Employee Turnover Prediction by Machine Learning Techniques

Ang Chyh Kae[1], Chew XinYing[1], Johnson Olanrewaju Victor[1] and Khaw Khai Wah[2]
[1]*School of Computer Sciences, Universiti Sains Malaysia, 11800, Gelugor, Penang.*
[2]*School of Management, Universiti Sains Malaysia, 11800, Gelugor, Penang.*
xinying@usm.my

Abstract— **Employee turnover in Human Resource (HR) analytic is a term used to describe employees who leave the company due to termination, seek better job, or they are dealt with a bad working environment. Typically, a high turnover rate indicates that employees are dissatisfied with their current work environment. This leads to a high cost in terms of productivity, time and money for the company as they were required to hire, rehire, and retrain the new employees to accustom themselves with their new work environment as well as the tasks assigned. In this paper, we propose a hybrid of machine learning algorithms and a Power BI model to design an Employee Turnover Prediction (ETP) application. Main factor influencing employee exit decisions and employee retention periods will be identified and the retention period for the employees or new applicants will be predicted. Employee dataset with the relevant features will be collected, processed, and analyzed. The analytics results (retention period) act as a benchmark for companies to determine whether they should hire applicants which also would possibly benefit to reduce the turnover rate of their company.**

Index Terms— **HR Analytic, HR Attrition, Machine Learning, Data Analytics, Data Science, Retention Period, Prediction**

I. Introduction

Employee turnover occurs in every company, no matter what their business, whether large-scale or small-scale. Employee turnover or employee churn is a costly problem for companies. The actual cost of replacing an employee is often reasonably large, depending on the experience and skillsets the employee possesses. According to a study by the Center for American Progress, around 20% of an employee's salary was typically paid by the companies to replace that employee, and the cost could significantly increase if higher executives or highest-paid employees are to be replaced [1].

A high employee turnover rate normally costs a company by reducing their productivity, increase in monetary costs for the hiring process of new employees, and wasting their time replacing the employee. A high employee turnover rate is usually an indicator that a company is likely having a problem. Usually, employees will not simply leave their workplace if they were satisfied with their current job. However, it is not easy to identify the main factors that cause the employee turnover.

According to, [2] some of the identifiable factors may include: i.) business morale problems, ii.) lack of opportunity for advancement or growth, iii.) unequal or substandard wage structures, iv.) being overworked, v.) lack of feedback and recognition, vi.) little opportunity for decision-making, vii.) poor new employee selection, etc. [3]. Whenever an employee leaves a company either voluntary or involuntary, they will be replaced. If the employee is leaving under involuntary turnover, it may help the company to save its cost. However, if the employee is leaving under voluntary turnover, it may possibly cost about twice the employee's salary to locate and hire a replacement. The cost varies across different industries, but for some employers, it can be even higher since some employees could possess certain key skill sets that are difficult to obtain at the market. Replacement for these employees is always at a huge cost and time to hire, rehire, and train the new employees, in as much the recruitment agency fee is also inclusive [4]. Therefore, it would be better if the employee retention period could be predicted earlier to avoid frequent re-cycling of the process. Apart from the reasons employees could withdraw and start looking for other opportunities, employee turnover caused by working environment issues, poor employee welfare, poor employee relation, peer pressure, etc. can be better managed by the retention plan system. A retention plan system, if well-structured and strategically deployed could provide the company better opportunities to utilize the latent potential of its data, identify the root causes of employee turnover, take appropriate measures to decrease the rate of turnover and overall save the reputation of the company [5].

Machine Learning (ML) with its evolving capabilities has been applied in similar fields such as banking [6], fraud detection [7], retail [8], online gaming [9], insurance service provider [10], etc. In literature, diverse types of ML-based approaches are underscored for developing employee retention plan models. Most of these approaches consider the application of a single ML technique for churn prediction. Such techniques include but are not limited to Decision Trees (DT), Artificial Neural Network (ANN), Naive Bayes and Support Vector Machines (SVM) [11 – 12]. Furthermore, the classification technique was also received great attention [13 – 15].

However, in literature, much emphasis is laid on prediction using classifier methods with little or no consideration for visualization techniques for data exploration. In addition, there is no existing application model that could filter and rank the root cause of employee turnover and employee retention period, thereby providing prediction on the retention period of employees. Therefore, by using ML methods and the dataset of previous employees, we proposed an Employee Turnover Prediction (ETP) application model to predict employee retention period by identifying the main factors that influence both the turnover and the retention period of the employees. We further explored Power BI for visual exploration of the dataset to consolidate the prediction of the selected ML to gain better insight into the "why" an employee leaves a company. The use of test cases in our experimental findings indicates that several methods of producing useful classifications with visualization technique used benefit in providing a better insight into employee turnover.

The contribution of this paper can be summarized into two-fold. To addressed the gaps found in the literature, (1) authors proposed an Employee Turnover Prediction (ETP) application model to predict employee retention period by identifying the main factors that influence both the turnover and the retention period of the employees. (2) Authors also explored Power BI for visual exploration of the dataset to consolidate the prediction of the selected ML to gain better insight into the "why" an employee leaves a company. The proposed Hybrid model able to identify the main factor that influences employees' leaving decisions; predict employee retention period and also the retention period for the employees or new applicants. The contribution of this research is particularly important for the industry as the results of analytics (retention period) could act as a benchmark for companies to determine whether they should hire the applicants which would also possibly help to reduce the turnover rate of their company. Besides, it helps to minimize the cost to hire, rehire, and retrain the new employees to accustom to their new working environment and their given tasks. Finally, the use of test cases in our experimental findings indicates that several new methods produce useful classifications, optimization, feature selections, with the visualization techniques used helpful in providing better insight into employee turnover use cases.

The remainder of the paper is organized as follows: In Section II we first discuss the literature review and related works, after which we explain data mining and related ML classifiers used in the paper in Section III. We then present preprocessing, feature selection, model selection and ETP system design architecture under methodology in section IV. In section V, we discuss the result and the findings of our ETP application and conclude in Section VI.

II. LITERATURE REVIEW

A. Employee Turnover versus Attrition

Before proceeding with further discussion of proposed ETP, there is a need to understand two synonymous concepts: employee turnover and employee attrition. According to [16], employee attrition refers to the loss of employees through a natural process, such as retirement, resignation, elimination of a position, personal health, or other similar reasons. In other words, when it comes to employee attrition, the reason they leave the company is not due to a problem with their jobs, but it is a matter of life unfolding. With employee attrition, an employer will not fill the vacancy left by the former employee. Therefore, employee attrition is often considered as a way that companies could decrease their labour costs when they are facing financial distress. Figure 1 on the other hand, described employee turnover consisting of two different types: voluntary turnover and involuntary turnover. Both terms though indicate the loss of an employee, with the organization's intent to fill the position vacancy [16]. Employee turnover is often viewed as a negative impact on a company and as a burden for employers. Voluntary turnover refers to an employee leaving the organization with their own decision and reasons, probably getting a better deal from another company, lacking growth opportunities in their current role, or having a hostile working environment. While, involuntary turnover refers to an employee terminated or fired from the company dues to poor performance, unethical behavior, or more. Typically, a high employee turnover rate indicates the working conditions are not optimal, poor salary offered, or employees were unsatisfied with their current job and vice versa.

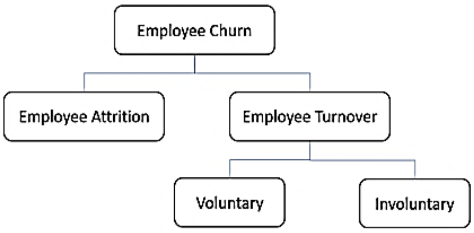

Figure 1. Attrition vs Turnover in Employee Management [16]

B. Related Works

[11] presented in their paper an employee churn prediction model by employing five classifiers namely DT, Naïve Bayes, ANN and SVM. A subset of employees and customers involved in a specific client unit within a large organization dataset was collected over a year and a half. A random split of 80/20, train and test set were performed on the dataset. Each of the classifiers was set up with unique tuning parameters. Total prediction accuracy was underscored for the five classifiers with SVM having overall performance accuracy of 99.83%. A substantive employee retention rate was achieved.

In a similar vein, [17] presented time-series-based classifiers to calculate the client-side inventory and based on periods of inactivity of usage and the pre-purchase of the unitary service classifies the clients either as lost or retained. Different lengths of the time series were explored as input to the selected ML, period of inactivity labels client as lost or otherwise, and different forecast horizons for client loss being detected. Both linear and radial SVM performance were benchmarked against KNN, RF and AdaBoost for the time-series length, inactivity period's length, and the forecast horizon. RF with 92% specificity has the best performance for predicting lost clients.

III. DATA MINING PROCESS AND MACHINE LEARNING

Before the feature engineering process and model building in the data science lifecycle, the Extract, Transform, and Load (ETL) process is performed to clean the data for further use. ETL is a generic process in which the data is first obtained, then processed, and finally loaded into a data warehouse or databases or other files such as Excel as shown in Figure 2. In the real world, not all data extracted is usable in its native form; the data obtained would contain a null value or inappropriate value. To eliminate these values, data cleansing and data manipulation will be performed such as replacing these values with corresponding reasonable numerical or categorical values by using mean or median. After data transformation, these data will load into the target data source or data warehouse. This process is considered important as an inappropriate or null value would bring inaccuracy of the predicted results by the ML algorithm. As earlier stated, to avoid having a high employee turnover rate in one's company, the retention period of the employee could be predicted in advance. By predicting the employee retention period, the employer could decide whether they should hire the employee based on the results of the predicted retention period. Historical employee data and ML algorithms were used to predict the employee retention period. Besides, the main factors that caused the employee turnover and the main factors that affect the employee retention period will also be identified using correlation analysis. We describe in the following section the common ML algorithms used for the employee turnover prediction problems.

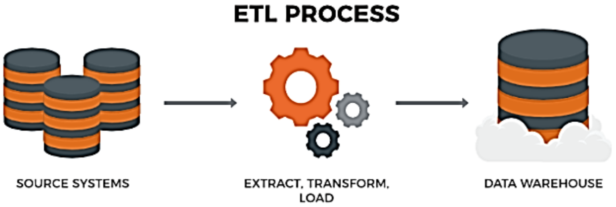

Figure 2: ETL Process in Data Science Life Cycle

A. Logistic regression (LR)

LR as represented in equation 1, is a supervised ML algorithm that models binary response variables, where the values are 1/0, male/female, yes/no [18]. To avoid overfitting problems, LR is normally used with regularization in the form of penalties based on L1-norm or L2-norm [19]. LR could be treated as a special case of linear regression, where the log of odds will be used as a dependent variable to get the categorical outcome. In other words, the occurrence of the event will be predicted by fitting the data to logit function or log-odds [20]. LR could be used to predict the turnover of an employee. [19] Opined that the logistic regression method outperformed other ML algorithms used in training run time based on their model results.

$$Odds = e^{\beta_0+\beta_1 x_1+\ldots+\beta_n x_n} \xrightarrow{applying\ log} \log(Odds) = \beta_0+\beta_1 x_1 + \cdots + \beta_n x_n \quad (1)$$

B. Support Vector Machines

SVM is a type of supervised ML algorithm based on a statistical learning algorithm. It has been widely employed for pattern classification and regression problems [12]. SVM draws a hyperplane to separate the classes based on the support vectors, where the support vector is the closest data point to the hyperplane (see Figure 3).

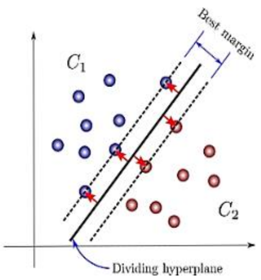

Figure 3: SVM with Hyperplane and Margin [12]

The margin that has the maximum distance from the support vectors to the hyperplane is considered the most optimum hyperplane. For non-linear classification, SVM could use kernel trick to implicitly map their inputs into high-dimensional features spaces [21]. [12] underscores the SVM model to having higher total prediction accuracy of 84.38% on employee turnover based on job performance, compared to the logit model and probit model in which both models have 71.9% of total prediction accuracy. Thus, from the result, they conclude that the SVM model has a better generalization ability to predict employee turnover as compared to other aforementioned models.

C. K-Nearest Neighbor (KNN)

KNN is an ML algorithm that could be used for both classification and regression predictive problems [22]. KNN is known as a lazy learner as it does not require learning from the data. It makes a prediction based on the closest distance of the new example with the training dataset [23]. In KNN, the k value indicates the number of nearest neighbours that will take a vote from (see Figure 4). By increasing the value of K, the class boundaries that separate classes will become smoother [22]. [23] revealed in their paper that though KNN is not the optimal model compared to other algorithms used, overall it has quite good performance metrics such as high accuracy, high precision, high recall, and high f-measure.

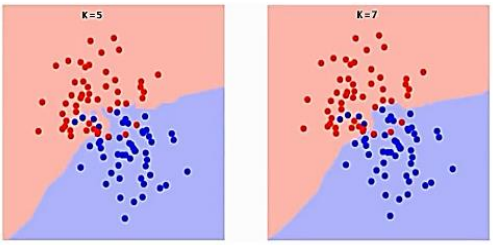

Figure 4: Effect of K-value on KNN [22]

D. Decision Tree (DT)

DT is one of the supervised learning algorithms that has been widely used in classification problems. It applies to both categorical and continuous variables. There are two types of DT, which are categorical variable DTs and continuous variable DTs. Tree representation is used to solve the problem, where each internal node of the tree represents an attribute, which each left node represents a class label as shown in Figure 5 [24]. C4.5 classifiers is used in Human Talent Prediction. Based on the result analysis, a great potential for performance prediction was gained. Thus, it can be used to determine the potential of the employee to be promoted based on their performance as stated in [25].

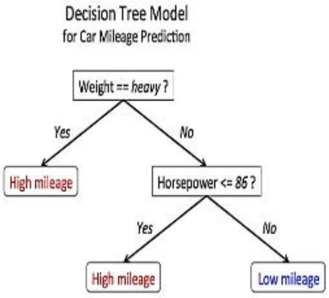

Figure 5: Decision Tree Model Diagram [24]

E. Naïve Bayes

Naïve Bayes is a supervised and probabilistic ML algorithm that is used for classification problems. The principle of the Naïve Bayes Classifier is based on the Bayes Theorem as stated in equation 2. It is so-called naïve as the features are independent and do not affect the other. There are three types of Naïve Bayes classifier, which are Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Gaussian Naïve Bayes [26]. In [27], Naïve Bayes was used to predicting the sales agents' performance of a call centre exclusively to activities of sales and telemarketing.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

A. Random Forest (RF)

RF is an ensemble model which is made up of many decision trees. There are two main concepts used by the RF in making the prediction, which is the forest made by random sampling of training data point and secondly each DT in a forest considers a random features' subset. Prediction is made by averaging all the estimations made by the individual DT. Since the prediction made by the random forest is based on the average of all individual decision trees, the result will be more accurate compared to the result made from a single decision tree. However, the more diversity of the random forest, the more robust the overall prediction is made [28]. [23] revealed in their paper that the RF classifier has the highest accuracy and precision made on employee attrition prediction compared to other ML algorithms used based on the HR analytics dataset.

IV. METHODOLOGY

A. Data

The data could be collected through different methods, either via surveys, interviews, web scraping tools such as beautifulsoup, or through open datasets from online such as Kaggle. We used in this paper the employee dataset obtainable from Kaggle as stated in [29].

B. Missing Values and Feature Selection

For the data cleaning process, the missing values are replaced with median or mode based on the nature of the features, either continuous or ordinal. The median value is selected to replace the missing value for continuous features. From the dataset that has been collected, the graph of the features skewed right and their statistics show their mean greater than their median. Therefore, the median has become a better option compared to the mean. For the ordinal feature, the mode value is the best measure of central tendency.

Various methods can be applied for feature selection such as wrapper methods, filter methods, and embedded methods, etc. We used the embedded method and correlation with the target method. The ML algorithm performs the feature selection process itself in embedded methods. It takes advantage of filter and wrapper methods by having consideration of the interaction of features like wrapper methods, shorter processing time and higher accuracy than filter methods, and less prone to overfitting [30]. Through the embedded methods, important features are identified through importance of tree-based features. Combined with correlation analysis with the target, features that show higher correlation to target and are important are selected to train the ML models.

C. Machine Learning Models Selection

The different split ratios of 80/20, 70/30 and 60/40 train-to-test sets were used with 10-fold cross-validation to build the classifier models. Besides, supervised ML models used in ETP applications are the ML models suggested in [19], [23] regarding employee attrition. However, the ML models used in those papers were classification models, therefore, to fit with our context, the classification models were replaced with the regression models. Supervised ML can be categorized into regression and classification models. The classification model is used to predict the discrete or categorical label, while the regression model is used to predict the numerical or continuous label. For the ETP application, a regression model is selected, as the main objective is to predict the retention period of the employee, where the retention period is a continuous value. Metrics used to evaluate the accuracy of the regression models are R^2 (RSquared), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) in equations 1, 2 and 3. R^2 is used to measure the closeness of data to the fitted regression line. It measures the variation of the dependent variable explained by the independent variable in the regression model. In general, the model is better fitted to the data when it has a higher R^2 value. RMSE and MAE are used to measure the average magnitude of the error. The difference between RMSE and MAE is RMSE tends to penalize large errors more on outliers and avoids the use of absolute value, which is undesirable in many mathematical calculations [28] as in equations 3, 4, and 5.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y}_i)^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \tag{5}$$

D. Employee Turnover Prediction (ETP) Application

We proposed a prediction application system that predicts the retention period of the employee and identifies the main factors that influence the employee turnover and employee retention period. The prediction system is trained with historical data of employees. Different ML algorithms are implemented in the ETP application. Each ML accuracy and other performance metrics evaluated is scored to identify the optimal model. Hyperparameter optimization on each model is also considered. The system has three actors which are admins, applicants and employees. Admins could use Power BI Desktop to perform data analysis such as univariate analysis, bivariate analysis, or multivariate analysis. These analyses allow the employer to get a better insight of the factors that cause employee turnover. The dashboard is created by the employer to share the analysis of their important results with others. Most researchers solely depend on packages such as Plotly and Seaborn for data visualization, whereas we considered the integration of Power BI Desktop, which is a business intelligence tool that can build data models, create reports, and share the work by publishing to the Power BI with the selected ML algorithms.

Figure 6 describes the architectural framework of the proposed ETP based on the System Development Life Cycle (SDLC) approach, a software development standard implemented in the design.

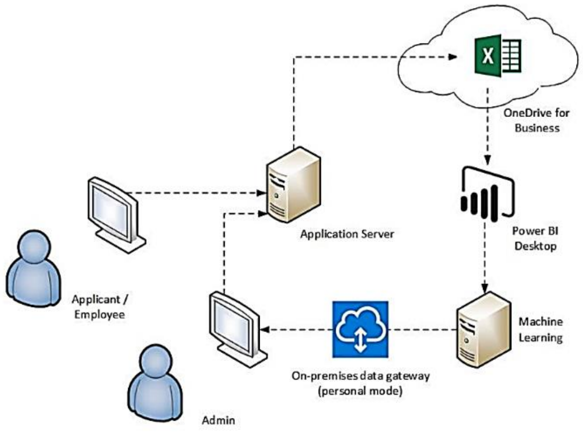

Figure 6: System Design Architecture of ETP Application

From the user interface, users submit their information to the server using the preparation methods and store results into excel in OneDrive for further business analysis. From that excel, the data is then passed through Power BI Desktop for running the python script with optimal machine learning model obtained to generate a prediction of the employee retention period. The prediction generated is transferred and visualized at Power BI through an on-premises data gateway (personal mode). Using Power BI Desktop, an analysis report is created through visualization with different charts. The main factors for employee turnover and employee retention period are visualized through Python Visual. Lastly, a dashboard is created at Power BI by using the published analysis report. The ETP application was developed and tested with a minimum requirement of Intel i5 processor and 8GB of RAM, Jupyter Notebook (version 6.0.3), Power BI Desktop (version: 2.81.5831.821 64-bit), Power Apps (version: 3.20053.18), On-premises data gateway (personal mode) (version number: 3000.37.35), Microsoft Excel 2016, OneDrive for Business, Python (3.7.3), VBA and M.

E. Data Visualization Method

Data visualization can be performed in different ways, either through creating charts in web apps or by using data visualization tools. In the ETP application, Power BI Desktop is selected as a data visualization tool as a wide range of custom visualization are provided and interactive charts can be created efficiently and effectively. It also integrates with Python to create interactive charts. Furthermore, it provides flexibility in accepting various types of data sources such as excel, web, database, etc. We performed feature engineering in Power BI Desktop by modifying the features in Power Query Editor, such as transforming the data or running python scripting. We also implement Power Query by using M language, the dashboard for overview reporting and Power Apps for creating a form for information management.

V. Results and Discussion

Although various testing including unit testing, integration testing and system testing for our developed ETP application were performed, we concentrate on presenting the result of ML and Power BI of the ETP.

Through the collected dataset, the main factors that caused employee turnover are identified through correlation analysis. Figure 7 shows the main factors that caused employee turnover as follows: OverTime, BusinessTravel, EnvironmentSatisfaction, JobSatisfaction, JobInvolvement, YearsAtCompany, StockOptionLevel, YearsWithCurrManager, Age, YearsInCurrentRole, JobLevel, TotalWorkingYears as shown in Figure. Based on the main factors identified, these factors can be classified as factors that caused employee turnover voluntarily. OverTime has the highest positive correlation score, which shows that most of the employees are reluctant to work overtime. It can be understood that working overtime will increase the stress levels of the employees as they probably are required to complete their task within the due date. If the employees are requested to work overtime frequently, probably the chance for the employees to turnover will be very high. On the other hand, TotalWorkingYears has the highest negative correlation score, which indicates that employees with fewer total working years tend to turn over easily. The reason that may explain this situation is probably the employees are still seeking for the right company to work.

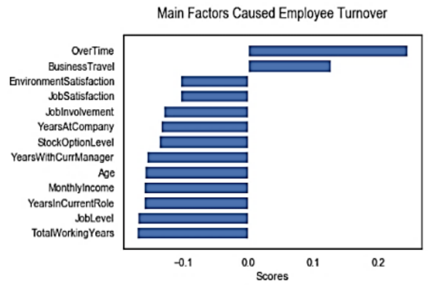

Figure 7: Bar Chart of Main Factors Causing Employee Turnover

Besides, as shown in Figure 8, the identified factors that affect employee retention period include YearsWithCurrManager, YearsInCurrentRole, TotalWorkingYears, YearsSinceLastPromotion, JobLevel, MonthlyIncome, Age, NumCompaniesWorked. YearsWithCurrManager shows the highest correlations to the retention period followed by YearsInCurrentRole. This implies that employees that are loyal to their manager tend to have a longer retention period as the bond between them has been formed. Employees who have more experience in their current role also tend to have a longer retention period as they are familiar with the job, and possibly gain more knowledge about the job and hone their job skills. On the other hand, NumCompaniesWorked has the highest negative correlation score, which implies that employees with a low number of companies worked tend to have a higher retention period. This can be explained through employees who may be loyal to the company or willing to spend more time learning new knowledge in their company rather than keep switching to the other company.

Initially, RF Regressor and Gradient Boosting (GB) Regressor are chosen as the ML models for hyperparameter optimization. These ML models outperformed other ML models such as KNN Regressor, DT Regressor, and LR on cross-validation results (see the boxplot in Figure 9). RF Regressor scored 0.94 R^2 mean and -1.5 RMSE mean, while GB Regressor scored 0.92 R^2 mean and -1.68 RMSE mean. RF Regressor and GB Regressor scored higher on cross-validation results as these models are ensemble methods that combine several weak learners into strong learners, thus increasing the accuracy of the model. These ML models also tend to have higher scores on a larger splitting ratio, which is 80/20 compared to 70/30 and 60/40. With a larger training split, the model can train on more data, thus improving the accuracy of the model. The overfitting problem of the training model is solved through cross-validation. Cross-validation with 10 folds is performed to provide a more accurate indication of the ML model's generalization to other parts of data through repeating 10 times on different parts of the training set as the validation set.

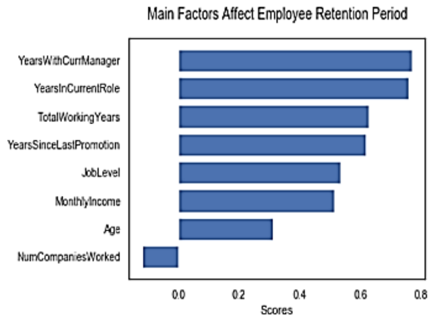

Figure 8: Bar Chart of Main Factors Affecting Employee Retention Period

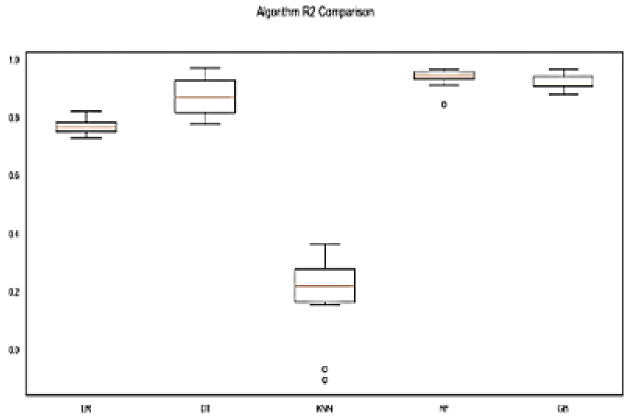

Figure 9: Boxplot of R^2 scores of Cross-Validation Results on 80/20 Splitting Ratio

RF Regressor is finally chosen as the predictive model during hyperparameter optimization. By running GridSearchCV on RF Regressor and GB Regressor, RF Regressor got 0.95 for R^2 score and used 0.2Hr for the time taken, while GB Regressor got 0.94 for R^2 score and used 5.9Hr for the time taken. GB Regressor took a long time as it builds tree sequentially and combine results along the way (Boosting) while RF Regressor builds tree independently and combine results at the end of the process (Bagging). The time taken for the model is affected by the parameters n_estimators and learning_rate. In this case, RF Regressor score better in terms of R^2 score and time-taken. The results of RF Regressor, GB Regressor and other classifiers are presented in Table 1. Parameters that are used for fine-tuning by RF Regressors are n_estimators, min_samples_split, min_samples_leaf, and max_depth. Among three times hyperparameter optimization on RF Regressor, although third tuning scores lower compared to first and second tuning, however, the max depth of the parameter shows lower on third tuning compared to first two tunings as presented in Tabel 2 and 3. By having a lower max depth of RF Regressor, it can avoid overfitting of the model to the training set. Parameters are adjusted according to the result of the previous tuning during hyperparameter optimization. First tuning shows an identical score to second tuning in terms of best score on train set and R2 score on the test. This phenomenon occurred because that parameter shows optimal results compared to the other parameters used for hyperparameter optimization. The first two tunings show higher $\boldsymbol{R^2}$ score on the test set than third tuning because the max_depth value is higher in the first two tunings. Normally, a higher max_depth value will increase the fitting of the training model, thus yielding a higher $\boldsymbol{R^2}$ score. However, too high a max_depth value will lead to overfitting of the model and probably will perform poorer prediction on the test set. Therefore, parameters with {'max_depth': 13, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 125, ‘random_state’: 0} is accepted as optimal parameters for RF Regressor in predicting employee retention period.

Table 1
Cross-validation Result of ML Models on 80:20 Splitting Ratio

	Algorithm	$\boldsymbol{R^2}$ mean	$\boldsymbol{R^2}$std	RMSE mean	RMSE std	MAE mean	MAE std
0	LR	0.77	0.03	-2.90	0.19	-1.88	0.11
1	DT	0.87	0.06	-2.07	0.61	0.92	0.22
2	KNN	0.18	0.14	-5.44	0.35	-3.80	0.25
3	RF	**0.94**	0.03	**-1.50**	0.38	-0.76	0.11
4	GB	0.92	0.03	-1.68	0.27	-0.94	0.09

Table 2
R^2 Score of RF Regressor on Further Hyperparameter Optimization

	1st Tuning	2nd Tuning	3rd Tuning
RF-10f R^2score	0.954144	0.954144	0.952269

Table 3
Cross-validation Result of RF Regressor after Further Hyperparameter Optimization

	Algorithm	R^2 mean	R^2std	RMSE mean	RMSE std	MAE mean	MAE std
0	RF 1st Tuning	0.94	0.03	-1.5	0.37	-0.76	0.11
1	RF 1st Tuning	0.94	0.03	-1.5	0.37	-0.76	0.11

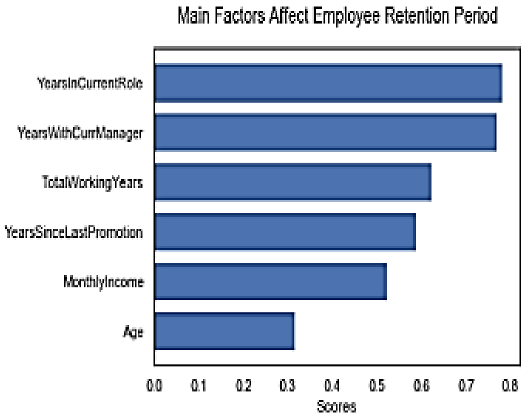

Figure 10: Correlation of Main factors affecting Employee Retention Period

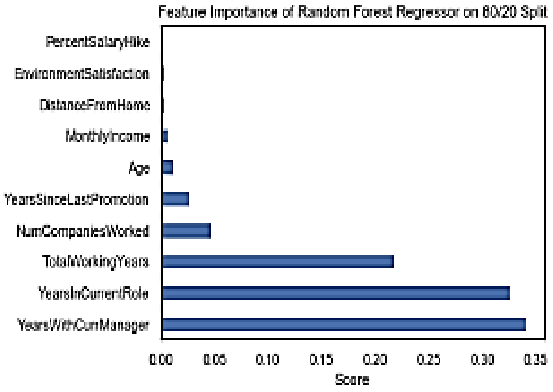

Figure 11: Feature Importance of RF on 80/20 split.

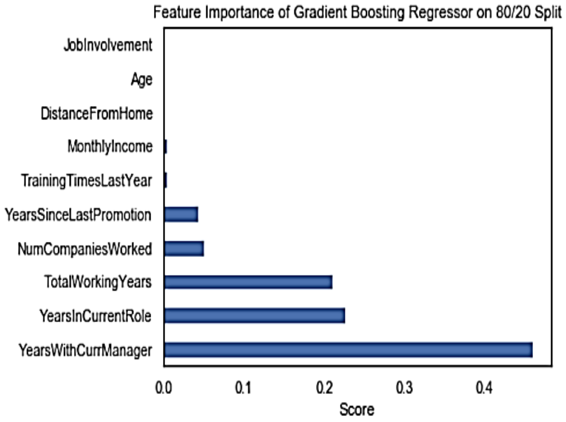

Figure 12: Feature Importance of GB on 80/20 split.

Different analyses were visualized through chart types available in Power BI Desktop or a Power BI Desktop marketplace that provides more chart types that are not pre-installed. Python Visual in Power BI Desktop supports visualization of the data through Python. It is especially useful when certain analyses can only be performed through Python code, such as the correlation analysis and feature importance of the RF model. Through data visualization, employers could obtain insights that are useful to their organization. The rest of the visualization results of our work is presented in appendix I-Figures 13-16.

VI. CONCLUSION

We have designed an ETP application and performed data exploration to identify data distribution. A suitable value for imputing missing value is identified through identifying skewness of the data distribution and the quantitative or qualitative data. The main factors that caused employee turnover are identified through correlation analysis (see Figure 11 and 12 for RF and GB). Features that showed a higher correlation to turnover features are ranked and visualized. Main factors affecting employee retention period are also identified through correlation analysis. Besides, for the feature selection process, the embedded methods are used together with the correlation analysis with the target to identify the suitable features for predictive modelling (see Figure 10). ML models are cross-validated with different splitting ratios and different features to determine the suitable ML models for further analysis. Hyperparameter optimization is performed to identify the optimal ML model and optimal parameters for predicting the retention period. Lastly, data visualization is performed by generating different charts for analysis and creating a dashboard an overview. New measures can be created with Power Query and visualized through Power BI Desktop. An example of a new measure created in the ETP application is the number of employees with early turnover, where the employees spent not more than a year in the company and left. Early turnover is one of the useful metrics for HR analytic. Our RFregressor+PowerBI model with appropriate hyperparameter optimization has an overall performance score of 0.94 on $\boldsymbol{R^2}$ mean and -1.5 on RMSE mean, while the optimal parameters for predicting employee retention period are identified. For future work, the dataset used in this ETP application can be considered by collecting through the data from industry. More employee information and features can be collected for further experiments. Besides, different methods of feature selection such as Recursive Feature Elimination (RFE) can be applied to identify the optimal features for predictive modelling. Finally, deep learning techniques such as neural networks can also be considered to predict the retention period of the employees.

APPENDIX

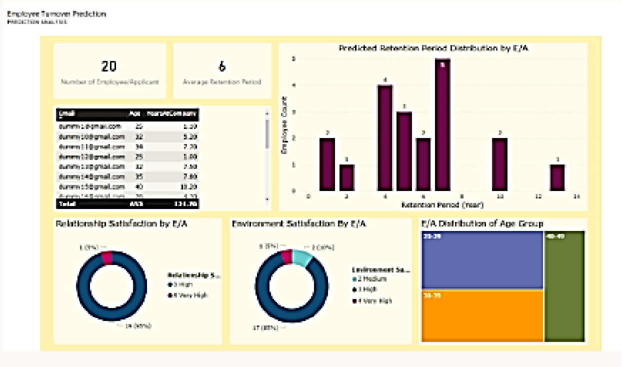
Employee Turnover Prediction
PREDICTION ANALYSIS
20
Number of Employee/Applicant
6
Average Retention Period
Email
Age
YearsAtCompany
Total
Predicted Retention Period Distribution by E/A
Employee Count
Retention Period (Year)
Relationship Satisfaction by E/A
Environment Satisfaction By E/A
E/A Distribution of Age Group
3 High
4 Very High
2 Medium

Figure 13: Screenshot of Prediction Page

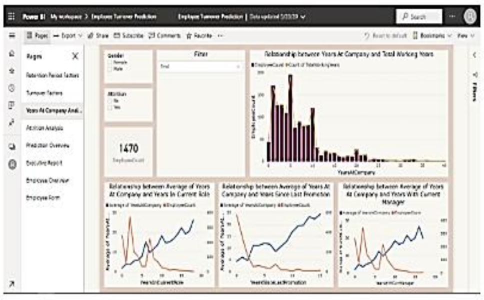

Figure 14: Screenshot of Analysis Report Page

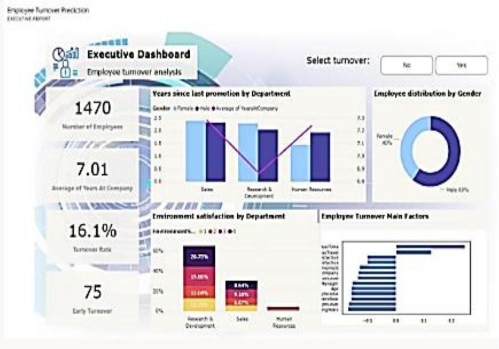
Employee Turnover Prediction
EXECUTIVE REPORT
Executive Dashboard
Employee turnover analysis
Select turnover:
No
Yes
1470
Number of Employees
7.01
Average of Years At Company
16.1%
Turnover Rate
75
Early Turnover
Years since last promotion by Department
Employee distribution by Gender
Environment satisfaction by Department
Employee Turnover Main Factors
Sales
Research & Development
Human Resources

Figure 15: Screenshot of Executive Report Dashboard Page

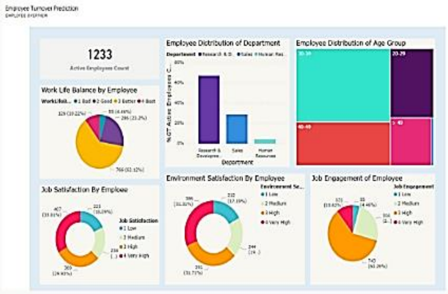
Employee Turnover Prediction
1233
Active Employee Count
Employee Distribution of Department
Employee Distribution of Age Group
Work Life Balance by Employee
Job Satisfaction By Employee
Environment Satisfaction By Employee
Job Engagement of Employee

Figure 16: Screenshot of Employee Information Overview Dashboard Page

ACKNOWLEDGMENT

This work is supported by the RUI (Individual) Grant by the Universiti Sains Malaysia (USM) under Grant No. 1001/PKOMP/8011136 for the project entitled "A New Adaptive Kernel-Distance-Based Control Chart by Machine Learning Techniques to Support Industry 4.0 Sustainable Smart Manufacturing".

REFERENCES

[1] H. Boushey and H. J. S. Glynn, “There Are Significant Business Costs to Replacing Employees,” *Center for American Progress*. November 2012. https://www.americanprogress.org/issues/economy/reports/2012/11/16/44464/thereare-significant-business-costs-to-replacing-employees/

[2] Identifying and Addressing Employee Turnover Issues. (n.d.). *Wolters Kluwer*. https://www.bizfilings.com/toolkit/research-topics/office-hr/identifying-andaddressing-employee-turnover-issues.

[3] K. Martinelli, “Causes of Employee Turnover and Strategies to Reduce it,” *High Speed Training*. October 13, 2017. https://www.highspeedtraining.co.uk/hub/causes-ofemployee-turnover/

[4] D. Whitelegg, “How do Recruitment Agencies Get Paid (and How Much),” *Agency Central*. Retrieved October, 2016. https://www.agencycentral.co.uk/articles/2016-10/howrecruitment-agencies getpaid.htm#targetText=The%20cost%20of%20a%20recruitment,for%20hard%20to%20fill%20positions

[5] Alienor, “What is a Data Silo and Why is It Bad for Your Organization?” *Plixer*. 2018. https://www.plixer.com/blog/data-silo-what-is-it-why-is-it-bad/

[6] Z. A. Bilal. "Predicting customer churn in banking industry using neural networks." *Interdisciplinary Description of Complex Systems: INDECS* 14.2 (2016): 116-124.

[7] P.C Patel, et al. "Retail store churn and performance–The moderating role of sales amplitude and unpredictability," *International Journal of Production Economics* 222 (2020): 107510.

[8] G. G. Sundarkumar, R. Vadlamani and V. Siddeshwar, "One-class Support Vector Machine Based Undersampling: Application to churn prediction and insurance fraud detection," *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE*, 2015.

[9] M. N. Z Milošević and A. Igor, "Early churn prediction with personalized targeting in mobile social games," *Expert Systems with Applications* 83 (2017): 326-332.

[10] C. Günther, et al., "Modelling and Predicting Customer Churn from An Insurance Company." *Scandinavian Actuarial Journal* 2014.1 (2014): 58-71.

[11] S. H. Dolatabadi and F. Keynia, "Designing of customer and employee churn prediction model based on data mining method and neural predictor," *2017 2nd International Conference on Computer and Communication Systems (ICCCS)*, 2017, pp. 74-77, doi: 10.1109/CCOMS.2017.8075270.

[12] W.C Hong, P.F. Pai, Y.Y. Huang, and S.L. Yang, “Application of Support Vector Machines in Predicting Employee Turnover Based on Job Performance,” *International Conference on Natural Computation*. 2005. 668-674. Springer.

[13] A. H. Ali, Z. F. Hussain and S. N. Abd, “Big Data Classification Efficiency Based on Linear Discriminant Analysis,” *Iraqi Journal for Computer Science and Mathematics*. 2020. 1(1), 7-12.

[14] A. H. Ali and M. Z. Abdullah, “A Novel Approach for Big Data Classification based on Hybrid Parallel Dimensionality Reduction using Spark Cluster,” *Computer Science*. 2019. 20(4).

[15] A. H. Ali and M. Z. Abdullah, “A Parallel Grid Optimization of SVM Hyperparameter for Big Data Classification using Spark Radoop,” *Karbala International Journal of Modern Science*. 2020. 6(2), Article 3.

[16] A. Huber, “Staff Attrition vs. Staff Turnover: What's the Difference?” *Jobzology*. Retrieved March 28, 2018. https://jobzology.com/staff-attrition-vs-staff-turnover-whats-the-difference/

[17] S. E. Schaeffer and S. V. R. Sanchez, “Forecasting Client Retention: A Machine Learning Approach,” *Journal of Retailing and Consumer Services*. 2020. 52. (C)

[18] V. Bewick, L. Cheek, and J. Ball, “Statistics review 14: Logistic regression.” *Critical care*. 2005. 9(1), 112.

[19] R. Punnoose and P. Ajit, “Prediction of Employee Turnover in Organizations Using Machine Learning Algorithms,” *Algorithms*. 2016. 4(5), C5.

[20] P. Chandrayan, “Logistic Regression for Dummies: A Detailed Explanation,” *Towards Data Science*. Retrieved August 5, 2019. https://towardsdatascience.com/logisticregression-for-dummies-a-detailed-explanation-9597f76edf46.

[21] N. Sharma, “People Analytics with Attrition Predictions,” *Towards Data Science*. Retrieved May 18. https://towardsdatascience.com/people-analytics-with-attritionpredictions-12adcce9573f

[22] T. Srivastava, “Introduction to KNN, K-Nearest Neighbors: Simplified,” *Analytics Vidhya*. Retrieved March 26, 2018.

[23] D. S. Sisodia, S. Vishwakarma, and A. Pujahari, “Evaluation of Machine Learning Models for Employee Churn Prediction,” *International Conference on Inventive Computing and Informatics (ICICI)*. 2017. 1016-1020. IEEE.

[24] R. S. Brid, “Decision Trees‐A simple way to visualize a decision Medium,” 2018. https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-adecision-dc506a403aeb

[25] H. Jantan, A. R. Hamdan, and Z. A. Othman, “Human Talent Prediction in HRM Using C4.5 Classification Algorithm,” *International Journal on Computer Science and Engineering*. 2010. 2(8), 2526-2534.

[26] R. Gandhi, “Naive Bayes Classifier,” *Towards Data Science*. Retrieved May 6, 2018. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.

[27] M. A. Valle, S. Varas, and G. A. Ruz, “Job performance prediction in a call center using a naive Bayes classifier,” *Expert Systems with Applications*. 2012, 39(11), 9939-9945.

[28] W. Koehrsen, “Random Forest Simple Explanation. Medium. Retrieved December 28, 2017. https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d.

[29] Pavansubash, “IBM HR Analytics Employee Attrition & Performance,” *Kaggle*. Retrieved March 31, 2017. https://www.kaggle.com/pavansubhasht/ibm-hr-analyticsattrition-dataset

[30] Y. Charfaoui, “Hands-on with Feature Selection Techniques: Embedded Methods,” *Medium*. Retrieved January 18, 2020. https://heartbeat.fritz.ai/hands-on-with-featureselection-techniques-embedded-methods-84747e814dab.