Predicting Absenteeism (Attrition) in Construction Industry in Sri Lanka

M. K. Thushara Sampath
Dept. of Computer Science
University Of Moratuwa
Moratuwa, Sri Lanka
thushara.23@cse.mrt.ac.lk

***Abstract*—Employee attrition can have significant financial impact to the employers of the any organization. The Absenteeism/Attrition of site staff in construction leads to project delays, increased costs, and decreased productivity of any construction project.**

The outcome of this study will present valuable insights to construction companies to minimize the impact of unplanned early attrition on the construction operations. The intention of this study is to build a ML model which could predict the employee early attrition in construction industry in Sri Lanka in this research.

I. Introduction

The objective of the author is to recognize the signs of the attractions of the employees and predict the employee early attrition in the construction industry in Sri Lanka.

Employee attrition refers to the discontinuation of the employees in an organization due to any circumstances. The employee attrition can be either voluntary or involuntary. The voluntary attrition occurs when an employee decides to resign from the employment by their own will, whereas involuntary attrition occurs when the company decides to terminate the employment by the management of the organization. Employee early Attrition is considered in this research as the discontinuation of the service of employees from the organization due his/her voluntary resignation during her/his first 3 months.

The Construction industry is one of the key contributors to the GDP in Sri Lanka. It is also observed that the construction industry employment is perceived as less attractive due to the temporary and fragmented nature of the employment which demonstrates by the above graph as well. The majority of the attrition happens within first 6 months.

The skill shortage is more prevalent in the construction industry compared to other industries in Sri Lanka and employee retention is very vital. Even if the machinery involvement is significant, the human resources are required for operational purposes. Further, considerable labor at many levels is required and supervision is crucial for continuous operation across the entire construction industry. The human resource is a significant factor in the performance of any construction projects. Hence, early employee attrition acts a significant role in this industry.

The objective of the research is to discover the pattern of the attendance, leaves, and income data of employees against early attrition and predict the attrition based on the trained Machine Learning model(s).

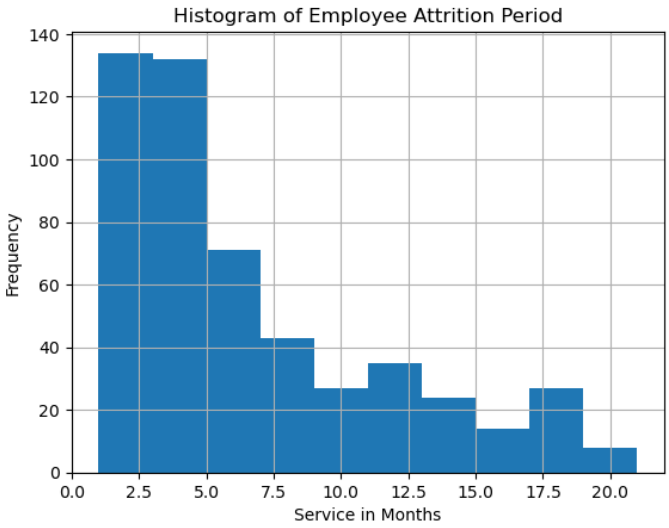

Fig. 1. Past Attrition Details

A. Machine Learning Models

The supervised learning could be considered as labels of data are known in the provided data set. The objective of supervised learning model is to learn the mapping between inputs and outputs so that accurate predictions can be made on new data. The classification based supervised models can be used for prediction as the dependent variable (IsEarlyAttrition) is a categorical variable in our study. In this case, whether a given employee would early absent can be taken as a classification problem. The following classification models are considered and used in the study.

1) Logistic Regression: Logistic Regression is a statistical method for analyzing the dataset in which there is one dependent binary variable and one or more independent variables that determine the outcome. It is used for classification problems and gives an estimation of the probability of an event occurring based on the values of the independent variables. The result is then transformed into a binary outcome using a threshold value. Logistic Regression could be used as the dependent variable is binary and the independent variables

are either continuous or categorical. Logistic regression can be evaluated for early attrition prediction since whether an employee would resign or not is a binary classification.

2) Random Forest: Random Forest can be used to solve a wide range of classification problems, including binary classification problems, multi-class classification problems, and multi-label classification problems. It is considered a powerful algorithm that can handle large amounts of data and can perform well on problems with complex relationships between the features and target variable.

II. Methodology

A. Data Description

The study is based on the historical data of Cirrus PMS which is a construction project management system developed and Implemented by Sky Management Systems (Pvt) Ltd. The Database of Cirrus PMS was used to extract the information of many construction sites in Sri Lanka. The details of employee profiles, attendance, salary, leave and holidays have been extracted.

The original data-set contains data from 35 different construction sites in Sri Lanka. The employee table contains records from 1988 which have been entered at the time of commissioning the attendance system. The attendance table however only contains records from 2022 where the attendance system was commissioned.

In order to construct the final data-set, four tables have been joined together including attendance, employee and salary and leaves tables.

Dataset	# Records	Dimension
Employee	1,273	Employee No, Employee Code, Name, Title, Address, Date of Birth, Gender, Religion ID, Marital Status, Designation ID, Date Joined, Date Resigned, Status, Status Reason, Inactive Date, Reporting employee, Employment Category, Employment Type, Religion, Designation
Leaves	1,073	Employee No, leave date, Type(Halfday/Full Day), Applied Date, Remarks, Apply type(Annual/Casual)
Salary	207,749	Employee No, Factor Name, Amount, Month, Year
Attendance	239,963	Site No, date, out date, employee no, in time, out time, Hourly Time, Shift Start, Shift End

Fig. 2. Data Description Details

B. Prepossessing

It is known fact that data prepossessing stage is the most crucial phase of an analysis, in order to retrieve correct results from any kind of analysis, a correct data set should be given as the input. The main focus of this study was on data cleaning and integrating parts under the data prepossessing phase.

Employee table had records with non-standard null values and this was rectified during data loading. Employee table also had malformed records which was skipped during data loading since they were only a small 1.86%. When checking the data types of the variables, it is clearly shown that all the variables are in the correct format. As shown in the table, missing values could be found in the variables including marital status, inactive date, resigned date and status reason. Among these variables, missing values that could be found in the variable; marital status should be handled because all the other 3 variables contain values for the resigned employees. This proportion of the missing values is 0.097 and since the marital status is a categorical variables which have only two levels, those missing values can be imputed using the mode of this variable which is 'Married'. When searching for duplicated rows, it is found out that the final data-set is free from duplicated rows.

There few numerical variable in the data-set, which is leave count, Salary Earning, Salary Deduction, No of late coming days and No of worked days which can be derived from the leaves, salary and attendance tables. When exploring the outliers as shown in the below box-plots, it is shown that there are some outliers which are slightly above than the upper bound and those outliers can be acceptable. But another two significant outliers can be detected which have higher deviations. When considering the nature of the data-set, there cannot be data entry issues in updating the employee data and when diving deep into that it was found that those two employees were under no pay leaves with special approvals. Those were removed from the study.

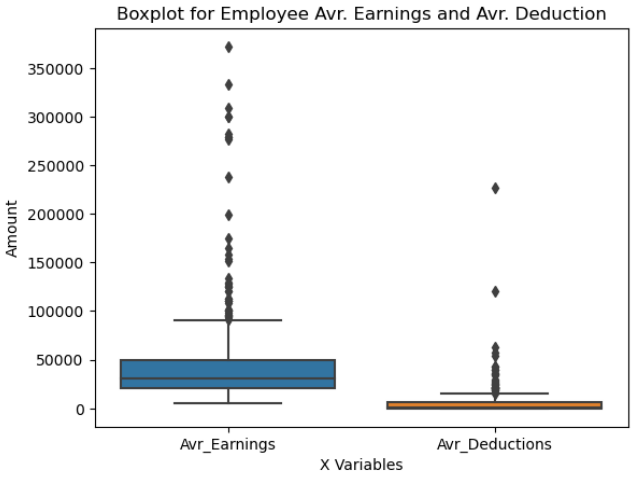

Fig. 3. Identifying outliers using box-plot

C. Descriptive Analysis

The main objective is on predicting weather an employee is likely to attrition or not, descriptive analysis were done around the employee table. The other tables were joined using the Employee_No attribute. Employee table contains 1261 employees, out of which 1241 are on contract basis and only 20 are permanent employees. The percentage of active employees is 81.6%, while 18.4% have resigned. The earliest join date in the data-set is January 27th, 1988, while the latest join date is December 20th, 2022. Although the employee table contains employee records from 1988, the attendance system has been implemented only in 2020. Therefore our analysis is narrowed down to the employees who were active in the company after February 2020. This data-set was extracted

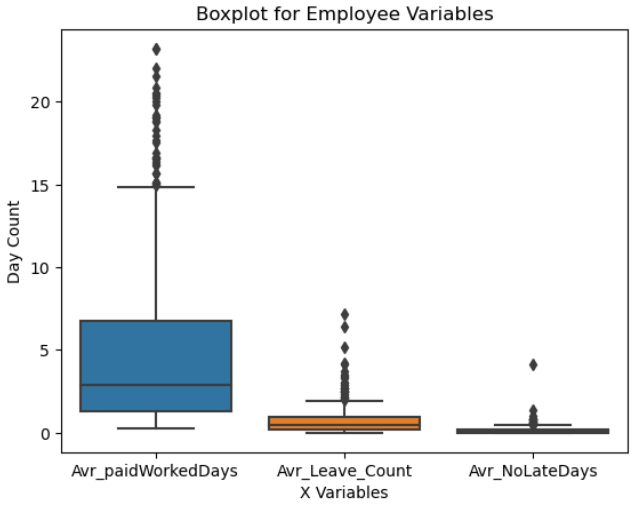

Fig. 4. Identifying outliers using box-plot

on mid January 2023. Therefore we restricted the analysis period from February 2020 to December 2022. For the focused analysis period there are 796 employees and out of those 794 are contract based employees 2 permanent employees. Out of the 796 employees 643 are inactive and 153 employees. There are very limited number of permanent employees, this analysis is mainly based on the contract basis employees. In the given employee data set the dominant are the male employees who account for 738 while female employees are only 58.

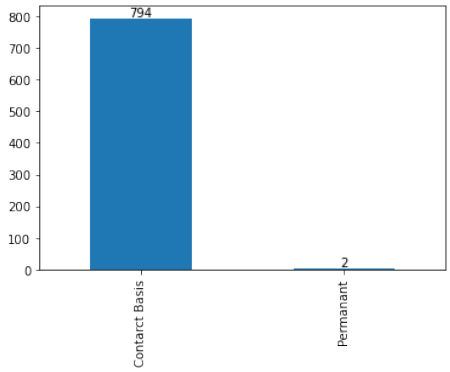

Fig. 5. Employee Type Breakdown

D. Feature Engineering

In the initial analysis some features were removed to simplify the model building. Employee_Code in employee table was removed since that attribute is not referred elsewhere. We have also removed less interesting Religion, Religion_ID, Address_Line_1, Address_Line_2, Address_Line_3, Reporting_emp_1, Reporting_emp_2 attributes from the analysis. Name attribute was also removed since that's already anonymized and uninteresting to the analysis. Designation_ID was removed since that's a duplicate feature of Designation attribute. Title attribute was also removed from the analysis

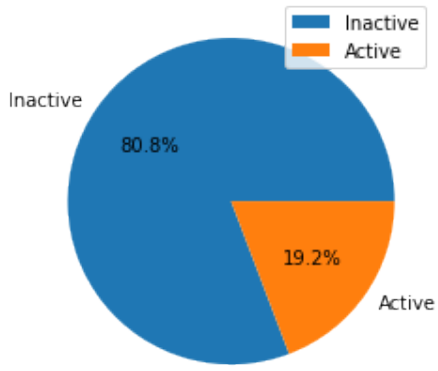

Fig. 6. Active Inactive Breakdown

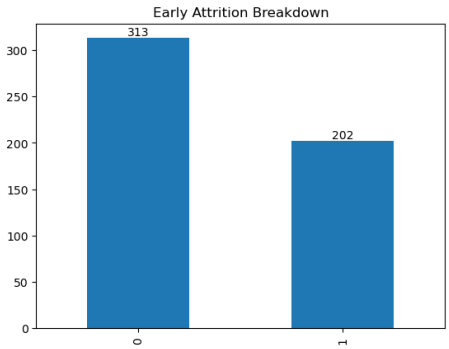

Fig. 7. Attrition Breakdown

since that is uninteresting and also a partial duplicate of the Gender attribute.

Several derived features Average Worked Days, Average Leave Count, Average Earnings, Average Deductions, Average Late Days were introduced by dividing the total count by the no of payroll months.

The objective of the study is to find whether a given employee is going to voluntarily resign or not within first 3 months. Therefore, the target variable is derived as IsEarlyAttrition. By default, target variable IsEarlyAttrition is zero for the employees who worked or working more than 3 months.

Leaves table was transformed into leave summary table by grouping by Employee_No with counting the leave factor (summation of leaves with assign 1 for full day leave and 0.5 for half day for each employee). The summary data-set was merged in to employee data-set with using Employee_No as join factor. The Same joining was applied for Earning, Deduction, Month count from salary and Late and Leave counts from attendance and Leave tables.

E. Model

Since the target variable is a binary variable which determines whether a given employee is going to be resigned, classification models are relevant. The known fact is that all the input variables for a machine learning model should

be numerical because machine learning algorithms are based on mathematical and statistical concepts. However, categorical variables also contain crucial information of the study. Therefore, all the categorical variables should be encoded accordingly.

In this data-set, six categorical variables are available which are not unique to each employee and have multiple levels. These variables are 'Gender', 'Marital_Status', 'Status', 'Status_Reason', 'Employment_Category' and 'Designation'. Among these categorical variables, it is not feasible to have encoded variables for 'Status_Reason' and 'Designation', because of the typographic inconsistencies and errors. Therefore, new encoded variables have been generated for other variables namely 'Gender_Code', 'Marital_Status_Code', 'Status_Code' and 'Employment_Category_Code'.

It is known that the features of a model should be free from multicollinearity for the model to perform well. Therefore, the multicollinearity is checked using heat-maps. That means multicollinearity is there and those variables should be eliminated when fitting the model. When considering the heat-maps which are illustrated below, it shows some signs of muliticollinearity of Month count and Staff Category. Those were removed in feature engineering.

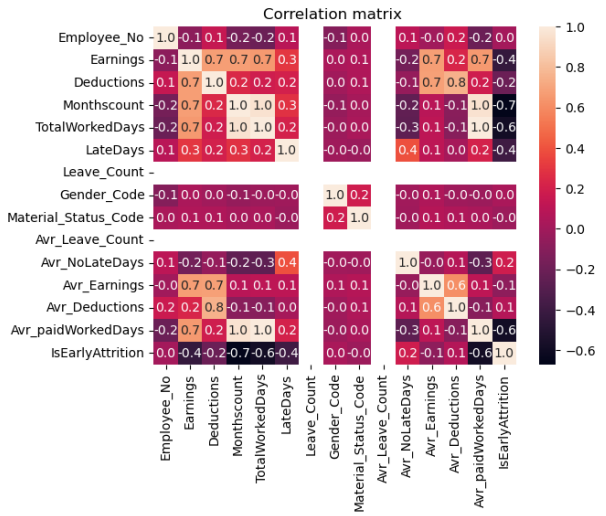

Fig. 8. Heat Map for Numerical Variables

Logistic regression and Rain Forest statistical methods were used to model the relationship between a binary outcome variable and predictor variables. When fitting the models, train and test sets have been divided into the proportion of 0.8 to 0.2. Here the predictor variables that have been included are encoded variables for Gender Code, Material Status Code, Average Worked Days, Average Leave Count, Average Earnings, Average Deductions, Average Late Days. The value are normalized in order to achieve better results.

III. Conclusion

Both Logistic regression and Rain Forest have produced high F1 scores(0.95 and 0.94). High F1 score indicates that the

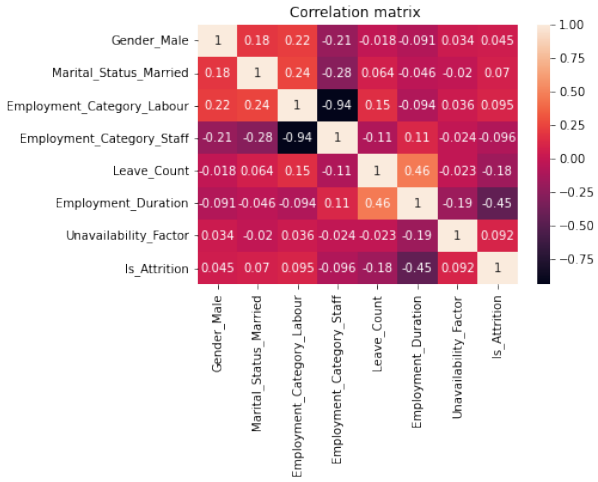

Fig. 9. Heat Map for Categorical Variables

model has both high precision and high recall, which means that it can accurately identify positive examples while avoiding false positives. Since F1 scores are approximately equal and somewhat higher in both classes, it can be concluded that both classes have been predicted correctly (Logistic Regression - 0.9223 and Random Forest - 0.9302). Further, Random Forest model shows a better fit for this use case.

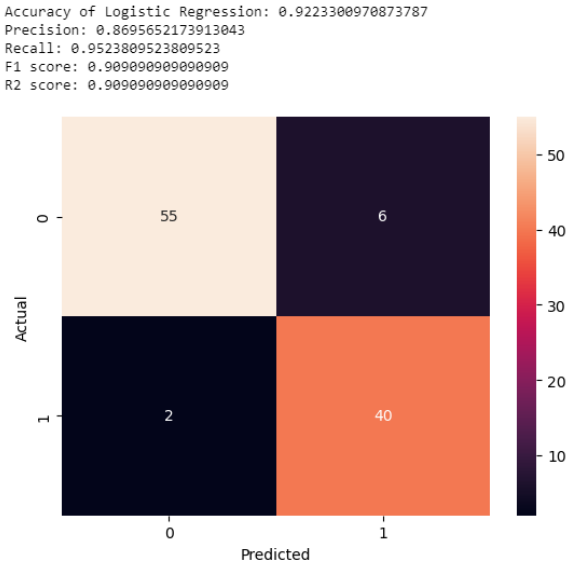

Fig. 10. Results of Logistic Regression

The selected data set in construction domain is perform better in Random forest algorithm and it predicts the attrition of any given employee at 92 percent accuracy.

The employee attrition may vary based on the different domains and different age groups. Therefore These models can be extended to analyse different designations/age groups in different organizations.

New features can be created for the model training. For an example, variables to track the behavior in the latest

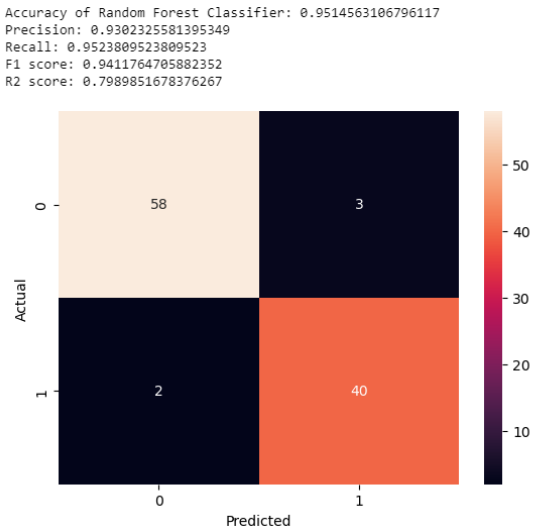

Fig. 11. Results of Random Forest Classifier

month before the resignation of the resigned employees can be created and evaluated. The same data of the employee shall be recreated by considering the cut of date into previous months.

Further the same feature set can be extended to predict the probability of resigning employees in 3 months, 6 months, 12 months etc. Predicting the service period of an employee using similar kind of historical data will be an other area which could be studied.