TAYLOR'S UNIVERSITY
Wisdom · Integrity · Excellence

Data Science Principle

ITS65704

GROUP ASSIGNMENT SEPTEMBER 2025

[GROUP 25]

Kryštof Vávra 0383088

Rafia Rabbani Easha 0366837

lin zishu 0370947

Liu Yifan 0374114

1.0 Project Background and Project Goal:

Employee attrition in which employees may move out of an organization involuntarily or voluntarily is one of the issues that arise in the process of managing talent in human resources in the modern settings. A productive working environment can be disrupted due to a high turnover; hiring and training will be disrupted and costly; morale demoralization of the staff members. The trend has now changed and companies are resorting to data-driven solutions to know more about the pattern of the attrition and hence be in a position to retain the best talent and form healthier workplaces.

The present project is concerned with the analysis of the employee turnover on HR actual data. It consists of 1470 rows and 35 columns such as age, gender, department, job role, job satisfaction, monthly income, years at the company and others. These features open a lot of possibility of conducting exploratory data analysis (EDA) and predictive modelling. All the analyses and modeling have been done in Python on Google Colab Environment that ensures an open and smooth process.

The aim of the given project comprises the identification of trends and characteristics associated with employee turnover using data science on the provided Human Resource data. Using critical EDA and statistical analysis we will seek to go deep in an attempt of coming up with critical variables that will affect the determination of whether an employee will leave a company or remain in it. This would subsequently be followed by development of predictive machine learning models that will be used to study the likelihood of whether, based on the use of the attributes of the workers, the occurrence of the situation of attrition would be present.

Among the deliverables our project results will comprise are implementable recommendations and a working model that can be implemented by the HR departments in decision making and employee retention. Reproducibility and transparency were achieved within Google Colab back-to-back cleaning, visualization, training, and evaluation of the data.

2.Dataset Description

This dataset used IBM HR Analytics Employee Attrition & Performance dataset. It is included 1470 employees and 35 columns. These data combined demographic, job-related, performance-related attributes. This dataset is for employee attrition to analyze patterns and build predictive models

Here’s are some key features:

Attrition(Target): Indicates whether the employee left the company or not

Age: employee age

MonthlyIncome: monthly income in USD

JobRole: employee role

OverTime: Whether the employee works overtime

JobSatisfaction: Satisfaction rating from 1 (Low) to 4 (Very High)

YearsAtCompany: How many years the employee worked in the company

WorkLifeBalance: Work-life balance rating

Gender, MaritalStatus, and BusinessTravel: Categorical personal details

3. Summary Statistics and Exploratory Data Analysis (EDA)

First the review of the structure and distributions in the dataset was done. There are 1470 employee records and 35 features, both numerical and categorical. Around 16% of the employees had left the company, meaning that the target variable (Attrition) is imbalanced.

The univariate analysis showed that most employees are between 30 and 40 years old, have mid-range monthly income, and are fewer than 5 years at the company. Boxplots revealed that the ones who left tend to be younger, less experienced and with lower incomes.

The bivariate analysis showed that those who worked overtime, frequently travelled for business and were single tended to have a higher leaving rate. Job positions affected by the attrition the most are Laboratory Technician and Sales Executive.

The correlation heatmap showed moderate negative correlation of our target variable (Attrition) with features like MonthlyIncome, Age, JobLevel, and YearsAtCompany. Expected relationships, such as JobLevel and MonthlyIncome were also shown. A pair plot showed that employees who left tend to have lower income and spend shorter time in the company.

4. Data Preprocessing and Issues

First the columns with no predictive value were removed from the dataset. These were: EmployeeCount, StandardHours, Over18, and EmployeeNumber, which were either constant or identifiers. The target variable Attrition was converted from Yes/No to binary (1 = left, 0 = stayed). Binary variables like OverTime and Gender were also encoded as 1/0. Then the categorical variables with multiple values (e.g., JobRole, BusinessTravel, Department), were split using one-hot encoding with drop_first=True to avoid multicollinearity. All numerical columns were kept as-is.

No missing values were found in the dataset. Scaling was applied using StandardScaler, because our chosen model (Logistic Regression) is sensitive to feature scale. The dataset was then split into training and testing sets using an 80/20 ratio with stratified sampling to preserve class balance.

5. Model selection and interpretation

In this project, we selected three different models to approach the employee attrition prediction task: Logistic Regression, Random Forest, and XGBoost.

We began with Logistic Regression as a baseline model, largely due to what appears to be its simplicity and transparency. It tends to reveal how individual features may influence the likelihood of attrition. Given that the dataset is imbalanced, we applied the class_weight='balanced' parameter to avoid overemphasizing the majority class.

We then used Random Forest, which seems to offer better flexibility in capturing nonlinear patterns. It handles both numerical and categorical data without requiring scaling, and because it aggregates multiple decision trees, it appears to reduce overfitting. We again applied class weighting here to deal with the imbalance in the target variable.

Finally, we incorporated XGBoost as a more advanced method. It builds trees sequentially to correct earlier errors, and includes the scale_pos_weight parameter, which seems particularly useful in cases where one class is significantly underrepresented—as is true in our dataset. Though more complex, it tends to perform well with structured data and often produces strong predictive results when tuned properly.

Using all three models allowed us to compare different approaches—from simple and interpretable to more complex and performance-driven. This helped us explore which features seemed most influential and how well each method handled the imbalance. Rather than focusing solely on accuracy, we emphasized metrics like precision, recall, and F1-score, which provide a clearer view of real performance in this context.

6. Model Validation

In order to validate how well our three classification models—Logistic Regression, Random Forest, and XGBoost their performance. To know whether the employees leave or not we use these models to predict whether employees leave or stay, and help us understand each model's accuracy

Logistic Regression:

This model gave balanced results. The ROC-AUC score is high which means that it works really well to tell the difference between the two groups overall, and also contributes a lot in finding who employees who left and stayed

Random Forest:

this model overall has good accuracy, but it also missed a lot of employees who left. However the ROC-AUC score is about 0.78 which means okay to predict whether it's stay or leave

XGBoost

This model does a better job in finding the employees who left compared to Random Forest. It also shows more true positives in the confusion matrix. The overall ROC-AUC score was lower, around 0.76 which means that this model is not stronger than other two models

Compared to these three models, Logistic Regression got the highest score and predicts employee stay or left well. Random Forest was close, but missed some people who left. XGBoost was better than Random Forest in finding those who left. Overall Logistic Regression is the best choice for predicting whether employees stay or leave

7.0 Conclusion and business recommendations:

The analysis of HR employee attrition dataset has provided certain valuable information related to the reasons of employee turnover. Using the Exploratory Data Analysis (EDA) we succeeded in establishing that there is high correlation between attrition and other variables that included job satisfaction, overtime, monthly income, years at the firm and work life balance. The likely chance of an individual leaving the organization was high where workers had to do overtime, where the workers had poor job satisfaction or where they had served less time in the organization.

We were able to successfully predict attrition based on the machine learning models with acceptable accuracy that has proved the fact that data science can be applied successfully as a useful tool in human resource decision-making. The given project demonstrates the success of data-based approach that is able to uncover the hidden pattern and techniques to retain the higher rate of employees. The awareness of the main factors that lead to attrition will enable companies

to take actions to prevent attrition by taking prior activities of improving employee engagement in an organization, reducing the turnover level, and ultimately, cutting costs.

Based on the results of the present project, the following recommendations are given concerning the HR and the organizational leadership:

1: Measure, and enhance Job Welling-Robustly gauge, and raise satisfaction of employees at regular intervals through surveys, or feedforward setups. The characteristic which the ex-employees of the organization possessed- was low satisfaction. HR should identify the dissatisfaction sources and implement the suitable changes to attract them.

2: Overtime Workload- Overtime was also linked to attrition to a significant extent. You can think of clear overtime policies and good working life. Promote the well-being culture and suppress burnout.

3: New Employee Retention Programs- All the employees who have not worked at the company too long would tend to quit. Add mentorship and onboarding care to make new company members feel more welcomed and integrate themselves in the company better.

4: Review Compensation structures- It was noted that the levels of attrition were higher among the low-income staff. Ensure there is equity and competitiveness of low-turnover jobs in terms of checking salaries.

5: Individual Career Development Plans- Offer career growth, training and recognition to encourage terminal growth. Once the employees have a view that they can have a career in the company they will hardly leave the company.

6: Use Predictive Models in HR Decisions- Utilise the trained machine learning model on the HR analytics systems of the employees to score the employees based on high risk of attrition. This helps to intervene and lend a helping hand before resigning.

8.0 Appendix:

Colab file and dataset file: