Stay or Stray: Machine Learning Decodes the Attrition Mystery

Group 10:

Kenny Lim Yong Xiang (A0262753B) | Park Jung Hyun (A0262660H) | Shaun Heng Wei Hao (A0201678A)

Soon Qing Rong (A0262852B) | Yep Yan Ling (A0205736E)

Github Link: https://github.com/takoyakee/BT5153_Group10 | **Demo Link:** https://attrition- analytics.herokuapp.com/

1. Introduction

1.1. Background

Employee attrition, regardless of organisations and industries, can negatively impact the productivity and morale of remaining employees, which can directly lead to an increase in costs. It is especially critical to the management of a business when the attrition rate is significantly high. Predicting attrition risk and addressing the issue timely are therefore crucial for management to maintain a stable workforce and minimise employee turnover. Companies can develop effective retention strategies and implement appropriate policies to foster a more conducive work environment for employees by identifying employees at risk of attrition and investigating the cause of their dissatisfaction.

1.2. Business Value

The business values derived from our study can be harnessed by organisations in different ways. First, companies can integrate the developed machine learning model into the organisation's HR analytics system to continuously monitor employee attrition risks. The insights provided by the model and the feature importance analysis can then be deployed to develop targeted retention strategies, such as flexible work arrangements and the provision of additional support for specific concerns. Finally, companies can tailor the cost-benefit matrix to their specific priorities and cost structures and obtain optimised prediction strategies.

The aim of this study is to empower organisations with insightful data to enhance employee loyalty and satisfaction, ultimately leading to cost reductions with minimised turnover. By identifying and addressing the key factors contributing to attrition, companies can not only reduce expenses related to recruitment and onboarding but also foster a more engaged and content workforce, thereby benefiting both the organisation and its employees.

1.3. Objective

Our project aims to create a robust machine learning solution that helps organisations tackle employee attrition efficiently by identifying attrition risks and providing interpretable and actionable insights for management. This will be achieved by pre-processing the raw data, exploring various classification models with the expected value framework, analysing models with interpretability methods, and optimising decision thresholds according to business needs.

2. Data

The dataset[1] consists of 1470 rows (employees) and 35 columns (including the target variable - Attrition). It contains various aspects of an employee's work experience such as demographics, work environment, job and career progression, compensation and benefits and finally work-life balance and time management. The features are shown in Appendix - Table A1.

2.1. EDA and Feature Engineering

The aim of exploratory data analysis (EDA) is to gain a better understanding of the observed data to identify traits of employees that left the company. Most importantly, the exploration should inform the selection of appropriate features, preprocessing techniques and machine learning algorithms

[1]IBM HR Analytics Employee Attrition & Performance Dataset: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

to predict employees with an inclination to leave.

The features were systematically analysed to identify patterns in the following sequence; univariate analysis e.g. distribution of features etc, bivariate analysis which reveals the distribution of both the positive and negative class and multivariate analysis that visualises relationships between features, separated by the target variable. Finally, we check for correlations amongst the features to determine features to include in our model.

The target variable (Attrition) has a positive class of approximately 16% (237 out of 1470), making the dataset imbalance. This will be addressed in the data preprocessing section using Synthetic Minority Oversampling TEchnique (SMOTE).

Upon quick inspection of the remaining features, several columns such as *EmployeeNumber*, *Over18*, *EmployeeCount* and *StandardHours* either contain one value for all rows or contain a unique value for each row. These columns were dropped as they are not likely to provide information to the model and instead negatively impact model performance.

We hypothesise that salary is a strong factor for employee attrition amongst other factors that could reduce the true pay i.e. taking into account overtime and the range of peers in similar job levels and roles.

The key findings of EDA are presented below: (a) younger people/singles in the early parts of their career have higher attrition rates, (b) overtime seems to have no bearing on attrition rates and (c) lower income employees have a high probability of leaving but generally changes as job level crosses a threshold e.g. managers & directors.

For (a), the distribution of various demographic features show that the younger generation who have less working experience and are single are more likely to leave the company. Younger employees with no dependents are likely to be more ambitious (sensitive to price/salary changes) and bigger risk takers, having less commitments. To better capture the demographics, we binned[2] the *Age* Feature by generations and created a *AgeMarital* feature that combines the age category and the marital status of an employee.

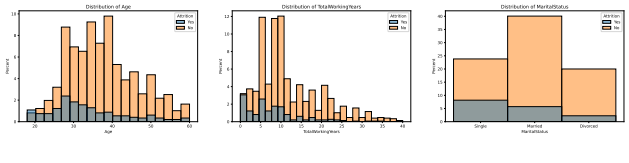

Figure 1. Charts for Demographics

For (b), we investigate the work-life balance and time management component and its effects on attrition rates. The histogram on the left of Figure 2 showing the distribution of workers who work overtime, is inconclusive as the percentage of workers ($\sim$10%) who left the company is the same for both workers who do overtime and workers who do not do overtime. This is counterintuitive as overtime diminishes an employee's real pay as overtime increases. Based on this information alone, we cannot conclude that overtime is important in deciding if a worker leaves the company. However, it suggests that there may be other factors such as Monthly Income that are more important than overtime taken out of context.

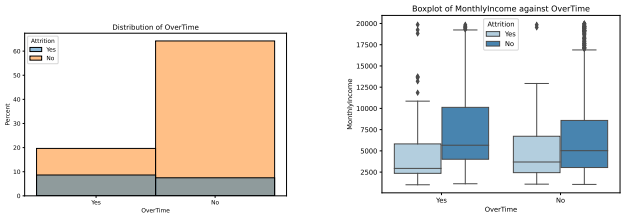

Figure 2. Charts for *OverTime*

To dig deeper, the boxplot of *MonthlyIncome* against *OverTime* (Right Chart in Figure 2) shows the effects of overtime changes. We see that attrition is constant regardless of overtime and consistently are being lower paid than non-attrited employees. However, following our business intuition, we leave the feature in for the machine to pick up more subtle patterns that we may have missed.

Finally for (c), we observed that Attrition seems to decrease as *JobLevel* increases (Left Chart in Figure 3). This could be due to a smaller sample at the top of the corporate ladder and that management generally takes it on the chin to overtime. To get a clearer picture, we see in the right chart in Figure 3 that the salary of employees regardless of attrition tend to be more comparable (except at level 4). This gave us an intuition that *JobLevel* may have influence on the predictions of the model. To accurately model the influence of "true pay" on attrition rates, we engineered two other features that pegs *MonthlyIncome* to (i) *JobLevel* and (ii) *JobRoles*.

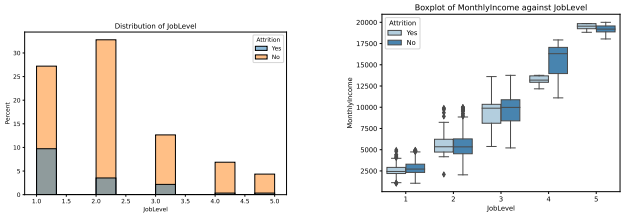

Figure 3. Charts for *Attrition* over *JobLevel*

Lastly, correlation amongst the features was calculated to identify features that are highly correlated and could cause

[2]Millenials ($<$ 37), Gen X ($<$ 54), Boomers ($<$ 73), Silent ($>=$ 73)

multicollinearity issues. Figure 4 shows a heatmap of the correlated features. Most features are weakly either positively correlated or negatively correlated. The stronger correlated features are shown in lighter colours with the one that are closest to white hot being most positively correlated.

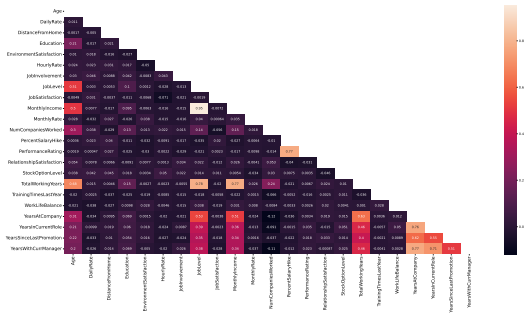

Figure 4. Correlation Heatmap for all Numerical Features

Based on the correlation heatmap, several features will be dropped. In particular, *YearsInCurrentRole* and *YearsWithCurrManager* are highly correlated to *YearsInCompany* and provide minimal information about a person's career. Instead, a new column *TenureRatio* will be created from dividing *YearsAtCompany* by *TotalWorkingYears*.

Other features are also created to capture job history, such as *FirstJob* (where *NumCompaniesWorked* == 0) and Job Hopping Tendency (calculated as *TotalWorkingYears*/*NumCompaniesWorked*).

Within the Compensation and Benefits category of features, we have *'DailyRate'*, *'MonthlyRate'*, *'HourlyRate'* which seem related to income but do not provide significant information upon closer inspection. We found no correlations to *MonthlyIncome* and decided to drop them for the normalised income feature which takes the ratio of *MonthlyIncome* to the median salary of the job role/level mentioned above.

There is also a group of features that are naturally correlated such as *Age* and *JobLevel*, *PerformanceRating* and *PercentSalaryHike* etc. For this group of features that make sense, we will leave them as is.

2.2. Summary of final features

In summary, 9 features from the original dataset were dropped[3] and 8 new features added[4]. The new dataset with 32 feature variables and 1 target variable is shown in Table 1:

[3] *'Over18', 'EmployeeNumber', 'StandardHours', 'EmployeeCount', 'DailyRate', 'MonthlyRate', 'HourlyRate', 'YearsInCurrentRole', 'YearsWithCurrManager'*

[4] *'AgeMarital', 'FirstJob', 'JobSatis_WLB', 'JobHopIndex', 'TenureRatio', 'MonthlyIncomeInJobLevel', 'MonthlyIncomeInJobRole', 'Age'*

Table 1. Features of Dataset

CATEGORY	FEATURES
DEMOGRAPHICS	AGE, GENDER, MARITAL STATUS, EDUCATION, EDUCATIONFIELD, AGEMARITAL
WORK ENVIRONMENT	BUSINESSTRAVEL, DEPARTMENT, DISTANCEFROMHOME, ENVIRONMENTSATISFACTION, JOBINVOLVEMENT, JOBROLE, JOBSATISFACTION
JOB AND CAREER PROGRESSION	JOBLEVEL, NUMCOMPANIESWORKED, TOTALWORKINGYEARS, YEARSATCOMPANY, YEARSSINCELASTPROMOTION, JOBHOPPINGTENDENCY, FIRSTJOB, TENURERATIO
COMPENSATION AND BENEFITS	MONTHLYINCOME, PERCENTAGESALARYHIKE, PERFORMANCERATING, STOCKOPTIONLEVEL, MONTHLYINCOMEIINJOBLEVEL, MONTHLYINCOMEINJOBROLE
WORK-LIFE BALANCE	WORKLIFEBALANCE, OVERTIME, TRAININGTIMESLASTYEAR, JOBSATIS_WLB, RELATIONSHIPSATISFACTION

2.3. Data Preprocessing

After finalising the set of features, data preprocessing will be done to restructure the dataset such that it is ready for model training.

Firstly, the categorical variables (e.g. *Department*, *EducationField*) will be transformed using one hot encoding, ordinal variables (e.g. *BusinessTravel*) will be transformed to preserve the relative magnitude of each category while numerical variables will be standardised by its mean and variance. This gave a total of 61 variables after this stage of preprocessing.

Secondly, the dataset will be split into 80% training set and 20% test set, while using stratified sampling.

Lastly, SMOTE oversampling is used to overcome the class imbalance (Attrition has a positive class of 16%), by increasing the number of training samples with positive classes. This is to ensure that the resultant model is not biased against positive classes which are smaller in size.

3. Methodology

Models will be trained on the pre-processed data to generate valuable insights for businesses. In this section, we describe the chosen models, the process for model selection, and the expected value framework, which will be used to optimise for specific business requirements.

3.1. Models and Hyperparameter Tuning

In this binary classification problem, several machine learning models have been utilised, such as Logistic Regression (LR), Support Vector Classifier (SVC), K Nearest Neighbours (KNN), Decision Tree (DT), Gradient Boosting classifier (GB) and Random Forest classifier (RF). The models were selected as they can capture complex relationships, improve accuracy through ensemble techniques and provide interpretability. The dataset was divided into 80% training set and 20% test set and 5-fold cross-validation was performed for hyperparameter tuning. Table 2 below summarises the range of parameters used.

Table 2. Range of Hyperparameters Tuned

ML MODEL	HYPERPARAMETER TUNED	RANGE
KNN	N_NEIGHBORS	1-5
LR	PENALTY	NONE, L2
	C	0.1-2
SVM	KERNEL	LINEAR, POLY RBF,
	C	1-2
DT	MAX_DEPTH	10-15
	MIN_SAMPLES_SPLIT	2-5
	MAX_LEAF_NODES	2-5
GB	N_ESTIMATORS	50-300
	LEARNING_RATE	0.1-1
RF	MAX_DEPTH	10-15
	MIN_SAMPLES_SPLIT	2-5

A custom scoring method extending the expected value framework will be used to benchmark models. The custom score constitutes the sum product of the normalised confusion matrix and the cost-benefit matrix (from expected value framework) which will be discussed in Section 3.3. For each type of ML model, the set of hyperparameters corresponding to the best expected value is chosen. A grid search algorithm is used to exhaustively find these optimal sets. After which, each model is fitted on the train set with the best hyperparameters.

3.2. Model Selection Process

The model selection process consists of four stages: 1. Model Training and Tuning (Section 3.1), 2. Model Evaluation, 3. Feature Selection, and 4. Final Training. First, the model is trained and tuned. Then, the evaluation of the model's performance will be on its expected value on the test set and behaviours in the various performance evaluation plots (e.g., ROC-plot, gain plot, expected value plot, and precision-recall plot). The model with the best overall performance will be chosen. Next, Recursive Feature Elimination (RFE) with 5-fold cross-validation is employed to identify the most critical features. Finally, the selected base model is retrained using the optimal hyperparameters and feature set.

3.3. Business value

As previously mentioned in Section 3.1, the expected value framework is employed to maximise business value. This is done by weighing the relative cost of incorrect predictions for businesses via a cost-benefit matrix, which in turns guides the models to intrinsically optimise for the desired business outcomes. Each element in the cost-benefit matrix corresponds to the cost/benefit for the respective entry in the confusion matrix. The derivation for the cost-benefit is explained below where we also made the following assumptions:

1. The dataset is without treatment

2. If employee is predicted to leave, the company will provide treatment and if employee is predicted to stay, the company will not provide treatment

3. If treatment is provided, there is a 50% chance the employee will stay, which varies according to company's treatment effectiveness

4. Treatment cost is 3 months worth of salary [1]

5. Replacement cost is 12 months worth of salary [2]

6. Company must find a replacement if an employee goes

7. Salary is constant

The company has two decisions to make, which is treatment (e.g., monetary incentives), or no treatment, based on the predictions. The value derived from each decision varies according to the actual employee behaviour. Combining the assumptions and the 4 different situations, we define our cost-benefit matrix as seen in Figure 5.

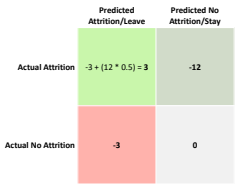

Figure 5. Proposed Cost-Benefit Matrix

One observation is the priority in minimising False Negatives (predicting employees to stay when they leave) due to the hefty cost of training a new employee as compared to retaining the employee. This preference for employee retention is corroborated by many companies such as AT&T (American Telephone & Telegraph) [3] and Costco [4], which justifies our cost-benefit matrix. With this cost-benefit matrix, models will learn to optimise business values effectively.

4. Predicting employee attrition

In this section, we present the final model selection as discussed earlier for the task of attrition prediction. We also highlight key predictors of attrition through model interpretability for intervention and suggest attrition prediction strategies for businesses.

4.1. Model Selection and Performance

Following the steps outlined in Section 3.2, results for each stage will be presented. For Model Training and Tuning, performance metrics like expected value (EV), precision (p), recall (r), f1 and Area Under the ROC Curve (AUROC) using models with tuned hyperparameters were obtained.

Table 3. Best Fitted Models

ML Model	Hyperparameter Tuned	Tuned Values
KNN	n_neighbors	1
LR	penalty	None
	C	1.44
SVM	kernel	Linear
	C	1.94
DT	max_depth	10
	min_samples_split	2
	max_leaf_nodes	2
GB	n_estimators	250
	learning_rate	0.8
RF	max_depth	14
	min_samples_split	4

Table 4. Results of best fitted models

ML Model	EV	P	R	F1	AUROC
KNN	-1.33	0.22	0.40	0.28	0.63
LR	-0.99	0.60	0.45	0.51	0.70
SVM	-1.16	0.53	0.38	0.44	0.71
DT	-1.51	0.22	0.60	0.32	0.60
GB	-0.96	0.52	0.49	0.51	0.73
RF	-0.99	0.39	0.60	0.47	0.72

Next, various performance evaluation curves were plotted in Figure 6 for a multi-dimensional perspective on the models' performance. DT and KNN have the worst performance for expected value and in the various plots. RF was chosen for having the second highest expected value and for acceptable trade-offs in precision and recall at the optimal threshold. Although the EV framework captured the trade-off in precision and recall, their values are still considered for a well-rounded performance. The imbalanced dataset results in sub-optimal f1 values, but a recall of 0.6 suggests that 60% of people who will leave are predicted, which is a steep improvement from random predictions that will only identify the attrition proportion of 16%.

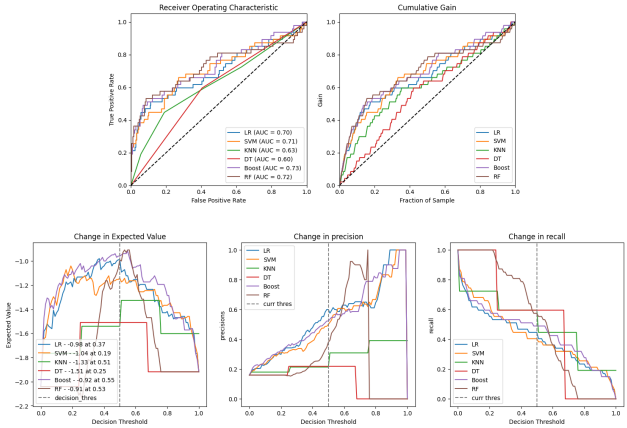

Figure 6. Performance evaluation curves

Next, 20 of 61 features were selected using RFE with cross validation. Figure 7 shows the variation in expected value and the optimal number of features were selected by balancing trade off in performance and interpretability. Table 5 summarised the dropped features according to their categories identified during EDA.

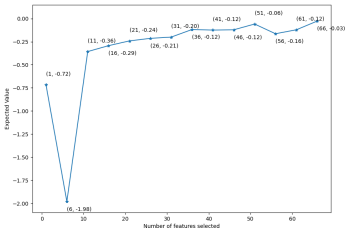

Figure 7. Change in Expected value based on number of features

Table 5. Dropped features

CATEGORY	DROPPED FEATURES
DEMOGRAPHICS	GENDER, EDUCATION, EDUCATIONFIELD, AGEBOOMERS
WORK ENVIRONMENT	BUSINESSTRAVEL, DEPARTMENT, JOBINVOLVEMENT, JOBROLE, JOBSATISFACTION
JOB AND CAREER PROGRESSION	NUMCOMPANIESWORKED, YEARSSINCELASTPROMOTION, TENURERATIO
COMPENSATION AND BENEFITS	PERCENTAGESALARYHIKE, PERFORMANCERATING, MONTHLYINCOMEIINJOBLEVEL, MONTHLYINCOMEINJOBROLE
WORK-LIFE BALANCE	WORKLIFEBALANCE, TRAININGTIMESLASTYEAR

Lastly, the model is retrained on the optimal hyperparameters and feature set. The final threshold is selected at 0.56 for optimal expected value of -0.87 as seen in Figure 8. Feature selection and threshold optimising has improved the expected value from -0.99 to -0.87. Threshold refers to the threshold for attrition prediction (i.e., for threshold of 0.56, an employee is predicted to leave if their attrition risk is 56% or higher). Further discussions of thresholds when business' values change will be presented in Section 4.5

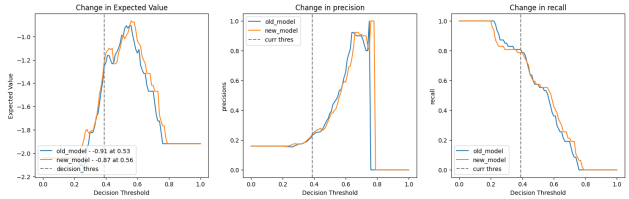

Figure 8. Change in metrics using different decision thresholds

4.2. Model Interpretability

In this project, we employed both global interpretability (feature importance) and local interpretability (local interpretable model-agnostic explanations (LIME)) to enhance the model's transparency and support data-driven decision-making for stakeholders. Feature importance provides insights into key factors driving attrition (or retention) of the company, which can guide general intervention strategies. Meanwhile, LIME offers insights into the unique factors influencing individual employees, enabling the creation of tailored interventions for maximum effectiveness.

4.3. Feature importance

Impurity based and permutation feature importance methods as complementary approaches. The result is as shown in Figure 9.

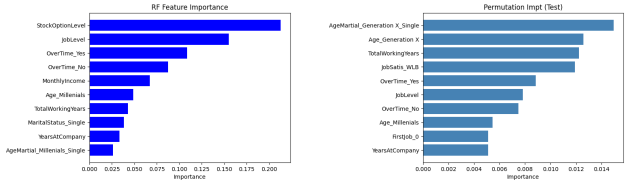

Figure 9. Feature and Permutation Importance

Few features (e.g., *OverTime, YearsAtCompany, TotalWorkingYears, JobLevel*) were important in both test and train sets, bolstering the models' credibility. Demographic features allow the company to identify the target group, which may be Millennials or Generation X singles, or individuals in the HR department. Compensation or work intensity-related features are actionable insights that highlight potential areas for improvement to reduce attrition risk. For example, offering a pay raise or decreasing overtime would lead to lower predicted attrition risk and represent feasible solutions for implementations.

4.4. LIME

LIME analyses individual employee records and returns key factors driving their attrition decisions. The results could be processed and be deployed in batch or real-time (depending on data update frequency and compatibility with systems) for HR management. A sample web app has been deployed at https://attrition-analytics.herokuapp.com/ to demonstrate its usefulness in formulating targeted strategies. Screenshots of dashboard and analytics for an employees are displayed in Figure 10.

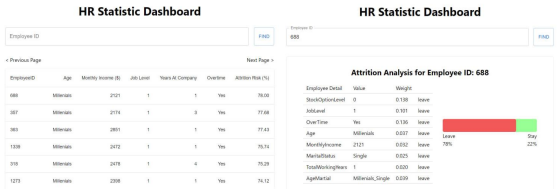

Figure 10. Screenshots of dashboard and analytics for 1 employee

4.5. Business Use Case

The final model incorporates the costs and benefits of businesses and provides a decision threshold to optimise expected value. The success of treatment was defined as 50%. However, given the dynamic nature of business, this confidence level may vary across organisations and time periods. The model is robust and can adjust its thresholds when there is a change to cost/benefits, which translates into taking on a more or less conservative approach based on the confidence in retention success. By capturing patterns in employee attributes, the model has demonstrated superior performance compared to naive strategies, such as random target, no target and target all. The result of each strategy is illustrated in Table 6, Table 7 and Table 8.

Table 6. Results (Success of treatment (%): **50**, Benefit of TP: **3**)

RANDOM TARGET	NO TARGET	TARGET ALL	RF MODEL	
EV	EV	EV	EV	THREHOLD
-2.077	-2.307	-1.847	**-0.87**	0.56

Table 7. Results (Success of treatment (%): **80**, Benefit of TP: **6.6**)

RANDOM TARGET	NO TARGET	TARGET ALL	RF MODEL	
EV	EV	EV	EV	THREHOLD
-1.731	-2.307	-1.155	**-0.59**	0.53

Table 8. Results (Success of treatment (%): **20**, Benefit of TP: **-0.6**)

RANDOM TARGET	NO TARGET	TARGET ALL	RF MODEL	
EV	EV	EV	EV	THREHOLD
-2.423	-2.307	-2.539	**-1.20**	0.56

When the confidence in success of treatment is high, the optimal threshold decreases slightly. The company is less conservative in predicting attrition and would likely target more people with the aim of increasing retention and capturing cost-savings benefits.

5. Conclusion

Our project has demonstrated the potential of machine learning algorithms in predicting employee attrition according to business needs, with Random Forest emerging as the chosen model. Feature elimination and threshold optimisation have led to improvements in the model's expected value. The incorporation of global and local feature importance for model interpretability allows stakeholders to gain different levels of actionable insights regarding employee retention strategies and policy adjustment. A sample dashboard effectively demonstrates this practical application.

5.1. Limitations

However,

1. The dataset used in the project may not be representative of all industries and organisations, which can potentially limit the generalizability of the model.

2. The values assigned in the cost-benefit matrix are based on assumptions, and therefore, they may not hold true for every company. Different companies may have to adjust the matrix according to their unique cost structures and employee valuation frameworks. In particular, it is important to acknowledge that individual employees possess distinct value within an organisation in real life; however, the cost-benefit matrix employed in this study operates under the assumption of uniform valuation for all employees.

3. Our analysis reflects a snapshot of employee data, which may not capture the dynamic nature of employee satisfaction and attrition risks over time.

5.2. Future Works

Future research directions are contingent upon the accessibility of different data types, including temporal and more granular information. The current dataset lacks a temporal dimension. As employee attrition is influenced by financial cycles, predictive models would improve with real-time or seasonally adjusted data. Furthermore, an investigation of attrition patterns across different departments would enable department-level attrition predictions. Department-specific information will enable large organisations to achieve more precise and relevant model outcomes appropriate for their departmental structures.

References

[1] "What is a retention bonus? (how it works plus benefits)." https://sg.indeed.com/career-advice/pay-salary/retention-bonus. Accessed: 2023-04-21.

[2] T. Marsden, "What is the true cost of attrition?," *Strategic HR Review*, vol. 15, no. 4, pp. 189–190, 2016.

[3] "At&t's $1 billion gambit: Retraining nearly half its workforce for jobs of the future." https://www.cnbc.com/2018/03/13/atts-1-billion-gambit-retraining-nearly-half-its-workforce.html. Accessed: 2023-04-21.

[4] "Costco is raising its minimum wage. here's why it matters to investors.." https://www.barrons.com/articles/costco-is-raising-its-minimum-wage-heres-why-it-matters-to-investors-51614286634. Accessed: 2023-04-21.

Appendix

Table A1. Features of Dataset (Original)

CATEGORY	FEATURES
DEMOGRAPHICS	AGE, GENDER MARITALSTATUS, LEVELOFEDUCATION, EDUCATIONFIELD
WORK ENVIRONMENT	FREQUENCYOFBUSINESSTRAVEL, DEPARTMENT, DISTANCEFROMHOME, ENVIRONMENTSATISFACTION, JOBINVOLVEMENT, JOBROLE, JOBSATISFACTION
JOB AND CAREER PROGRESSION	JOBLEVEL, NUMCOMPANIESWORKED, TOTALWORKINGYEARS, YEARSATCOMPANY, YEARSINCURRENTROLE, YEARSSINCELASTPROMOTION, YEARSWITHCURRENTMANAGER,
COMPENSATION AND BENEFITS	MONTHLYINCOME, PERCENTAGESALARYHIKE, PERFORMANCERATING, STOCKOPTIONLEVEL, DAILYRATE, HOURLYRATE, MONTHLYRATE
WORK-LIFE BALANCE	WORKLIFEBALANCE, OVERTIME, NUMOFTRAININGLASTYEAR, RELATIONSHIPSATISFACTION