novelty of this study is the use of time series forecasting techniques to forecast future attrition trends specifically based on attrition data, which has not been explored much. This forecasted data can be used to better workload management which in turn is expected to reduce missed SLAs and penalties.

Keywords: Forecasting, Timeseries, ARIMA, Seasonal ARIMA, Exponential Smoothing, Holt-Winters, Data Discovery, Pareto, Trend Analysis, Regression, Moving Average, LSTM

I. Introduction

For any organization, finding the right candidate is of paramount importance. However, this is just the first step. Then there starts a long journey of fitting the candidate into the right job. This is not only time intensive, but also involves a huge cost of orientation and training [1]. However, when the same candidate decides to leave the organization within a year, it creates a huge void jeopardizing the entire setup with a cascading effect going down till contract termination.

That is the condition the transportation LOB is going through. With attrition going up the roof, the transportation LOB is struggling to meet its SLAs resulting in a huge outflow of money in form of penalties.

The hypothesis here is that if probable attrition is forecasted well in advance at various levels, a plan can be put in place to hire employees and make them available whenever there is demand from contract managers.

In recent times, human resource departments of various organizations have been trying to map the employee life cycle which involves a lot of phases [2]. Using machine learning algorithms, predictions are being made about employee tenure within organizations [3]. However, when there is a lack of good Human Resource (HR) data, the scope becomes limited.

The study aims to find an easy-to-use attrition forecast solution which in turn could help individual contract managers manage their workload efficiently, and reduce their missed SLAs, thereby reducing penalties.

To have a better understanding of the attrition problem and various time series techniques, a detailed literature review has been conducted as discussed in the next section.

II. LITERATURE REVIEW

Employee attrition is a very costly affair for any industry. The direct costs of workforce turnover include the cost of hiring new employees, the cost of training new employees, the time it takes to transition, the cost of temporary employees, the cost of lost expertise, and the cost of the job itself [4].

Several studies have been conducted on employee attritions. However, most of the employee attrition studies have concentrated on using various Machine Learning Algorithms using several factors. In the study [5], several machine learning algorithms like Decision Tree, Support Vector Machine (SVM), Random Forests have been used to estimate if an employee will leave or not.

In another study [6], Gaussian Naïve Bayes classifier has been used to classify if an employee will attrit or not. XGBoost classifier has also been used to classify employee attrition [7].

As seen in most of the studies related to attrition predictions, classification is the most used approach. However, using time series techniques to forecast future

[2] URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.

attrition has not been explored enough based on the observations during the literature review. This establishes a unique opportunity for this study.

Generating scientific projections based on data with historical time stamps is known as time series forecasting. It entails creating models through historical study, using them to draw conclusions and guide strategic decision-making in the future [8].

Timeseries analysis and forecasting are important for a variety of applications, including business, the stock market and exchange, the weather, electricity demand, cost, and usages of goods like fuels and electricity, etc., and in any setting where there are periodic, seasonal variations [9].

There are several time series techniques available, notably Moving Average, Exponential Smoothing, Holt-Winters Smoothing method, ARIMA, Seasonal ARIMA, LSTM, etc.

One well-known technical indicator, the moving average, is employed in time series analysis to forecast future data. Researchers have produced numerous variations and implementations of it during its evolution [10]. Another variation of the moving average is the Ratio to Moving Average, which is superior to the simple average method and is predicated on the idea that seasonal variance for any given month is a continuous component of the trend. Moving average methods reduce periodic movements if any [11].

Another immensely popular time series technique is Exponential Smoothing. Its popularity is based on the fact that surprisingly accurate forecasts can be obtained with minimal effort. This has been proved in this study as well where time series forecasting methods are used via an Excel (FORECAST.ETS) function. The superior efficacy of this model has been nicely illustrated in the paper by Dewi Rahardja. With this Excel function, forecasting is simple and quick while considering the model's level (intercept), trend (slope), and seasonality [12].

Another variation of the Exponential Smoothing technique popularly known as Holt-Winters after the name of the inventors is very effective when there is both trend and seasonality in the data. The two primary HW models are multiplicative for time series exhibiting multiplicative seasonality and additive for time series exhibiting additive seasonality [13].

However, there are times when, for the same number of data, a Long Short-Term Memory (LSTM) multivariate machine learning model outperforms a Holt-Winters univariate statistical model [14]. By utilizing the nonlinearities of a particular dataset, LSTM networks can overcome the constraints of conventional time series forecasting methods and produce state-of-the-art outcomes on temporal data [15].

From a statistical modeling perspective, another time series technique that produces a robust result in short-term prediction is ARIMA, first introduced by Box and Jenkins in 1970. It consists of a series of steps for locating, calculating, and diagnosing ARIMA models using time series data. A very well-researched paper available in this context is [16], which shows ARIMA’s strength in predicting future stock prices. The limitation of ARIMA model is however the number of periods. It is recommended to employ at least 50 and ideally 100 observations [17].

While ARIMA has its own strength, when it comes to seasonal data, there is a variation of ARIMA available commonly known as SARIMA or seasonal ARIMA. For climate-related data, SARIMA has been a valuable tool [18].

With a lot of statistical techniques already being widely used in forecasting techniques, recent studies have now been conducted using Deep Learning (DL) models showing outstanding results when compared to traditional forecasting techniques. A comparative study [19] has shown that the DL method significantly improves upon Machine Learning (ML) models.

I. Problem statement

It has been observed that, in the last 14 months, close to $2,082k has been paid in terms of penalties in transportation LOB. Penalties are incurred when SLAs are not being met.

One of the main reasons identified is “attrition” which presently stands at 34%. This has almost remained constant for the last three years.

With constant churn at top management and lack of HR support, there seems almost no headway in managing attritions. Ironically, the lack of HR support is due to the fact that HR department has failed to manage its own attrition rate.

The problem is further exacerbated due to the lack of proper data collection at the HR end. With the limited data and considering the present problem area, this study aims to do the following:

1. Identify the required attrition dataset. In this case, only the attrition data for the last 14 months are available.

2. Study data at various levels and categories to identify trends, patterns, and top contributors which then will be used to create subsets of the main dataset for modeling.

3. Explore various time series techniques to identify the best time series forecasting model which can be used to forecast future attrition.

II. Methodology

This study uses Cross-Industry Standard Process for Data Mining (CRSIP-DM) framework which is discussed below.

The first phase is to understand the business in context. For this study, the transportation LOB is considered which is reeling under huge attrition for the last 3 years.

The next phase in CRISP-DM process is data understanding. Here, the data is the employee attrition record captured for the last 14 months. The other data points are missed SLA numbers at contract levels and penalties paid at contract levels for the last 14 months.

The third phase involves data preparation. The goal here is to identify if the data is fit for time series analysis. This involves looking at trends to see if there are any strong upward or downward trends. Along with this, the data is further analysed for any seasonal trends. Based on the findings, data transformation can be done to make the data stationary.

The main approach in the modeling phase is to select the best time series forecasting technique like Moving Average, ARIMA, SARIMA, Exponential Smoothing, and Holt Winters.

Post modeling technique, the evaluation phase starts. In this phase, the efficacy of the various time series forecasting techniques is assessed. The one which most accurately mimics the test data would be finalized.

In the deployment phase, the forecasted numbers can be used to hire a pool of employees. These employees will then be suitably placed in various contracts based on the current need.

A. *Business Understanding*

The transportation LOB is one of the most profitable units in this organization. It provides the following solutions as depicted in TABLE 1 to its client.

TABLE 1. Transportation Solutions

Solutions	Description
Automated Tolling	Captures vehicle details when a tolling both is crossed and bills customer accordingly. A team also works on dispute resolutions pertaining to technical failure, failed auto-debit attempts, customer complaints, etc.
Automated Parking	It provided intelligent parking solutions mainly for governments. The solution involves fee collections, dynamic pricing, enforcement solutions, etc.
Public Safety	It provides automated photo enforcement, traffic violation solutions, etc.

However, the biggest challenge is the abnormal attrition rate for the last 3 years which on average is around 34 %. Managers are finding it difficult to manage their workload efficiently.

Apart from providing technological solutions, there are dedicated teams working on resolving disputes pertaining to technological failures, missed billing, payment disputes, failed payments, etc. There are agreed Turnaround Time (TAT) for resolving disputes which are part of SLAs with individual contracts.

There are also several SLAs linked to penalties. A missed SLA incurs penalties to be paid to the client. It has been observed that almost $ 2,082k penalties is paid in the last 14 months by the accounts reeling under huge attritions.

Since dispute resolution is manually intensive work, more employees are needed during peak times. But with a lack of available manpower due to high attrition, dispute resolutions often lead to missed TAT, and missed TAT leads to penalties being levied at contract levels.

This study aims to forecast probable attrition, which can help plan workload management efficiently thereby can help control missed SLAs and bring down penalties to an acceptable level.

B. *Data Understanding*

The attrition data collected for this study contains some important parameters as described below.

1. Employee details – ID, Employee Name, Salary, Last Performance Rating
2. Employment Details - Employee Type (Regular or Contract), Joining Date, Termination Date, Employee Level, Type of Termination, Termination Code, Cost Centre, Job Name
3. Contract Details – Contract Name, Sector, Business Category, Location City, Country

For this analysis, the focus is on the contract level. The following plot as appears in Fig. 1 shows a Pareto Chart of Attrition vs Contracts. It clearly shows top 6 which represents 15% of overall contracts are contributing to more than 80% of attrition.

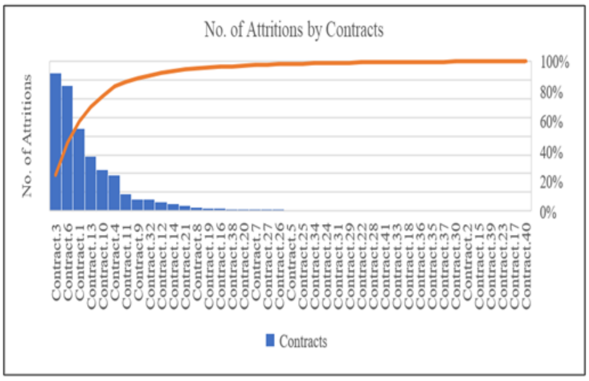


Fig. 1. Pareto – No. of Attrition by Contracts

It is also observed that the penalties paid by the top 6 contracts which total to $1,051k representing ~50% of overall penalties.

There is also a strong correlation of 0.93 exists between contract level attrition numbers and missed SLAs. Fig. 2 shows a scatter plot relation between attrition numbers and missed SLAs at contract levels.

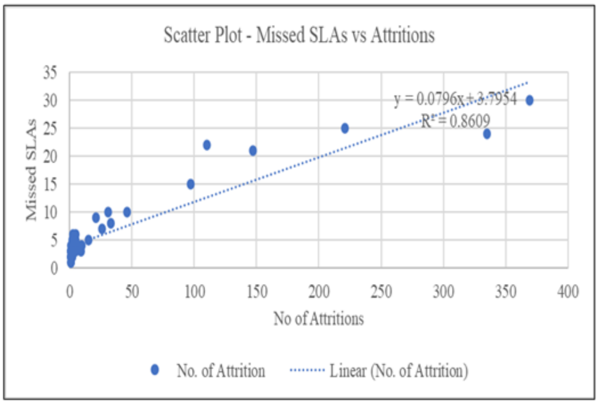


Fig. 2. Scatter Plot - Missed SLAs vs Attrition

An almost similar strong correlation is detected between missed SLAs and penalties paid at contract levels as shown in Fig. 3.

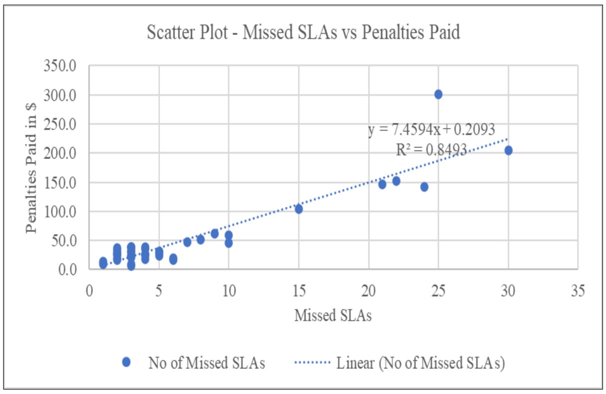


Fig. 3. Scatter Plot - Missed SLAs vs Penalties

Another key point that came out prominently is that close to 90% of attrition is at the junior most level (C01). Fig. 4 highlights this fact.

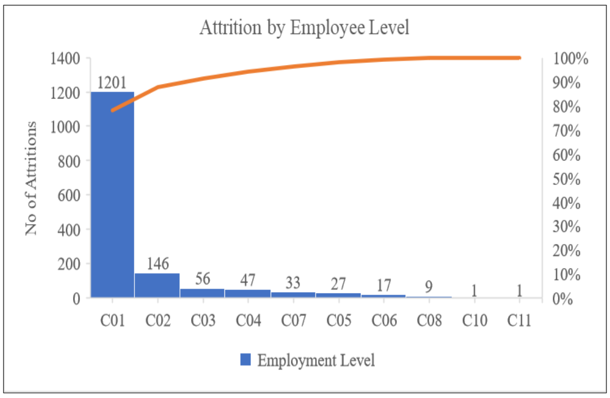


Fig. 4. Pareto Graph - Attrition contribution by Employee Level

This finding remains consistent with the overall contract level attrition trend.

A few other levels that are considered for this study are location-wise attrition and salary range. At various locations, Fig. 5 shows that the top 11 cities contributed to 80% of attritions in the last 14 months.

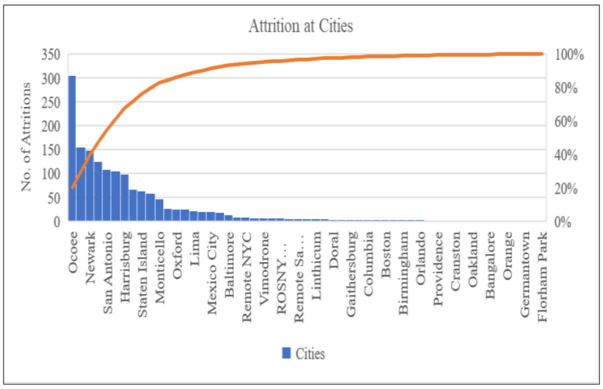


Fig. 5. Pareto Graph - Attrition contribution by cities

Attrition at the pay level is the last set of data that is examined. The salary bucket is created to identify a certain range causing attrition. The salary range of $20k to $40k is the one that causes the most attrition as shown in Fig. 6.

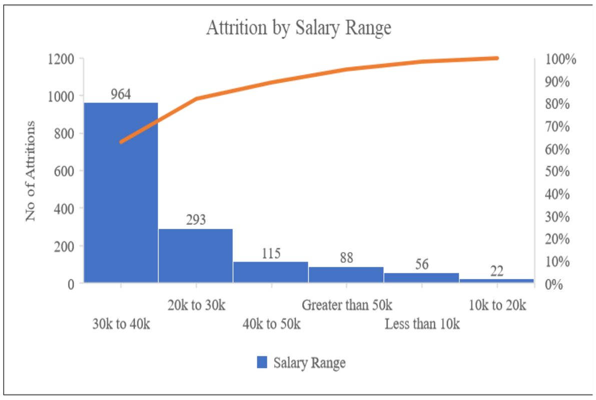


Fig. 6. Pareto Graph - Attrition contribution by salary bucket

A. *Data Preparation*

Based on termination date, the attrition data is divided into five qusrters with each quarters consist of three months of data. The first 12 months of data have been used for modeling and the last 2 months of data have been used for testing forecast accuracy.

Again, based on the termination date, the month with "MMM-YY" format is created. The quarter and the month form the basis of forecasting data.

The salary bucket is created to understand if there is any particular salary range that is contributing to high attrition. The buckets created here are "Less than 10k", "10k to 20k","20k to 30k", "30k to 40k", "40k to 50k" and "Greater than 50k".

To summarize, the data is divided into the following six categories for forecasting purpose. This was done in order to account for every level that could significantly affect the forecasting outcomes.

1. Overall attrition by quarter and month-wise.
2. Attrition by top contracts, quarter and month-wise.
3. Attrition by top employment levels, quarter and month-wise.
4. Attrition by top contracts and top employment level, quarter and month-wise.
5. Attrition by top cities, quarter and month-wise.
6. Attrition by top salary ranges, quarter, and month-wise.

The above categorizations form the basis of the forecast modeling.

B. *Modeling*

Considering the limitation of the available information in the dataset, various time series forecasting techniques are considered on each of the categorized levels of data and compared. The top technique is chosen to predict future attrition. The forecasted attrition result is further used as the input of regression to predict missed SLAs numbers and the predicted missed SLAs to predict probable penalties. These approaches are discussed in detail in the following sections.

The modeling approach starts by taking the overall data and by checking for stationarity of the data using Dickey-Fuller (DF) test. The low p-value of 0.0395 observed during DF test which is less than 0.05 indicates that the data has no unit root and is stationary. This result remains consistent for all other 5 datasets.

Then a quick check on the trend graph shows an overall upward trend as appears in the Fig. 7. This is a bit contradictory to the findings in Dickey-Fuller test. However, this trend is observed for the other 5 subsets of the data.

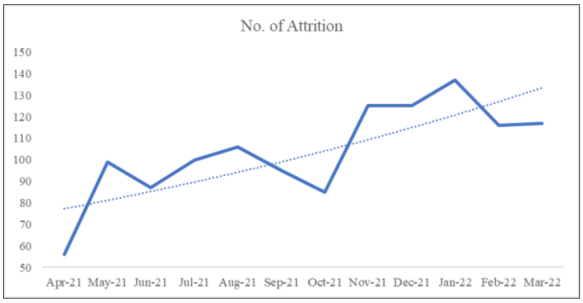


Fig. 7. Attrition trend

The first technique used is Moving Average (MA), which is followed by ratio-to-moving-average and Exponential Triple Smoothing (ETS) using Microsoft Excel's FORECAST.ETS formula.

In the MA, the moving average of three months is considered. This is done since the months are divided into quarters with each quarter consisting of three months.

Building on MA, the next technique that is used in ratio-to-moving-average. This contains some additional steps like deseasonalizing the data and building a regression model on the deseasonalized data to forecast. This method is frequently used to show the data's overall movement without taking seasonal effects into account.

The next forecasting technique considered is Microsoft Excel's in-built forecast algorithm. It uses FORECAST.ETS function and allows for auto-detection of seasonality. The other in-built function FORECAST.ETS. STAT is used to show some important stats related to FORECAST.ETS predictions.

The ARIMA, Holt-Winters (Smoothing 1, Smoothing 2 Additive & Multiplicative, Smoothing 3 Additive and Multiplicative), and LSTM techniques are also explored using python. For ARIMA, the auto ARIMA is used to find the best combination of the order (p,d,q). The best order found is (1,0,0). This order is used in ARIMA Model to forecast attrition.

Since ARIMA could not perform as expected, it led to another technique called LSTM. It is a kind of recurrent neural network that can pick up order dependence in situations involving sequence prediction. The data is divided into train and test data. After selecting test and train data, "MinMax" preprocessing technique is used on both datasets. The LSTM technique has also performed on the given dataset as expected and hence the technique is dropped for consideration on other datasets.

Finally, Holt-Winters smoothing technique is used to see if the forecast can be improved. It uses a modified version of exponential smoothing to account for a linear trend. Simple smoothing is used where the result observed is poor.

The Holt-Winter Exponential Double Smoothing is then tried to see if the forecast can be improved further. Though, the results improved but not as expected.

The last Holt-Winter Exponential Smoothing technique used is Triple Smoothing to see if the forecast can be further improved upon Excel's FORECAST.ETS function.

Finally, the forecasted attrition data is used to create a regression model to predict the missed SLAs as shown in Equation (1). The predicted missed SLAs then become the input to Equation (2) to predict penalties.

Predicted Missed SLAs = (0.07958 * Forecasted Attrition) + 3.7954 (1)

Predicted Penalties = (7.45942 * Predicted Missed SLAs) + 0.20933 (2)

A. *Model Evaluation*

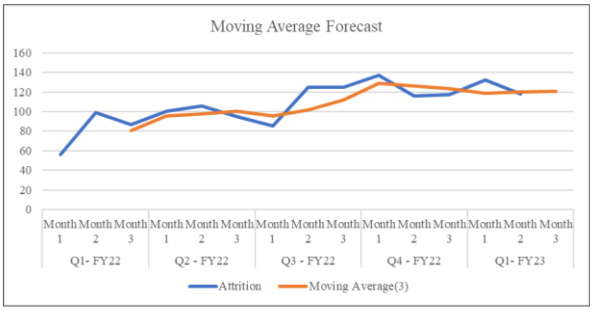
Once all the forecast techniques are used, a summary showing the model performance is presented in TABLE 2.

TABLE 2. Model Performance on overall attrition data

TS Models	MAD	RMSE	MAPE
Moving Average (3)	9.6	10.9	9%
Ratio-to-Moving-Average	11.0	12.9	12%
ETS	13.2	14.0	10%
ARIMA	15.8	18.7	14%
Holt Winters ES1	27.5	30.5	25%
Holt Winters ES2_ADD	12.1	13.5	13%
Holt Winters ES2_MUL	18.9	25.0	18%
Holt Winters ES3_ADD	16.9	19.1	16%
Holt Winters ES3_MUL	21.3	24.3	20%

When overall attrition data is considered, the best forecast technique is the Moving Average for which the MAPE is 9 %.

The graph in Fig. 8 shows how well the forecasted values perform against the original values.



The same approach is used for the other datasets to check if there is any significant difference between the various time series technique used.

I. Analysis and results

The forecasting techniques are tested on the actual attrition data of the following two months (Month 1 and Month 2). When the overall data is used, the Moving Average model is giving MAPE as 17% shown in TABLE 3 compared with the result received at the time of modeling which is 9%.

TABLE 3. Moving Average Model Outcome for Overall dataset

Overall Data					
Month	**Attrition - Actual**	**MA (Forecast)**	**MAD**	**RMSE**	**MAPE**
Month 1	132	118.8	**25.8**	**28.6**	**17%**
Month 2	158	119.7			
Month 3		120.6			

Since there is a difference of 8% between the actual versus the model prediction, the ETS is used to compare the results with the Moving Average. The ETS shows better performance on the actual numbers as shown in TABLE 4.

TABLE 4. ETS Model Outcome for Overall dataset

Overall Data					
Month	**Attrition - Actual**	**ETS (Forecast)**	**MAD**	**RMSE**	**MAPE**
Month 1	132	128.7	**14.0**	**17.5**	**9%**
Month 2	158	133.4			
Month 3		138.2			

For top contracts, both Moving Average and ETS can be used as their results are almost similar with MAPE for ETS is 6% whereas for Moving Average it is 5%. The results are shown in TABLE 5 and TABLE 6 respectively.

TABLE 5. Moving Average (MA) Model Outcome for Top 6 Contracts

Top 6 Contracts					
Month	**Attrition - Actual**	**MA (Forecast)**	**MAD**	**RMSE**	**MAPE**
Month 1	116	103.8	**6.3**	**8.6**	**5%**
Month 2	126	125.6			
Month 3		111.4			

TABLE 6. ETS Model Outcome for Top 6 Contracts

Top 6 Contracts					
Month	**Attrition - Actual**	**ETS (Forecast)**	**MAD**	**RMSE**	**MAPE**
Month 1	116	111.9	**7.1**	**7.6**	**6%**
Month 2	126	116.0			
Month 3		120.2			

For the rest of the datasets, the ETS model result is considered since it gives the best and most consistent results across all datasets.

Finally, using the regression Equation (1) and Equation (2), predicted missed SLAs and penalties on overall attrition data are calculated respectively as shown in TABLE 7. The same concept can be used for all other 5 datasets.

TABLE 7. Predicted Missed SLAs and Penalties

Overall Data			
Month	**ETS (Forecast)**	**Predicted Missed SLAs**	**Predicted Penalties in ($ k)**
Month 1	128.7	14	104.9
Month 2	133.4	14	107.7
Month 3	138.2	15	110.6
Total		**43**	**323.2**

II. Conclusions and Recommendations for future work

Attrition remains a burning problem in any sector, and it can have profound consequences when it is way above a tolerable limit.

Numerous research have been done to address this issue, however, when there is little information available, a simple solution can sometimes work wonders.

As seen here, the time series forecasting technique is used to predict future attritions across several datasets. Since this is a novel method for predicting attrition, multiple forecasting approaches are applied to a variety of datasets to determine the effectiveness of this method.

This study has shown that future attrition can be forecasted accurately even when only attrition statistics are available. The main forecasting method for all datasets has been FORECAST.ETS, a built-in Excel function.

To conclude, sometimes a seemingly tough problem can be tackled through simple approaches, in this case, attrition forecasting using time series techniques. This will help plan future workload effectively, and reduce missed SLAs and penalties.

However, this approach is suitable only when the data dimension is less. In an ideal scenario, there can be several factors that may affect a company's attrition, but with limited data, this approach is a way out as it uses monthly attrition data to forecast probable attritions for the next 3 months.

Just like with any other data modelling technique, this work has to be replicated on new datasets to determine its validity. More the data, the better the result expected as it may throw up additional trends which are probably missing in the current context.

Due to the novelty of this strategy in the current AI / ML era, the approach used in this study would open the door for similar studies in attrition predictions.

References

[1] W. Ner, "THE OFFICIAL PUBLICATION OF TRAINING MAGAZINE NETWORK Training Temperature Check," 2020. [Online]. Available: www.trainingmag.com

[2] L. Smither, "Managing Employee Life Cycles To Improve Labor Retention," 2003. [Online]. Available: www.thomas-staffing.com/survey99/retention_TABLE2.htm

[1] W. Ner, “THE OFFICIAL PUBLICATION OF TRAINING MAGAZINE NETWORK Training Temperature Check,” 2020. [Online]. Available: www.trainingmag.com

[2] L. Smither, “Managing Employee Life Cycles To Improve Labor Retention,” 2003. [Online]. Available: www.thomas-staffing.com/survey99/retention_TABLE2.htm

[3] D. Singh Sisodia, S. Vishwakarma, and A. Pujahari, *Evaluation of Machine Learning Models for Employee Churn Prediction*. 2017.

[4] R. Chakraborty, K. Mridha, R. Nath Shaw, and A. Ghosh, “Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches; Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches,” *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, 2021, doi: 10.1109/GUCON50781.2021.9573759.

[5] P. Kumar Jain, M. Jain, and R. Pamula, “Explaining and predicting employees’ attrition: a machine learning approach,” 123AD, doi: 10.1007/s42452-020-2519-4.

[6] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. de Luca, “Predicting Employee Attrition Using Machine Learning Techniques,” 2020, doi: 10.3390/computers9040086.

[7] R. Jain and A. Nayyar, “Predicting Employee Attrition using XGBoost Machine Learning Approach; Predicting Employee Attrition using XGBoost Machine Learning Approach,” 2018.

[8] “Time Series Forecasting: Definition & Examples | tableau,” 2020. https://www.tableau.com/learn/articles/time-series-forecasting (accessed Aug. 07, 2022).

[9] G. Mahalakshmi, S. Sridevi, and S. Rajaram, *A survey on forecasting of time series data; A survey on forecasting of time series data*. 2016. doi: 10.1109/ICCTIDE.2016.7725358.

[10] S. Hansun, “A New Approach of Moving Average Method in Time Series Analysis,” 2013. doi: 10.1109/CoNMedia.2013.6708545.

[11] M. Sailaja and A. R. Prasad, “Identification of Seasonal Effects through Ratio to Moving Average Method for the Number of Train Passengers and Income of South Central Railway Zone,” *International Journal of Mathematics Trends and Technology*, vol. 65, p. 11, 2019, [Online]. Available: http://www.ijmttjournal.org

[12] D. Rahardja, “Statistical Time-Series Forecast via Microsoft Excel (FORECAST.ETS) Built-In Function,” 2021. [Online]. Available: www.questjournals.org

[13] P. S. Kalekar, “Time series Forecasting using Holt-Winters Exponential Smoothing,” 2004.

[14] R. Ueno, D. Calitoiu, and D. Calitoiu@forces, “Forecasting Attrition from the Canadian Armed Forces using Multivariate LSTM; Forecasting Attrition from the Canadian Armed Forces using Multivariate LSTM,” 2020, doi: 10.1109/ICMLA51294.2020.00123.

[15] V. K. R. Chimmula and L. Zhang, “Time series forecasting of COVID-19 transmission in Canada using LSTM networks,” *Chaos Solitons Fractals*, vol. 135, Jun. 2020, doi: 10.1016/j.chaos.2020.109864.

[16] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, “Stock Price Prediction Using the ARIMA Model,” *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, 2014, doi: 10.1109/UKSim.2014.67.

[17] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, “Time Series Analysis Forecasting and Gontrol FOURTH EDITION,” 1976.

[18] T. Dimri, S. Ahmad, and M. Sharif, “Time series analysis of climate variables using seasonal ARIMA approach,” *Journal of Earth System Science*, vol. 129, no. 1, Dec. 2020, doi: 10.1007/s12040-020-01408-x.

[19] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, “Financial Time Series Forecasting with Deep Learning : A Systematic Literature Review: 2005-2019,” Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.13288

Any Additional Details

Github Link - https://github.com/Saumyadip/Timeseries-Analysis-Attrition-Forecasting