Analyzing Factors Associated with Employee Turnover in an Apparel Manufacturing Factory in Sri Lanka

Azeem Ameen (ahamed.20@cse.mrt.ac.lk) | Sampath Thennakoon (sampath.20@cse.mrt.ac.lk)

Department of Computer Science & Engineering
Faculty of Engineering
University of Moratuwa
Sri Lanka

***Abstract*—Apparel manufacturers in Sri Lanka are major players in Sri Lankan economy and immense contributors of employment to the workforce. The research investigates a key problem of such a manufacturer who produces and exports apparels of world-leading brands with a workforce of more than 1200 employees. The employee turnover rate of the factory has been on the rise and the first six-month termination rate is the critical contributor to it. This research attempts to identify the factors influencing the first six-month termination rate and predict such possible terminations in the future. The descriptive study, as well as the diagnostic study conducted using logistic regression and decision tree classification, proved that salary of an employee is a significant factor in determining the stay of an employee. It also revealed the importance of the demographic factors of the employees. The predictive model built using decision tree classification methodology will forecast employees who potentially could leave and help Human Resource teams to take necessary actions to mitigate the risk.**

***Index Terms*—Apparel Manufacturing, Employee Turnover, Employee Attrition Prediction**

I. Introduction

The apparel sector in Sri Lanka is a dynamic contributor to the Sri Lankan economy and has helped the country to grow as well as support many families to build their future. Apparels manufactured and exported in Sri Lanka is considered to be one of the best in the region and Sri Lanka is among the top apparel producing countries in the world relative to its population. Among the total exports in Sri Lanka, the apparel industry accounts for more than 50% of the share. In 2018, Sri Lankan apparel industry registered a $5.2b exports in revenue becoming the first industry to cross the $5b mark in the country[1]. The apparel manufacturing industry in Sri Lanka, employs about 15% [2] of the countries workforce. It utilizes a strong female workforce which provides direct employment opportunities to a substantial number of women in Sri Lanka. This industry has provided the rural poor families a platform to grow, succeed and build a life around it. The factory considered in the investigation is one of such factories that manufactures apparel garments and exports it to many parts of the world. Situated in the Western province of Sri Lanka, the factory has a labor force of more than 1200 employees. As one of the factories manufacturing world-class short lead time apparel products, one of the main problem factory facing is its instability of the workforce. Employee turnover has been the main concern that requires a solution. According to factory data, it has a 5% employee turnover rate and among it, the employee turnover rate of employees who have been working in the organization for less than six months has been the main concern. When employees leave, the factory must rehire and retrain which in hand will incur a high cost. The process of familiarizing the new employees to the factory floor and the methods will take time thus creating an instability in the production floor. This will lead to delays in fulfilling orders and a drop in the quality of the products which will create other bigger problems. Thus, minimizing or controlling the employee turnover rate is a paramount problem and requires a deep analysis. The objective of this study is to analyze the factors influencing termination within the first six months based on recruitment and payroll data provided. Also, to build two predictive models that can be used by plant recruitment and the Employee Relations team. The predictive model required to the recruitment team (Recruitment Model) should be able to predict whether a candidate if given an opportunity, will get terminated within the first six months or not. This information can be used in the recruitment decisions of the factory. The predictive model required to the Employee Relations team (ER Model) should be able to predict whether an existing employee will get terminated within the first six months or not thus the team can take necessary steps to avoid the risk. The result in the analysis may provide the recommendation to factory teams who are focusing on reducing the turnover problem.

II. Data Collection and Processing

Data provided for this analysis is obtained from the recruitment and the payroll teams of the factory. At the recruitment stage of an employee, recruitment teams collect certain information on demographics, previous engagements and personal impressions from the employee. The payroll team enters monthly payment details of each employee categorizing it into basic salary, incentive salary, overtime salary, and gross salary. For this investigation, a combined data set is build using the data provided by both teams. 1370 data points collected from January 2018 to December 2019 is used in this analysis. Among them, 283 observations had incomplete data and hence were omitted from the analysis. To achieve the objectives two sets of data were prepared. Since payroll data will not be available in the recruitment stage, to build the Recruitment Model, data set without salary variables is considered. For the ER Model, the data set with the salary variable is considered. When considering the salary variables, 2nd month salary is used for this analysis. Employees who did not receive a complete 2nd month salary is removed from this data set.

III. Methodology

To understand the factors influencing the termination of an employee within six months, a detailed study on the data collected was undertaken. A descriptive analysis and a diagnostic analysis using Logistic regression and Decision Tree was done to find the factors influencing the termination. Also, two predictive models (Recruitment Model and ER Model) were built using the two data sets to predict termination at different stages using Logistic Regression, Decision Tree and Random Forest classification methodologies.

IV. Analysis and Results

A. Identifying variables affecting the termination of an Employee

1) Descriptive Analysis: Data set contains 37 attributes, which provide information on demographic details, location details and previous engagement details of an employee. Only factory related variable considered here is the salary variable.The initial data analysis is done using well-known python libraries[3], [4], [5], [6].

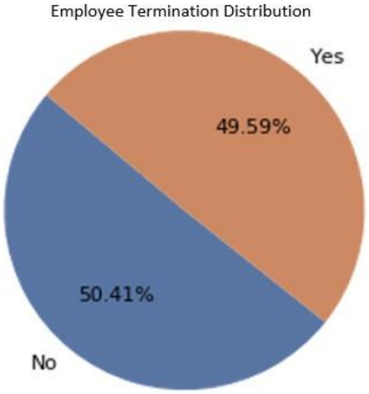

Fig. 1. Representation of the distribution of employee termination.

According to the figure 1, the first six-month termination is 49.59%. this is means that almost half of the new employees will get terminated in the first six months. Factory's concern over the first six-month termination is valid and requires focus attention.

Figure 2, elaborates on the relationship between the Age of an employee when joined with the employee stay. Most of the employees who have joined the factory are in the 20-25 Age category. When compared, the termination probability within age categories, employees who are in the age category less than 20 and 20-25 category have a high chance of leaving. When age increases, the percentage not leaving the factory increases.

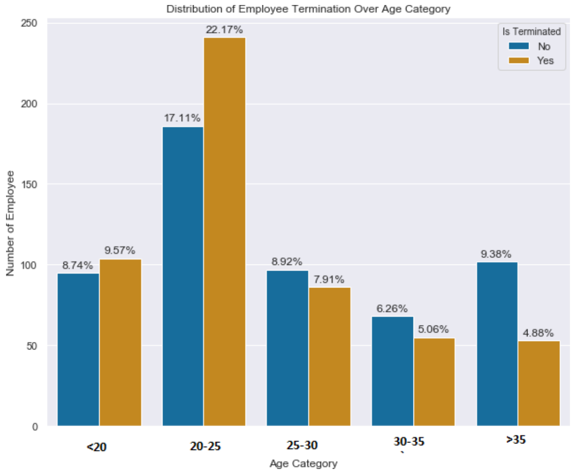

Fig. 2. Representation of the distribution of employee age categories.

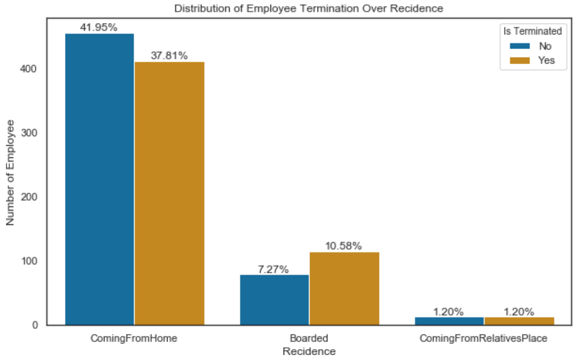

Fig. 3. Representation of the distribution of employee residencies.

Figure 3, shows how termination is affected by the location or the place of stay of an employee. Comparing the termination rate and non-termination rate in each case, people coming from Boarding places have a higher tendency to leave the factory. Employees who are coming from relatives' place has an equal probability of leaving and staying suggesting that it does not affect the problem.

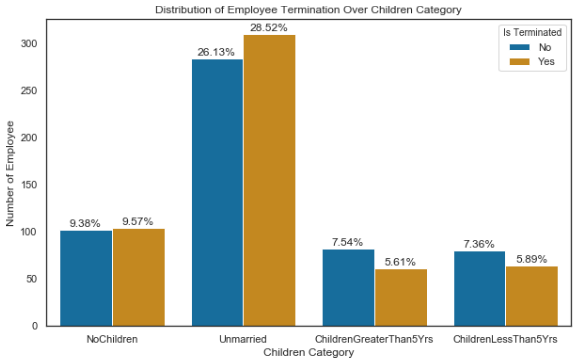

Fig. 4. Representation of the distribution of employee children categories.

Figure 4, represents the marital status and parental status of employees in one diagram. According to the graph, most of the employees joining the factory are unmarried. Comparing the termination rate and non-termination rate in each case, employees who have children have a lesser tendency to leave the organization. And when the number of children increases termination probability decreases.

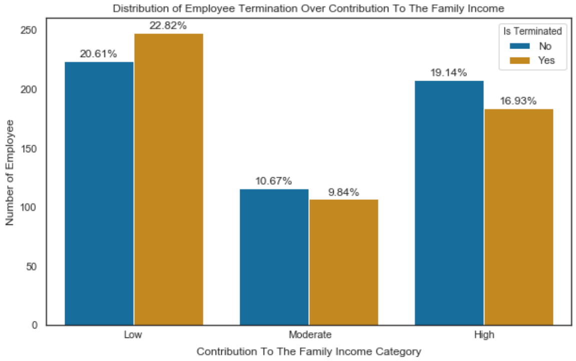

Fig. 5. Representation of the distribution of employee family income category.

Figure 5, displays the contribution of the salary to the family income of employees of both terminated and non-terminated categories. According to the figure, employees with low contribution has a higher probability of leaving. Comparing the termination and non-termination within categories when contribution increases the tendency of leaving reduces.

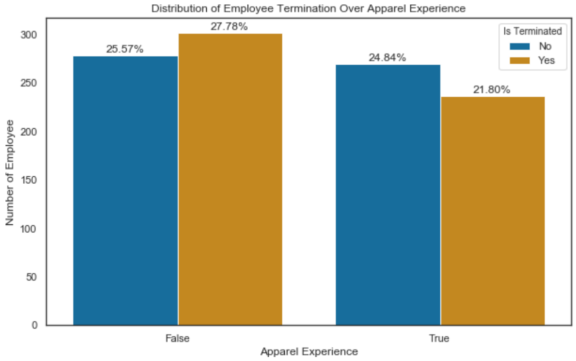

Fig. 6. Representation of the distribution of apparel experience.

According to figure 6, when compared to the termination and non-termination probability within each category, employees who have worked in an apparel factory and has experience on it have a less tendency to leave. Figure 7 shows that employees with sewing experience has a lower termination probability. This suggests that employees with apparel manufacturing skill survives more than who have no skill.

Figure 8 and 9 shows, the distribution of 2nd month gross salary plotted against the termination status. According to the box plot, the 2nd month gross salary distribution of employees who have left the factory is comparatively low when compared to employees who have stayed. According to figure 9,

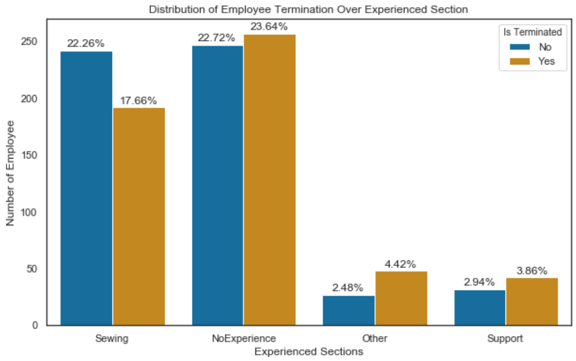

Fig. 7. Representation of the distribution of experience section.

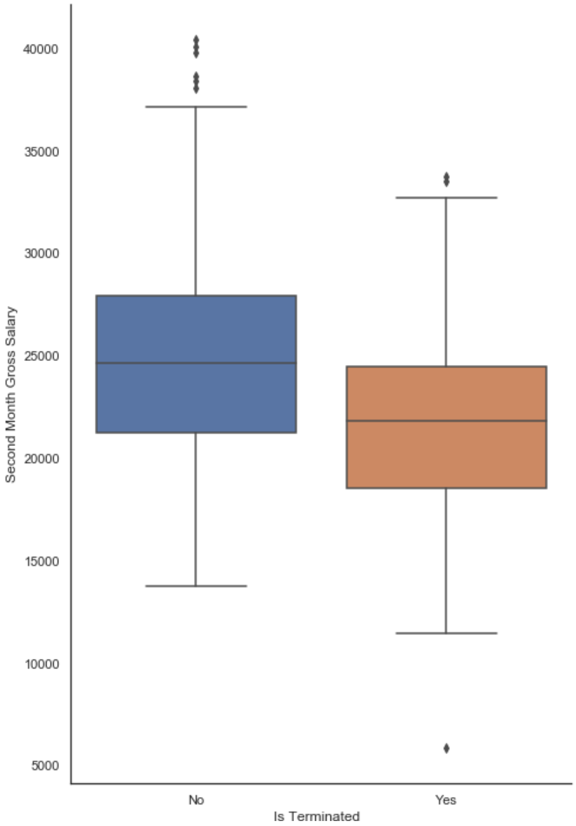

Fig. 8. Representation of the distribution of second month gross salary.

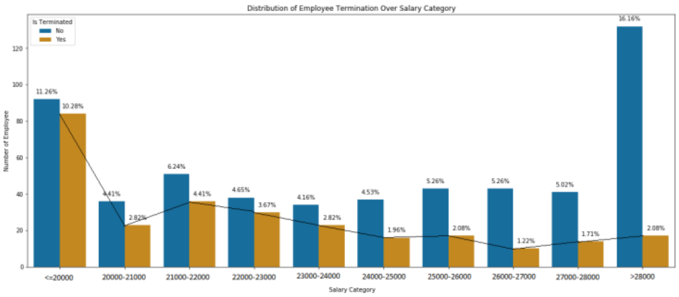

Fig. 9. Representation of the distribution of second month gross salary.

termination probability within each salary category reduces as the amount increases. The variability in termination status for employees who receive more than 28000 is very significant suggesting salary plays a key role in the determination of the stay of an employee. This diagram also shows the variability of incomes employees receive. 54% of the employees have obtained a salary of less than 24000 and 18.24% of the employees have received a salary of more than 28000. This variability in pay influences the termination thus needs actions to mitigate the risk involved.

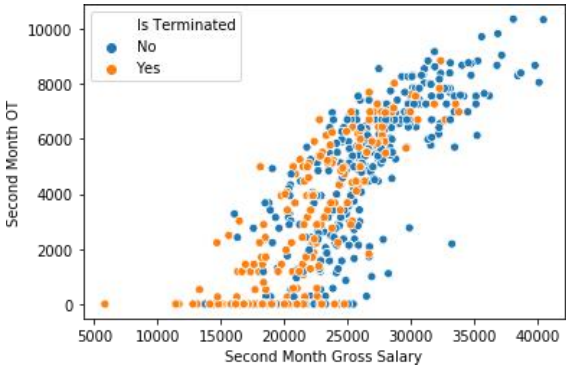

Fig. 10. Representation of the distribution of employee Overtime Salary with Gross Salary.

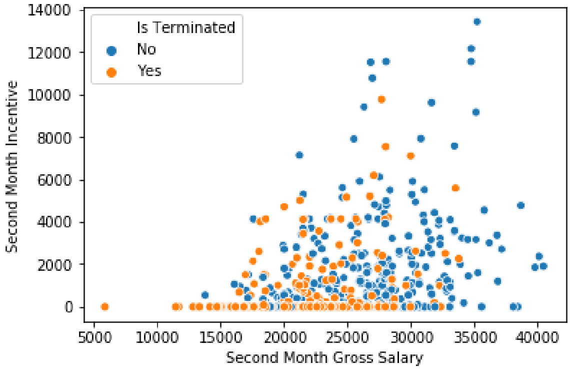

Fig. 11. Representation of the distribution of employee Incentive Salary with Gross Salary.

Figure 10, 11 shows the distribution of 2nd month overtime salary and 2nd month incentive salary plotted against the 2nd month gross salary respectively. According to figure 10, employees who have received a lower Overtime salary has terminated more. Some employees have been terminated despite getting high Overtime salary. This might be because of the workload and time management issues due to working overtime. According to figure 11, the number of terminated employees is high when the incentives received are low. Management has to optimize the workload content and balance the overtime working and review the incentive schemes to better suit the employees.

2) Data analysis using logistic regression: Logistic regression is a specialized form of regression used to predict and explain a categorical dependent variable. It works best when the dependent variable is a binary categorical variable. One special advantage of logistic regression is that it is not restricted by the normality assumption which is a basic assumption in the regression analysis[7].A logistic regression model was fitted to the data sets including salary variables and data set without including the salary variable.

The forward selection technique began with no variables in the model. Explanatory variables were subsequently added to the model one at a time. At each step, each variable that was not already in the model was tested for inclusion using BIC criteria. The likelihood ratio test was used to test the significance of the model with the new variable against the current model. By going through the iterations best model was obtained.

TABLE I

Logistic Model summary obtained for data set with Salary Variable

	Estimate	**Std. Error**	**z value**	**Pr(>|z|**
(Intercept)	-2.36E+01	1.67E+03	-0.014	0.98874
Gross Salary	-2.40E-04	4.31E-05	-5.56	2.70E-08
Overtime Salary	1.92E-04	6.39E-05	3.009	0.00262
Basic Salary	-2.40E-04	1.19E-04	-2.021	0.0433
Family Opinion About the Job (=Like)	1.56E+01	8.19E+02	0.019	0.98481
Extra-Curricular Activities (=True)	9.48E-01	4.25E-01	2.231	0.02565
Expectation of Doing the Job (=Short term)	-5.70E-01	3.09E-01	-1.844	0.06522
Permanent Residence (=Yes)	-3.46E-01	1.98E-01	-1.742	0.08155
Apparel Experience (=TRUE)	-2.71E-01	2.02E-01	-1.344	0.17901
Medical Test (=Passed)	1.59E+01	1.46E+03	0.011	0.99128
Weight	1.63E-02	1.25E-02	1.3	0.19346

Based on the model obtained, represented in Table I, at 99% confidence level variables gross salary, overtime salary, basic salary, extra-curricular activities, permanent residence, expectation of doing the Job variables became significant. At 95% confidence level gross salary, overtime salary, basic salary, extra-curricular activities are the variables that become significant.

Based on the above results it shows that Salary variables play a significant role in determining the stay of an employee. With variables extra-curricular activities, permanent residence, expectation of doing the Job variables becoming significant shows that demographic variables also play a role in determining the stay of an employee. Employees who are involved in extra-curricular activities has a high termination possibility compared to those who are not. Short term expectation employees have a higher tendency to stay compared to Long Term expectation employees. Also, employees who live close to the factory tends to stay more time in the organization.

For the model fitted to the data set without including salary variables represented in table II previous workplace, weight, height were the significant variables. Based on the results, compared to employees who have joined working in a different company factory, employees who have worked in the same factory have a lower tendency to leave and employees who have no previous job experience or worked in a different company have a higher tendency to leave. Employee’s weight and height also plays a

TABLE II
Logistic Model summary obtained for data set without Salary Variable

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	31.68302	1253.574	0.025	0.979836
Previous Workplace (=No Job)	0.26065	0.46663	0.559	0.576444
Previous Workplace (=Different Company)	0.78794	0.39351	2.002	0.045248
Previous Workplace (=Same Plant)	-1.35044	0.50712	-2.663	0.007745
Weight	0.03559	0.01012	3.519	0.000434
Height	-0.0243	0.01151	-2.111	0.034764
Retention Category(=B)	0.70812	1.22469	0.578	0.563127
Retention Category(=C)	1.33527	1.21621	1.098	0.27225
Retention Category(=D)	-26.3881	982.6491	-0.027	0.978576
Availability of Transport (=Yes)	-15.027	671.5021	-0.022	0.982146
Apparel Experience (=TRUE)	-0.47605	0.28072	-1.696	0.089923
Following External Courses (=Yes)	-15.8512	1058.55	-0.015	0.988053
Following External Courses (=No)	-16.6414	1058.55	-0.016	0.987457
Extra-Curricular Activities (=True)	0.51803	0.37663	1.375	0.168993

significant role in the duration of stay. In the manufacturing environment employees' physical conditions are tested and maybe the reason for the above result.

3) Data analysis using decision tree: Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated data sets without imposing a complicated parametric structure[8]. Unlike other data mining tools, the decision tree is a non-black box model that provides the reason and variables behind the classification which is useful for the Factory team when taking insights from data and to apply on the factory floor. Like in the logistic regression case, the decision tree model is fitted for both data sets.

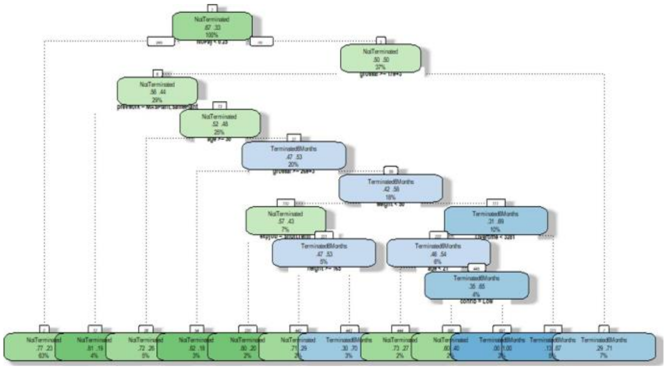

Fig. 14. Decision tree fitted to the data set including Salary Variable.

Based on the decision tree obtained, no paydays, gross salary, previous workplace, age, weight, expectation of doing the job, overtime salary, contribution to the family income variables becomes the variables required to classify the termination status of an employee.

For the model fitted without using the salary variable, IQ test results conducted at the recruitment stage, expected salary, reason for leaving the previous workplace, involvement with extracurricular activities, height, age, weight and no. of children are the variables required to classify the termination status of an employee.

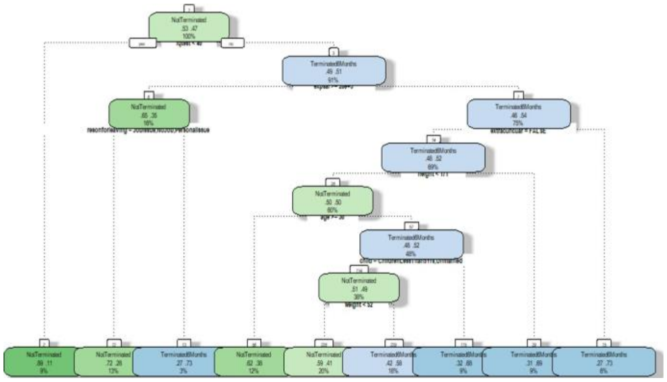

Fig. 15. Decision tree fitted to the data set without Salary Variable

Using the three-analysis conducted above, factors influencing the termination of an employee within six months can be identified. By looking at the variables, the factory team would be able to take the necessary steps and mitigate the risks involved.

B. Predicting Employee Turnover

To achieve the objectives of building predicting models for recruitment and employee relation teams, models built using logistic regression and decision trees were used. Also, a random forest model was fitted for both data sets.

TABLE III
Forecasting performance of different classification methods fitted for data set with Salary variable (ER Model)

Model	Accuracy	Recall	Precision
Logistic Regression	68%	31%	48.3%
Decision Tree	67.8%	39%	43%
Random Forest	74.1%	15.3%	34.5%

TABLE IV
Forecasting performance of different classification methods fitted for data set without Salary Variable (Recruitment Model)

Model	Accuracy	Recall	Precision
Logistic Regression	54.8%	75.7%	51.7%
Decision Tree	53.3%	14.5%	46.9%
Random Forest	32.6%	75.8%	34.7%

For the recruitment model, the highest accuracy rate is obtained by fitting the logistic regression model. Comparing with the ER model, the accuracy rate of each classification model is low for the recruitment model. This suggests, the variables captured in the recruitment phase-only are not enough to predict an employee termination and needs to add factory specific internal variables like salary to improve the result. This means that internal factors play a major role in employee termination than external factors. Ability to capture variables like employee workload, targets, skill, product difficulty, shift patterns, etc would help to improve the model. Considering the ER model, random forest gives a higher accuracy rate compared to other models. In this case, more than the accuracy focusing on recall is vital as identifying a true termination is the key factor here. Correctly predicting and identifying a possible termination will help the ER team to get involved and take necessary actions to reduce the risk, thus reducing the employee termination rate of the factory. Based on the recall, the decision tree classification model gives a better value compared to the other models. As suggested above, adding more factory specific internal variables of an employee would increase accuracy and recall values. The decision tree model will be a handy tool for the ER team to use against their quest of reducing the employee turnover rate.

V. Discussion and Summary

This data-driven approach to analyze the first six-month termination would be a solution to the employee turnover problem the factory is facing. Preliminary analysis showed, demographic and salary variables play a role in the first six-month termination of an employee. Diagnostic tests conducted using logistic regression and decision tree classifications further proved the results. Based on the tests, 2nd month salary variables, age, location or stay, contribution to the family income, apparel experience and physical conditions like height and weight are critical variables in determining the stay of an employee. Factory teams can use these results and check the ability to influence those variables in a manner that favors the factory. Predictive analysis conducted showed, demographic variables and the other variables collected in the recruitment stage only, are not enough to predict the termination of an employee, proving the importance of the salary variables. Analysis has shown that there is a huge variability in the salaries that employees receive. If factory teams can reduce the variability and increase the salary of employees, the factory would be able to reduce the employee turnover rate. Also including factory specific internal variables of an employee would help to increase the predictability of employee turnover.

VI. ACKNOWLEDGMENT

We would like to extend our gratitude to Dr. Shehan Perera and to Dr. Uthayasanker Thayasivam for their support and guidance to carry out this analytic study. Further, gratitude goes to the factory team to providing us a data and a domain problem worth solving.

References

1 “Sri Lanka apparel exports $ 5b milestone, what lessons to learn and what does it mean for SL?” [Online]. Available: http: //www.ft.lk/columns/Sri-Lanka-apparel-exports 5-b-milestone--what z -lessons-to-learn-and-what-does-it-mean-for-SL-/4-673218

2 “Sri Lankan Apparel.” [Online]. Available: https://www.srilankabusiness.com/apparel/

3 van Rossum, G., *Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python’s design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.* [Online]. Available: https://www.python.org/

4 McKinney, W., *Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool,built on top of the Python programming language.* [Online]. Available: https://pandas.pydata.org/

5 *Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.* [Online]. Available: https://seaborn.pydata.org/

6 Hunter, J. D., *Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.* [Online]. Available: https://matplotlib.org/3.2.1/index.html

7 Pregibon, D., “Logistic Regression Diagnostics. Ann. Statist.” vol. 9, pp. 705–724, 1981, doi:10.1214/aos/1176345513. [Online]. Available: https://projecteuclid.org/euclid.aos/1176345513

8 Song, . L. Y., “Decision tree methods: applications for classification and prediction.” vol. 27(2), p. 130–135, 2015. [Online]. Available: https://doi.org/10.11919/j.issn.1002-0829.215044