2021 21st International Conference on Advances in ICT for Emerging Regions (ICter) | 978-1-6654-6686-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICter53630.2021.9774825

Machine learning approach for predicting career suitability, career progression and attrition of IT graduates

B.M.D.E Bannaka*, D.M.H.S.G Dhanasekara†, M.K Sheena‡, Anuradha Karunasena§ and Nadeesa Pemadasa‖
Department of Information Technology, Faculty of Computing, Sri Lanka Institute of Information Technology, Sri Lanka
Email: *it18074796@my.sliit.lk, †it18188028@my.sliit.lk, ‡it18187106@my.sliit.lk,
§anuradha.k@sliit.lk, ‖nadeesa.p@sliit.lk

***Abstract*—The IT industry in Sri Lanka is associated with a massive work force consisting of skillful professionals and it also provides many job opportunities for fresh graduates at the present. For a fresh graduate entering the IT industry there is a wide variety of job opportunities available and in order to have a satisfactory and rewarding career they should identify the most suitable career for them. On the other hand, employees change their careers and regularly seeking for career advancements and more benefits while the employers struggle to retain employees. Under such circumstances, this research focuses on developing a career mentoring system which comprises of the prediction of career suitability, career and salary progression, and employee attrition to assist IT employees to achieve career goals by overcoming barriers in their career path. For this purpose, data are collected from IT employees, and several models were implemented using classification algorithms such as XGBoost, Random Forest, Support Vector Machine, K-Nearest Neighbors, Decision tree, Naive Bayes, and their performance are compared using accuracy, precision, recall, and F1-Score to select accurate models. XGBoost resulted with higher accuracies for prediction of career suitability, initial salary, career and salary progression with values of 92.31, 90.35, 86.45 and 88.76 respectively. Furthermore, for the prediction of professional courses and employee attrition, Random Forest resulted higher accuracies of 93.52 and 89.70. The ultimate goal of this research is to guide IT graduates and employees to have better performances and to assist them in embracing responsibilities throughout their career life.**

Keywords- attrition, career path, classification algorithms, progression, suitability

I. Introduction

Due to the enhanced upliftment of the IT industry, many higher educational institutes including those of state and non-state sectors are offering IT degree programs. Many students are enrolled in such programs. As per the University Grants Commission (UGC) of Sri Lanka's statistics in 2019, 12510 undergraduates have enrolled in the state universities to follow IT-related degree programs [1] and it was stated that 146,089 employees exists in the IT industry in 2019 according to a survey conducted by Information and Communication Technology Agency (ICTA) [2].

Currently, the IT industry is gaining popularity among the prevailing industries by attracting young individuals and talents to cooperate in its activities [3]. These provide a gateway to a pool of opportunities for the IT employees to achieve professional success since it consist of a wide range of job titles. A few of the job titles include Software Engineer, Quality Assurance Engineer, and Business Analyst [4], where each career path possesses a unique combination of skills. For example, a Software Engineer must possess strong skills like knowledge on programming languages and concepts, problem-solving, the ability to work as a team, and multitasking [5].

The upcoming generation's desirability towards higher studies in IT varies due to certain impacts, such as current trends and social influences from companions. This has affected the motivation of the IT undergraduates eventually causing them to inhibit their academic performance [6]. Furthermore, during their studentship they are not aware of the required skills to be developed to be a suitable candidate for a certain job role and what are the job roles they will be suitable for with their current level of skills. Therefore, IT graduates face setbacks after entering the industry under the prevailing competition [7]. This may even cause them to change their career path at the later stages due to the inability in performing the expected role [8].

Every employee desire to rise up faster in their respective career paths and enjoy the monetary benefits and satisfaction. Progression in the career of some employees is slow due to various factors such as lack of updated technical skills, non-technical skills, business knowledge sets, and disciplines [9]. Under such circumstances, the unpredictability of the number of years taken to promote from the current career stage to upcoming job stages in the organizational hierarchy and dissatisfaction of increments received with respect to the career promotions has caused the employees to be frustrated and demotivated [10]. Furthermore, employees retain in the same career position and are unable to reach their desired positions due to their inadequate knowledge of certain professional and technical skills to be enhanced within themselves.

IT industry has become a dynamic industry where employees change their employers quite frequently [12]. Employers spend much time and money in training employees and giving them promotions, which might not be worthy depending on the duration the employee has stayed employed under the company [13]. Furthermore, employers are unaware of the steps to be practiced in retaining the employees within the organization. Therefore, the employers face hardships while submitting the products on time due to the loss of required

978-1-6654-6686-8/21/$31.00 ©2021 IEEE

staff. Eventually, this causes the organization to incur losses and bad reputation.

Existing research has proposed various productive solutions for career suitability and progression prediction in a particular career path [5] - [11]. However, these research has evaluated limited aspects for IT career-related activities such as prediction of the initial career and career progression. Therefore, such research has provided limited assistance to IT graduates and undergraduates to improve themselves for better career opportunities and to identify suitable career paths according to their skills [8]. Moreover, for the prediction of attrition, there exists a limited number of solutions specific to the IT industry along with suggestions of remedial actions for the IT organization to retain their IT employees who have a higher possibility of leaving the organization in a short period of time [12], [13]. The aims of the present study are to implement a comprehensive solution for predicting the most suitable initial career and salary range; to predict their career progression as well as associated salary range progression and recommend professional courses to the employees in the field of Software Engineering; and to provide guidance to employers and organizations for taking remedial actions to retain employees for a longer period. To achieve these objectives initially data was collected from the existing IT employees. Then machine learning algorithms such as XGBoost, Random Forest, Support Vector Machine (SVM), Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Decision tree, and Naive Bayes were practiced on preprocessed data and compared their performances using accuracy, precision, recall, and F1-Score for selecting the most appropriate algorithm. Eventually, a web-based Responsive application was implemented by utilizing the selected algorithms for accomplishing the above-mentioned predictions. Overall, this solution could be used as a career mentor for IT employees throughout their career life.

II. Literature Review

Numerous research have been conducted using various approaches on predicting career suitability and progression in the field of IT. Key research found among such research are summarized below.

A career in which an individual can perform the best is the most suitable career path for that individual. Existing research has attempted to identify the most suitable careers for individuals in IT field. For an Example, Arafath et al. [6] carried out a study to assist the university administration to gain an accurate understanding regarding the Computer Science undergraduates by mining their academic, technical and interpersonal factors for predicting an appropriate career for them. The above research used multiple classification techniques such as Iterative Dichotomiser 3 (ID3), Classification and Regression Trees (CART), Random Forest, SVM and Multilayer Perceptron (MLP) Neural Networks to predict the career path of a student residing in the final year by analyzing the successful alumni data which was collected through a survey. A higher accuracy with a value of 95.24% was resulted for MLP. Furthermore, the study concluded that the success of the IT graduates does not only depend on their academic skills but also their technical performances, personal skills and their social skills. Roy and Roopkanth [7] considered the skills, talents, capabilities and interests to recommend a suitable role in a particular career path while recruiting employees. When recruiting candidates to IT organizations, candidates are thoroughly evaluated and the most appropriate roles in a selected career area are recommended. Machine learning algorithms were used to train a model to predict roles and the results of the research showed an accuracy of 88.3% for XGBoost and an accuracy of 90.3% for SVM. Therefore, SVM was used for the implementation of a web application to assist employers in recommending career paths for employees. Furthermore, Daharwal et al. [8] conducted a study to assist students in selecting a career and a field of employment based on their skills. The above research used several machine learning techniques to train and test a model where Naive Bayes algorithm resulted with a high accuracy of 93 %. Based on the developed model the researchers developed a career guidance system.

An employee's progression is determined by the professional practices and the performances he demonstrates. Therefore, they are often used to predict employee progression. For example, Liu et al. [9] conducted a study to predict career progression of an employee in a particular career path by using Multi-Source Learning Framework Lasso Penalty (MSLFL) model. The model was applied for career-oriented features including demographic features which were extracted from social networks. In this research MSLFL turned to provide a high accuracy than SVM and Regularized Least Square (RLS). In a study conducted by Li et al. [10] modelling a talented individual's career path was done where a model was trained to predict the career progression as well as the turn over. The research was conducted using data which were collected for a period of 48 months. Although, convectional survival models could not be directly used for the prediction of career progression in this research, they were utilized for depicting the changing event of a career stage for an employee. Furthermore, multitask learning based baseline methods (M-LASSO, M-L2,1, and MTLSAV2) were also used in this research to predict the relative career stage at a particular time interval. Then by adding ranking constraints to baseline methods, performance of the models was improved and comparison of methods used in the research revealed that the highest efficiency was provided from MTLSAV2+DF with a value of 83%.

Employee attrition is a key issue that needs to be addressed by organizations for a better productivity. A study was conducted by Punnoose and Ajith [12], to identify the most accurate prediction model for employee attrition. The Human Resource Information System (HRIS) data of the employees in a company located in Unites States which consisted of noisy demographic dataset were used to train the models by fitting on seven classifiers such as XGBoost, Logistic Regression, Naive Bayes, Random Forest, KNN, Linear Discriminant Analysis (LDA) and SVM. Upon comparison and evaluation of these classifiers it was found that, XGBoost outperformed with a

higher value of 0.88 for Area under the receiving operating characteristic Curve (AUC) with lower runtime and efficient memory utilization. Therefore, XGBoost was considered as best algorithm for the attrition prediction and implementation of real-life problems which also works well for noisy data due to its regularization capability. Alduayj and Rajpoot [13] also conducted a research where a model was developed to predict attrition of employees. Dataset used for the research consisted of HR data created by IBM Watson which contained imbalanced data and therefore two approaches were followed. Firstly, the original imbalanced dataset was trained using SVM, Random Forest and KNN where SVM scored highest F1-Score of 0.50. Secondly, Adaptive Synthetic Sampling (ADASYN) approach was applied to balance the two classes where Random Forest obtained F1-Score of 0.90. Finally, manual under sampling was done to have equal classes which resulted in loss of information. It was then concluded that overcoming the class imbalance problem by using ADASYN approach on Random Forest provides a better attrition prediction model.

Most of the previous approaches mentioned above have paid limited attention to certain aspects such as providing guidance for IT undergraduates in exploring career paths based on their skills and qualifications. Furthermore, the existing solutions provide insufficient recommendations regarding the professional skills to be developed by employees in the field of Software Engineering which assist them to have better prospects in career progression. Moreover, the existing literature indicates the absence of suggestions for precautionary steps for employers to retain their IT employees within the organization for a longer period. To address the above research gaps, this study proposes the prediction of the initial career and its salary by recommending ways to enhance a student's capabilities during their undergraduate period. Furthermore, IT employees are recommended extra professional courses to expedite career promotions by depicting the number of years at each career stage through the organizational hierarchy while outlining the salary progression ranges. Moreover, IT organizations will be assisted in predicting the employee attrition and the duration of the employees' tenure within the organization, while suggesting remedial actions to retain the employees for a longer period. Eventually, the Career Mentoring System can be used as an all-in-one solution for IT undergraduates, employees in the field of Software Engineering, and employers to achieve career satisfaction and improvement.

III. Methodology

This section outlines the techniques carried out for predicting the initial career for IT undergraduates, career and salary progression for IT employees in the field of Software Engineering and employee attrition for employers.

A. Data Collection

To proceed with the research, firstly two survey questionnaires were designed with the intention of collecting data. Here one survey was designed for the first two components namely career suitability and progression prediction with 26 questions while the other survey for employee attrition with 16 questions. The list of attributes collected from the questionnaires are depicted in above Table I. These questions were prepared based on the feature's literature revealed related to career suitability [6], [7] and progression prediction [9], [11]. With reference to studies [12] and [13] which listed 14 features as important for employee attrition prediction, more features such as workplace relationship, organizational commitment, work satisfactory levels, management satisfactory rating, benefits and rewards were also considered with the domain knowledge for increasing the efficiency of the model. After designing the surveys, the questionnaires were published using Survey Monkey Online Tool and shared among the current IT employees with random sampling method for collecting the required data. The questionnaires were filled by 1450 employees.

TABLE I
ATTRIBUTES USED FOR CAREER SUITABILITY, PROGRESSION AND EMPLOYEE ATTRITION PREDICTION

Demographic Features	Age, Race, Ethnicity, Gender[6],[9],[11],[12],[14]
Academic Features	A/L results, University, Degree, GPA, Professional qualifications [6],[9],[14]
Technical Skills	Knowledge of: programming languages and concepts, Software Engineering Concepts [9],[14]
Soft Skills	Problem Solving, Creativity, Self-Learning, Management, English Knowledge [6],[7]
Performance in relevant fields	Front End, Back End, Full Stack, Mobile application, UI/ UX Development, Software Testing and Quality Assurance [9]
Working History	No of Years, Joined Year, Company Scale, Salary Range [12]
Job Details	Education Level and Field, Job Level and Role, Monthly Salary, Company, KMs From Home, No of Trainings and Working Years [13]
Satisfactory Levels	Environment and Job Satisfaction, Workplace Relationship, Commitment, Satisfactory levels, Management Satisfaction, Benefits and Rewards

B. Data Preparation and Pre-processing

The raw data which were collected from the IT professionals were in an unorganized format which includes null values, invalid data and redundant data. With the aim of obtaining a proper input dataset and enhancing the performance of machine learning models, imputation of missing values, removal of outliers, scaling, normalization and One Hot Encoding were used. Firstly, rather than dropping the missing values in the dataset, imputation was considered a preferable alternative. The missing numerical values were handled by imputing with the mean calculated from the non-missing values in a column. Also, the missing categorical values were replaced by the maximum occurred value in the column. Furthermore, the removal of duplicate rows and outliers detected from scatter plots were carried out with the intention of obtaining a clean dataset. Then, normalization and standard scaling methods were used

as feature scaling techniques. One Hot Encoding method was applied to convert the categorical values into numerical format. After cleaning and scaling data, dimensionality reduction was carried out by using Principal Component Analysis (PCA). The dataset resulted as above, was divided as the training and testing set, where 80% of the original dataset was used to fit into the machine learning models for training, while 20% of the dataset was used for the evaluation of the fitted machine learning model.

C. Model Selection and Training

As shown in Table II, several supervised machine learning classification algorithms were used to select the most appropriate algorithm.

The following predictions were made on the career suitability of IT undergraduates:

- Prediction of initial career (Software Engineer (SE), Quality Assurance (QA) Engineer, Business Analyst (BA)).
- Prediction of initial salary range (Below LKR 25000, 25 000 - 50 000, 50 000 - 100 000).

The predictions given below were practiced by utilizing multi-output classification algorithms for determining career and salary progression of IT employees in the field of Software Engineering:

- Prediction of number of years at upcoming career stages of an employee in the Software Engineering profession (Associate Software Engineer (ASE), Software Engineer (SE), Senior Software Engineer (SSE), Tech Lead, and Architect).
- Prediction of salary range (Below LKR 25000, 25 000 - 50 000, 50 000 - 100 000, 100 000 - 200 000, Over 200 000) with career progression.
- Prediction of professional courses or practices (Programming, React, Mobile Application Development, Web Design, Database Administration,Test Automation, Project management) to expedite one's career progression.

The following predictions were obtained for discovering an employee's attrition and retention:

- Prediction of employee attrition as to whether the individual will leave the organization in the near future (Attrition- Yes, No).
- Prediction of the retention duration of an IT employee (less than 1, 1 - 2, 2 - 3, 3 - 4, 4 - 5 years).
- Prediction of the remedial actions (increase work satisfactory, organizational commitment, benefits and rewards, management satisfaction).

To accomplish the above predictions, various classification algorithms were utilized as depicted in Table II. For SVM and SVC algorithms, several Kernels such as linear, polynomial, and radial basis function (RBF) were compared. Thereafter, the polynomial kernel was chosen as it gave the highest accuracy compared to the other kernels. Then the degree and Regularization parameter (C) with default value of 1.0 was determined. Furthermore, to find out the optimal k-nearest neighbors (k value) of the KNN classifier, plots were generated between error rate and k value. Accordingly, for ID3 algorithm of Decision tree, the criteria including entropy and Gini were compared for their performances. Moreover, the splitter and maximum depth were set. Also, Multinomial Naive Bayes, a probabilistic learning method based on the Bayes theorem, was practiced for achieving speedy predictions. By using Random Forest Classifier, many trees were generated each with equal weights of leaves. Thereafter, n-estimators (the number of trees in the forest), the type of the criterion (Gini and entropy), and the maximum depth of the tree were compared with the aim of determining the better performing conditions. Finally, the implementation of gradient boosting decision trees was done by using XGBoost Classifier. Comparatively, the learning rate and the n-estimators were set to discover their optimum accomplishing conditions.

TABLE II
USED CLASSIFICATION ALGORITHMS

Prediction	Classification Algorithms
Initial career and salary range	XGBoost, Decision Tree, SVM, Random Forest, KNN, Naive Bayes
Employee attrition, retention and remedial action	
Career and salary range progression professional courses	XGBoost, Random Forest, KNN, SVC

D. Evaluation

Classification Reports, Confusion Matrix, Receiver Operation characteristic (ROC) Curve and Local Interpretable Model-agnostic Explanations (Lime) were used to evaluate the performance of the algorithms. Classification reports consisting of accuracy, precision, recall and F1-score were generated to evaluate the performance of the algorithms. Precision of classifiers demonstrates the accuracy of positive predictions [14] and recall has the ability to find all positive instances of a classifier [15]. F1-score is calculated with the combination of precision and recall of the predictive model, while assisting to compare the performance of each model [15]. Confusion matrix was also used to conceptualize the performance of each classification models on the test data. Furthermore, for the visualization of the output performance, AUC and ROC which denotes the True Positive (TP) rates against False Positive (FP) rate were acquired. Trained machine learning models were interpreted using LIME which has the potential of explaining any black box classifiers. Since these techniques have the capability for the application in binary classification, multi-class classification and multi-output classification it was effectively applied for the evaluation of the research objectives.

E. Deployment of the models

Machine learning models were deployed on flask by creating Representational state transfer (REST) Application Programming Interface (API) which receives inputs from the end users and provide the predictions as an output.

IV. Result and Discussion

This section describes the result and discussions of the study under three main components with their classification models and their respective performances.

A. Result Analysis of Prediction of initial career and salary

To predict the initial career and salary of IT undergraduates, a preprocessed dataset was trained using six different classification algorithms as shown in Table II. Among the considered trained models, XGBoost provided a better testing accuracy of 92.31% for the prediction of initial career, and 90.35% for the prediction of the initial salary range respectively as shown in Table III.

Figure 1 represents the classification report for model performance of initial career prediction, and it has three classes (1 – BA 2 – QA, 3 – SE). As given in the support column, there are more samples for SE careers than BA and QA careers. Therefore, when each class is considered individually with one vs all approach, class 1 and 2 become minority classes and class 3 becomes the majority class. Therefore, precision should be more focused when evaluating the performance with regard to class 1 and 2 while recall should be focused when evaluating the performance with regard to class 3. Here, the precision is less than the recall with regard to class 1 and 2 while it is higher than the precision with regard to class 3. However, all three classes have F1-score over 0.85, and model accuracy is greater than 0.90, which shows that this model has a good performance.

As shown in Figure 2 of the confusion matrix, for the initial career prediction, model class BA and QA have higher FP count than the False negative (FN) count because when it considers each class individually with one vs all approach, both classes become minority class. On the other hand, class SE has a higher FN count than FP count. This is because when it considers each class individually in one vs all approach, class SE becomes the majority class. Also, this model has a higher TP count than both FP and FN counts. Therefore, it further proves that this model performs well. According to Figure 3, the initial career prediction model has an excellent AUC value performance for all classes since the AUC values are greater than 90%. Therefore, according to the ROC curve and AUC values, it can be concluded that the trained model is performing effectively. According to the Lime interpretation, as shown in Figure 4, features such as GPA, degree, and English subject display a positive correlation while ethnicity and hometown correlate negatively for the outcome of the initial career being a Software Engineer.

	precision	recall	f1-score	support
1	0.91	0.98	0.95	44
2	0.80	0.96	0.87	49
3	0.98	0.89	0.94	141
accuracy			0.92	234
macro avg	0.90	0.94	0.92	234
weighted avg	0.93	0.92	0.92	234

Fig. 1. Classification Report for initial career Prediction

TABLE III
ACCURACY PERCENTAGES FOR PREDICTING INITIAL CAREER AND SALARY

Classification Algorithm	Initial Career	Initial Salary
XGBoost	92.31	90.35
Random Forest	85.31	85.71
SVM	76.92	88.11
KNN	79.02	82.21
Decision Tree	81.81	80.41
Naive Bayes	72.72	75.52

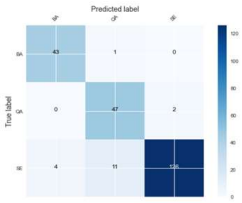

Fig. 2. Confusion Matrix for Initial Career Prediction

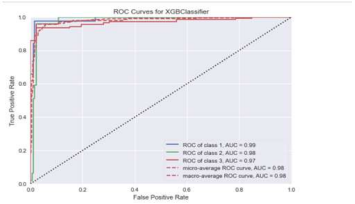

Fig. 3. ROC Curve for Initial Career Prediction

B. Result Analysis of Prediction of career and salary range progression and professional courses

To predict the career progression, salary range progression, and professional courses recommendation, preprocessed dataset was trained using four different multi-output classification algorithms as shown in Table II. After the comparison of the trained models, XGBoost Classifier resulted in higher testing accuracies for career progression prediction with a value of 86.45% and the prediction of the salary progression with a value of 88.76% respectively. Furthermore, the Random Forest classifier resulted with a highest testing accuracy of 93.52% for the prediction of professional courses, as shown in Table IV. Career progression prediction model contains multi-outputs such as ASE No of Years, SE No of Years, SSE No of Years, Tech Lead No of Years, and Architect No of Years. Therefore, only the prediction of ASE No of Years is considered here for the interpretation. Figure 5 represents the classification report for the model performance of prediction of ASE No of Years output. Based on the Figure, there are two classes (1 – 1 Year, 2 – 2 Years). As given in the support column there are more samples related to 1 Year than 2 Years. Therefore, the class which is related to

TABLE IV
ACCURACY PERCENTAGES FOR PREDICTING CAREER AND SALARY PROGRESSION AND PROFESSION COURSES

Classification Algorithm	Career Progression	Salary Progression	Professional Courses
XGBoost	86.45	88.76	91.47
Random Forest	80.83	82.65	93.52
SVC	80.87	81.95	92.50
KNN	74.40	78.18	87.50

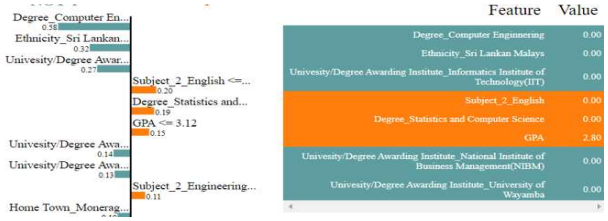

Fig. 4. Lime Explanation for initial career Prediction

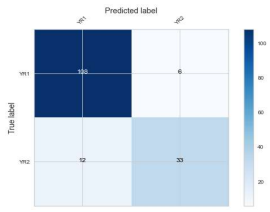

Fig. 6. Confusion Matrix for Prediction of ASE No of Years

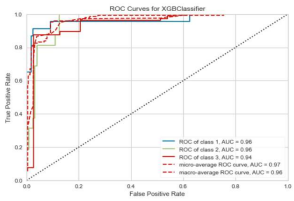

Fig. 7. ROC Curve for Prediction of SE No of Years

1 Year becomes majority class and 2 Years becomes minor class. As a result, precision should be more focused when evaluating the performance with regard to the 2 Years class while recall should be focused when evaluating the 1 Year class. However, both classes have F1-score over 0.95, and model accuracy is almost close to 0.90. Eventually, it can be concluded that the model predicting the ASE No of Years has a good performance.

Figure 6 represents the confusion matrix for the model performance of prediction of ASE No of Years output. Both classes of this model have much lower FP and FN value count when compared with TP count. Therefore, it can be concluded that the considered model has a good performance. Figure 7 demonstrates the ROC Curve received for the output of SE No of Years. This model has four classes namely, No of year 0 (0), No of year 1 (1), No of year 2 (2), and No of year 3 (3) since this is a multi-class classification. ROC curve generated for class 1 will be against class 2 and 3, as depicted in Figure. Similarly, ROC curves are generated against other classes. Therefore, it can be concluded that the model has a good performance since the curves are adjacent to the top-left corner. According to the Lime interpretation, features that comprised of positive and negative correlations were differentiated. As for the output of ASE number of years, the features such as Full Stack Development, Backend Development, Mobile Application Development, adapt to new technologies graduation year and GPA contributed positively while subject 1 agriculture and home town contributed negatively.

	precision	recall	f1-score	support
1	0.97	0.96	0.96	118
2	0.91	0.93	0.92	57
accuracy			0.95	175
macro avg	0.94	0.94	0.94	175
weighted avg	0.95	0.95	0.95	175

Fig. 5. Classification Report for Prediction of ASE No of Years

C. Result Analysis Predictive models for IT Employee Attrition, Retention Period and suggestion of remedial action

Similarly, to the above components, several classifiers were practiced in identifying the optimal models for the prediction of attrition, retention period, and remedial actions to retain

TABLE V
ACCURACY PERCENTAGES FOR PREDICTING EMPLOYEE ATTRITION AND RETENTION

Classification Algorithm	Employee Attrition	Employee Retention	Remedial action Prediction
XGBoost	86.02	90.44	93.85
Random Forest	89.70	89.68	92.14
SVM	73.52	75.47	87.14
KNN	82.35	81.42	87.14
Decision Tree	83.82	88.97	90.00
Naive Bayes	75.73	76.64	83.57

the IT employees for a longer duration. The preprocessed dataset was trained using six different classification algorithms as shown in Table II. Upon comparison of testing accuracy scores, the Random Forest classifier produced the highest accuracy of 89.70% for the prediction of employee attrition as displayed in Table V. Meanwhile XGBoost classifier provided the highest accuracies of 90.44% and 93.85% for prediction of the employee retention period and remedial action respectively. Figure 8 represents the classification report for the model performance of Employee Attrition Prediction Model. Based on Figure there are two classes (0 – No Attrition, 1 – Attrition). As given in the support column there are more samples related to No Attrition than Attrition. Therefore, class No Attrition become majority class. As a result, precision should be more focused when evaluating the performance with regard to Attrition while recall should be focused with regard to the No Attrition. However, both classes have F1-score over 0.90 and model accuracy over 89%. Eventually, it can be considered as a good performing model for predicting the employee attrition. Based on the confusion matrix as shown

	precision	recall	f1-score	support
0	0.87	0.99	0.93	75
1	0.98	0.82	0.89	61
accuracy			0.91	136
macro avg	0.93	0.90	0.91	136
weighted avg	0.92	0.91	0.91	136

Fig. 8. Classification Report for Prediction of Employee Attrition

in Figure 9, this model is considered as a better performing model since the count of TP is higher than both FP and FN count. As demonstrated in Figure 10, selected model for employee attrition prediction had a better performance since

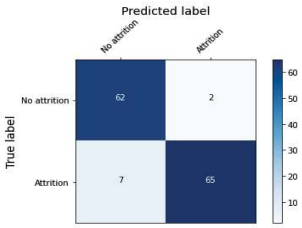

Fig. 9. Confusion Matrix for Employee Attrition Prediction

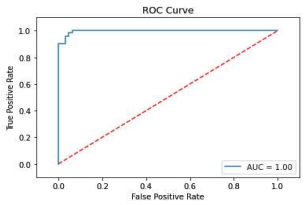

Fig. 10. ROC Curve for Employee Attrition Prediction

the curves were adjacent to the top-left corner. According to LIME interpretation of the IT employee attrition prediction model, features such as work environment satisfaction, number of companies worked and distance between home and company contributed positively while job satisfaction rate affected negatively for the prediction.

Conclusion and Future Works

The present study on career mentoring systems was conducted using emerging technologies and machine learning approach to help undergraduates by predicting their initial career based on their skills. The implemented system facilitates rapid career advancement of IT employees, and enables employers to manage the retention of IT employees via implementing the suggested remedial actions. Additionally, it could assist IT employees in the field of Software Engineering for enhancement of professional skills, which, in turn, will expedite their career promotions. To achieve these objectives, data was collected from existing IT employees, and models were trained using machine learning classification algorithms. This study proves that XGBoost algorithm is more suitable for initial career prediction of 92.31%, salary range of 90.35%, career progression of 86.45%, salary range progression of 88.76%, retention of 90.44%, and for suggesting remedies 93.85%. Moreover, the prediction of professional skills recommendation achieved via Random Forest, resulted in an accuracy rate of 93.52% while employee attrition prediction was 89.70%.

Based on the result analysis, it was concluded that there is a high impact from features such as GPA and degree on the prediction of initial career, while ability in full stack, backend, mobile application development, knowledge in programming languages, adapting to new technologies and GPA have a positive influence on the prediction of career progression. Moreover, for the prediction of employee attrition, features such as work environment satisfaction, the number of companies worked in, and the distance between home and workplace showed positive contribution.

In future studies, it is expected to collect more data from several IT organizations with the intention of obtaining higher accuracies. Furthermore, the scope of this study is planned to be expanded by providing career guidance to other Engineering Fields.

References

[1] Sri Lanka University Statistics 2019, Ugc.ac.lk, 2021. [Online]. Available: https://www.ugc.ac.lk/index.php

[2] Information and Communication Technology Agency — ICTA, - ICTA, 2021. [Online].Available:https://www.icta.lk/news/sri-lanka-aiming-200000-ict-workforce-by-2022/.

[3] K. Mlitz, "Global IT industry growth rate by segment 2018-2023," Statista, 27-Jan-2021. [Online]. Available: https://www.statista.com/statistics/967095/worldwide-it-industry-growth-rate-forecast-segment/.

[4] Alex, "Ten typical jobs graduates can do in IT," TARGETjobs, 27-Aug-2020. [Online]. Available: https://targetjobs.co.uk/career-sectors/it-and-technology/286189-ten-typical-jobs-graduates-can-do-in-it.

[5] Information Technology Jobs Descriptions," Mallory, 17-Jul-2021. [Online]. Available: https://mallory.com.au/information-technology-jobs-descriptions/.

[6] M. Y. Arafath, M. Saifuzzaman, S. Ahmed, and S. A. Hossain, "Predicting career using data mining," 2018 Int. Conf. Comput. Power Commun. Technol. GUCON 2018, pp. 889–894, 2019, doi: 10.1109/GUCON.2018.8674995.

[7] K. Sripath Roy, K. Roopkanth, V. Uday Teja, V. Bhavana, and J. Priyanka, "Student career prediction using advanced machine learning techniques," Int. J. Eng. Technol., vol. 7, no. 2, pp. 26–29, 2018, doi: 10.14419/ijet.v7i2.20.11738.

[8] A. Daharwal et al., "Career Guidance System using Machine Learning For Engineering Students (CS / IT)," no. June, pp. 3417–3420, 2020.

[9] Y. Liu, L. Zhang, L. Nie, Y. Yan, and D. S. Rosenblum, "Fortune teller: Predicting your career path," 30th AAAI Conf. Artif. Intell. AAAI 2016, pp. 201–207, 2016.

[10] H. Li, Y. Ge, H. Zhu, H. Xiong, and H. Zhao, "Prospecting the career development of talents: A survival analysis perspective," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. Part F129685, pp. 917–925, 2017, doi: 10.1145/3097983.3098107

[11] Q. A. Al-Radaideh and E. Al Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance," Int. J. Adv. Comput. Sci. Appl., vol. 3, no. 2, pp. 144–151, 2012.

[12] A. Qutub, A. Al-Mehmadi, M. Al-Hssan, R. Aljohani and H. S. Alghamdi, "Prediction of Employee Attrition Using Machine Learning," International Journal of Machine Learning and Computing, vol. 11, no. 2, p. 5, 2021.

[13] R. Punnoose and P. Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms," (IJARAI) International Journal of Advanced Research in Artificial Intelligence, vol. 5, no. 9, pp. 22-26, 2016.

[14] R Shiva Shankar, J Rajanikanth, V V Sivaramaraju and K VSSR Murthy, "PREDICTION OF EMPLOYEE ATTRITION USING DATAMINING," in IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), 2018

[15] Y. Pan, X. Peng, T. Hu, and J. Luo, "Understanding what affects career progression using linkedin and twitter data," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018-Janua, pp. 2047–2055, 2017, doi: 10.1109/BigData.2017.8258151.

[16] D. Haritha, "Smart Career Guidance and Recommendation System," vol. 7, no. 3, pp. 633–638, 2019.