Effective Ways to Mitigate Employee Turnover

Cheng Qian (1266297), Yufeng Liu (1266263),
Pengjiabei Tang (1323020), Junheng Yu (1324718), Xuan Ji (1405130)

May 23, 2024

Contents

1 Introduction

In the changing realm of managing personnel, foreseeing voluntary departures, such as those in the Australian Defence Force (ADF), presents a significant challenge with profound impacts on strategic human resource planning and retention. The complexity arises from the factors influencing personnel decisions, such as age, rank, marital status, and years of service. Dealing with sparse and noisy datasets adds to the complexity from a data science perspective.

To address these challenges, the project aims to create a model using a comprehensive dataset of anonymized personnel information. This involves utilizing advanced machine learning techniques and operational research methods like the Cox hazard model to forecast which individuals are likely to leave in the next 6 to 12 months. The goal is to pinpoint the drivers behind separations and identify data that could improve prediction accuracy.

This predictive analysis not only enhances retention strategies within the ADF but also contributes to advancing data science by tackling issues such as data scarcity and noise, commonly found in extensive human resource datasets. The project's goal is to predict employee departures, offering insights for making informed decisions in military personnel management both in the short term and long term.

2 Related work

2.1 The impact of organizational culture on employee turnover

The interplay between organizational culture and employee turnover is critical in enhancing job satisfaction and reducing turnover. Leadership styles significantly influence employee retention, where toxic leadership reduces satisfaction and increases turnover [1]. Organizational culture is divided into Clan and Market cultures. Clan Culture promotes a team-oriented environment that fosters personal development and reduces turnover [2]. Conversely, Market Culture, focusing on competitiveness, may increase stress and reduce retention. Compatibility between an employee's capabilities and organizational culture affects their stress and retention. Adaptable employees thrive in team-oriented environments, while customer-focused individuals excel in competitive settings [1, 3]. Optimizing these aspects is vital for improving organizational performance.

2.2 How to retain employees of different ages

In 2009, Ng and Feldman mentions people treat turnover is different at different times where older people prefer the original job position and young people do not have a chance to improve.[4] Moreover in 2021, Rajapakshe studies a new turnover model and found only age, experience and[5] family income significantly affect labor turnover in demographic variables. Thus, we consider age could be one of the important factors of employee turnover. Therefore, we consider a research question: How to retain employees of different ages? In 2022, Bajaba, Azim, and Uddin considers work-family could impact turnover. Moreover, we consider that people have different families and working environments at different ages, thus we try to research what kind of environment people focus on at different ages. After exploring the impact of age on turnovers, the details will be displayed in the third section.

2.3 Effective ways to mitigate employee turnover

Work satisfaction is complex and multifactorial. It significantly influences how engaged and productive the employees are within a company. Research indicates a clear correlation between higher job satisfaction and lower turnover rates among employees in 2014 by Men.[7] This is especially distinct in industries such as information technology and hospitality where positive emotional experiences at work can decrease the likelihood of employee turnover by Emerald Publishing Limited.[8] Moreover, organizational emotional intelligence is essential for improving job satisfaction since it helps create a supportive and understanding work environment. Research in 2023 by Kanchana and Jayathilaka indicates when employees feel supported, they are more likely to participate in their organization and be satisfied with their jobs, then reduces the probability that they would leave the company.[9]

Addressing these issues calls for a comprehensive approach that goes beyond superficial solutions. Kanchana and Jayathilaka has revealed that utilizing regular anonymous employee surveys is noted as an effective tool in tackling these concerns since it plays a key role in understanding the thoughts and feelings of the team, helping pinpoint areas where improvements are necessary.[9] In ScienceForWork demonstrates that establishing clear communication is crucial as well. Such guidelines promote open discussions between staff and management, ensuring that feedback is not just received but actively acknowledged.[10]

2.4 Three most relevant variables correlating to turnover rate

This study identified three key variables that are closely related to employee turnover:

- **Career growth and promotion opportunities:** Employees often pursue promotions and salary increases, and may leave their current positions if they feel there are no opportunities for growth or advancement in their current positions[11].
- **Job satisfaction and employee engagement:** Numerous studies show that job satisfaction often has a significant impact on turnover intentions. Job satisfaction includes the nature of the job, the work environment, relationships with colleagues and superiors, and the congruence of the work with personal values.
- **Compensation and benefits:** The literature related to compensation and turnover intention emphasize that compensation is one of the most important factors influencing the decision to leave[12]. If employees feel their compensation is unfair within the company or significantly below industry standards, they may seek more competitive pay.

3 Data analysis and preliminary model development

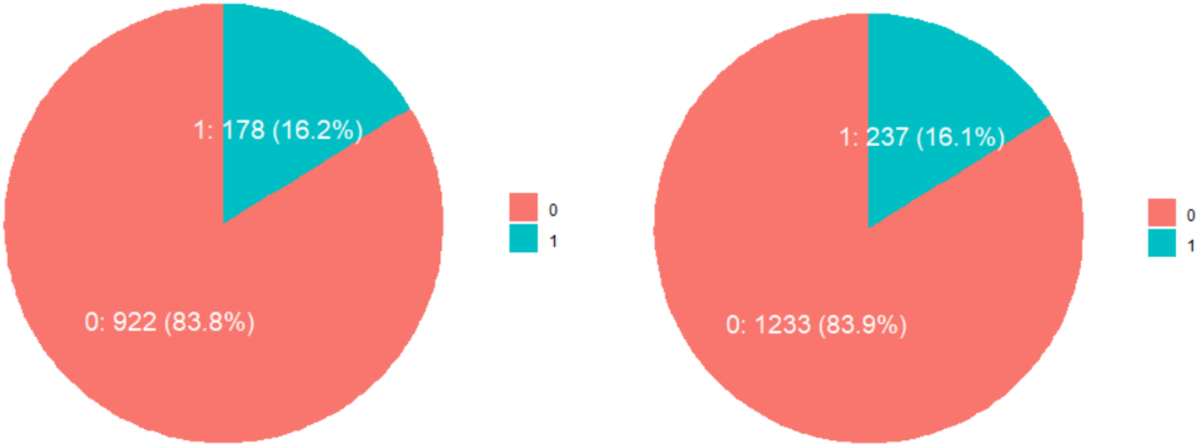
3.1 Data overview

We have not received the real datasets from our host, and we will actively communicate with them to obtain them in the next semester. Instead, we utilize two publicly available datasets from Kaggle for our analysis this semester, which were provided by the host.

Employee Separation Forecast Baseline Dataset (dataset 1) has 1000 records. IBM HR (dataset 2) Analytics Employee Attrition & Performance Dataset have 1500 records. We show the distribution of two datasets based on their labels. In the figure 1a and 1b, label 0 represents retention and 1 represents attrition.

We can see from the two datasets are imbalanced. This means that even if we predict all outcomes as retention, the model might still show a high accuracy due to the imbalance in the data.

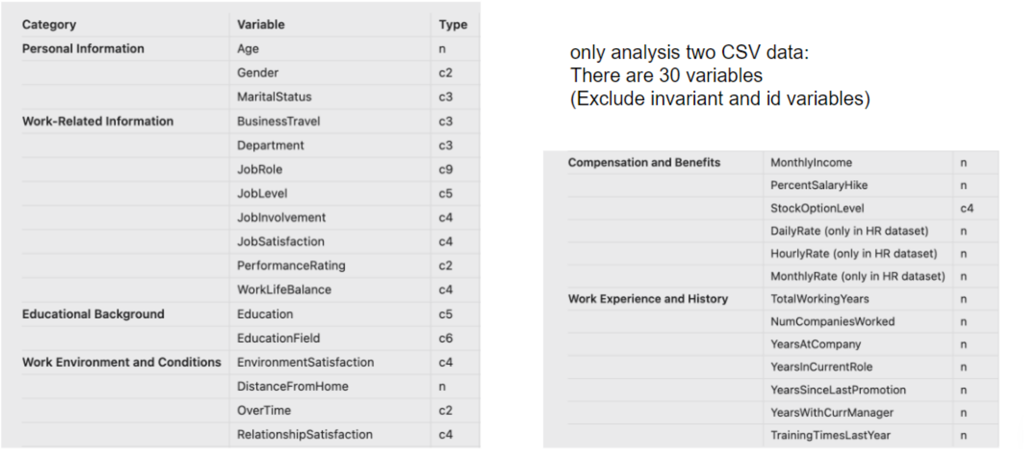
We have found the two datasets have a high degree of similarity. The introduction of variables is shown in figure 2.



(a) Dataset 1 distribution

(b) Dataset 2 distribution

Figure 1: Data distribution



only analysis two CSV data:
There are 30 variables
(Exclude invariant and id variables)

Category	Variable	Type
Personal Information	Age	n
	Gender	c2
	MaritalStatus	c3
Work-Related Information	BusinessTravel	c3
	Department	c3
	JobRole	c9
	JobLevel	c5
	JobInvolvement	c4
	JobSatisfaction	c4
	PerformanceRating	c2
	WorkLifeBalance	c4
Educational Background	Education	c5
	EducationField	c6
Work Environment and Conditions	EnvironmentSatisfaction	c4
	DistanceFromHome	n
	OverTime	c2
	RelationshipSatisfaction	c4
Compensation and Benefits	MonthlyIncome	n
	PercentSalaryHike	n
	StockOptionLevel	c4
	DailyRate (only in HR dataset)	n
	HourlyRate (only in HR dataset)	n
	MonthlyRate (only in HR dataset)	n
Work Experience and History	TotalWorkingYears	n
	NumCompaniesWorked	n
	YearsAtCompany	n
	YearsInCurrentRole	n
	YearsSinceLastPromotion	n
	YearsWithCurrManager	n
	TrainingTimesLastYear	n

Figure 2: Data variable

3.2 Data cleaning and pre-processing

In preparing our dataset for detailed analysis, we focused on ensuring data integrity and suitability through systematic cleaning and pre-processing. This involved addressing missing values and transforming categorical data into a numerical format, which are essential steps to make the dataset analytically viable.

- Missing values handling: We confirm no null values in data.
- Data transformation: The categorical variables were transformed into numerical values, using the One-Hot Encoding method to encode all categorical variables, and then merged back into the main dataset. By checking the data types, it was confirmed that all features have been correctly encoded and integrated, preparing for subsequent analysis tasks.

By these preprocessing steps not only preserve the data's integrity but also enrich its structure ready for further analysis.

3.3 Exploratory data analysis (EDA)

3.3.1 Features importance selection

In our preliminary features analysis, we have more than 30 features in our data. Thereby we have decided to select the top 10 most important features (figure 3) for turnover analysis. By focusing on these key features, we can simplify our data analysis processes, ensuring our subsequent analysis is efficient and targeted.

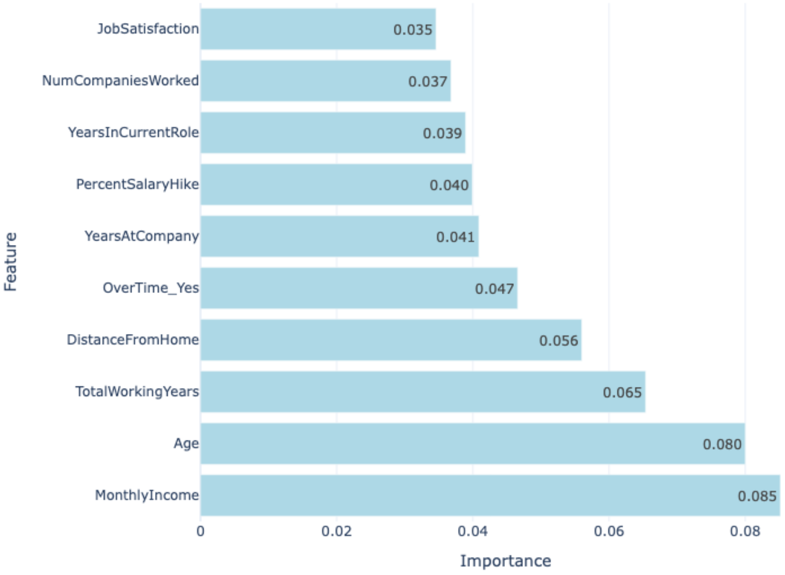


Figure 3: Bat chart of top 10 feature importances

We utilized a Random Forest Classifier, a machine learning method that is particularly good at handling a large number of features and identifying important ones. After training the model on our dataset, we extracted the feature importance by gathering the importance scores from each tree, and then gained a reliable estimate of each feature's contribution to the prediction.

The graph presents a clear visualization of the most influential factors determined by our predictive model. It illustrates that `Monthly Income` holds the highest importance score of 0.085. Following closely are `Age` and `Total Working Years`, with importance scores of 0.080 and 0.065 respectively.

3.3.2 Statistical summary for top 3 features

We performed a detailed statistical summary of the three most influential features: Monthly Income, Age, and Total Working Years. The box plots (figure 4) show the distribution characteristics of each feature.

- **Monthly income:** the data spread from a minimum of 1009 to a maximum of 19999 AUD, with a median income of 4857 AUD, while the interquartile range (IQR) from 2924.5 to 8354.5, indicating significant variability in employee earnings.
- **Age:** shows a narrower range, from 18 to 60 years, with a median of 36 years. The majority of employees are clustered between 30 and 43 years, as seen in the box plot, suggesting a workforce with a considerable proportion of mid-career professionals.

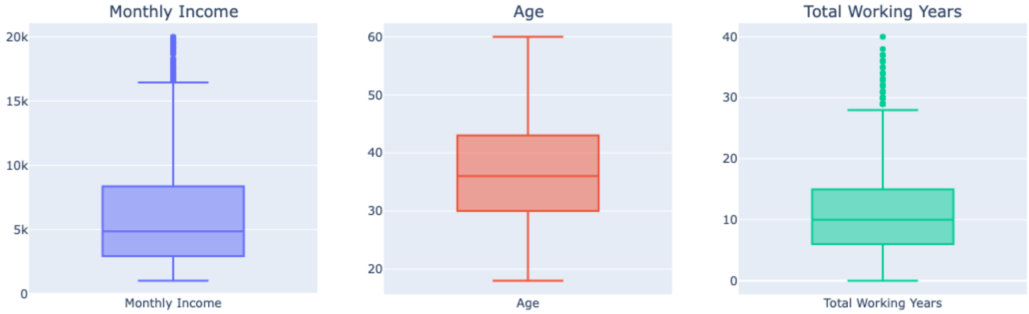


Figure 4: Box plots of monthly income, age, and total working years

- **Total working years:** varies from 0 to 40, with half of the workforce having between 6 and 15 years of experience, indicating a relative majority of seasoned employees.

These findings offer a foundational understanding of employee distribution, which is essential for further analysis and strategic employee retention planning.

3.3.3 Features correlation analysis

The heatmap graph (figure 5) of the features shows the correlations between variables in our dataset interacting with each other, indicating both strong and weak associations. Values closer to 1 indicate a strong positive correlation, while values closer to -1 indicate a strong negative correlation.

It's crucial to consider this interaction between variables during subsequent modeling phases. It is not only helpful for model accuracy but also assists with the management suggestion in the final project.

3.4 Model construction

3.4.1 Naive bayes

1. **Principle and implement**
   Naive Bayes is a type of probabilistic classifier in which one uses the known conditional probabilities to predict the probability of unknown events. Conditional probability is the probability of an event occurring when another condition is known. In simpler terms, once we know an employee is working for a specific department, we can find the probability of him leaving the company.

2. **Results and discussion**
   After implementation, we observed that the accuracy produced by the Naive Bayes model is 0.68, which is far less than the baseline model's accuracy of 0.83. This result is due to the poor distribution of our training data. As indicated, the "Stay" proportion was high at 0.83, while the "Leave" proportion was only 0.17. This data distribution highly distorts the prior probabilities of the Naive Bayes model. Furthermore, as the Naive Bayes model assumes independence, features that correlate highly with others must be removed from the pool. As shown in Figure 5, the "Job level" feature has a strong correlation to that of "Monthly Income" because the higher one goes in the pyramid construct, the higher the salaries. Additionally, there are still many variables

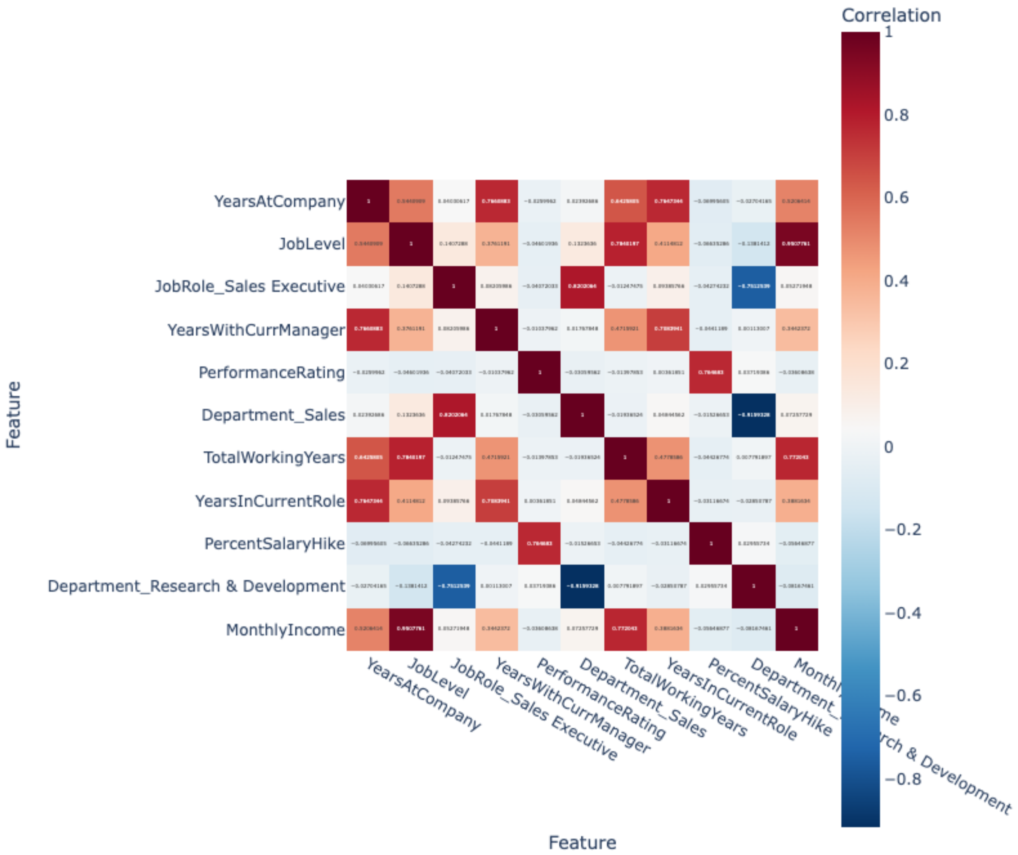


Figure 5: Heatmap of strong correlations between features

that are interrelated. Upon eliminating these interdependent factors, the accuracy increased marginally to 0.77 but still pales in comparison to the baseline model.

In light of the obtained results of the conducted analyses, we can conclude that the Naive Bayes model is not the most suitable approach to the given case. The model's main disadvantages, including hypersensitivity to imbalanced data and an unreasonable assumption about the independence of features, suggest that other machine-learning models are more appropriate. Even in case we intended to work with the Naive Bayes model, we could try to unnecessarily balance the dataset, which would cause another set of problems. As a result, the Naive Bayes model, while allowing for easy classification operations, has numerous drawbacks that need comprehensive analysis, specifically in the case of overcoming the limitations that are characteristic of the imbalanced and non-independent datasets. Therefore, additional research on other models should be conducted.

3.4.2 Support vector machines

1. **Principle and implement**
   Support Vector Machines (SVM) is a powerful classification method that finds the optimal hyper-

plane to separate data points into different classes. It is particularly effective in high-dimensional spaces and can handle linear and non-linear separations with different kernel functions. In our analysis, we handled over 30 variables and used one-hot encoding for categorical data and feature scaling to standardize the data. We implemented the SVM with a "linear" kernel, which effectively managed the complexity of the multidimensional data and ensured efficient convergence.

2. **Results and discussion**

   - **Validation Accuracy:** Achieved 85% accuracy. However, the recall for the positive class (employees predicted to leave) was relatively low at 31%, which means that the model may miss a significant number of actual attrition cases.
   - **Test Accuracy:** Achieved a higher accuracy of 89.25%. The precision for predicting attrition was 83%, and recall was 42%, reinforcing the problem regarding the model's sensitivity to the minority class.

   Moving forward, the focus will be on the imbalance in the dataset, possibly by using SMOTE for oversampling or other techniques to undersampling the data. Additionally, testing with different kernel functions like `rbf` or `poly` and tuning the model parameters.

3.4.3 XGBoost

1. **Principle and implement**
   XGBoost, standing for eXtreme Gradient Boosting, is used popularly for classification and regression problems. It is based on the Gradient Boosting framework, where a sequence of decision trees is built serially to improve the prediction. We built our model with XGBoost for the case of predicting attrition for employees. We perform feature engineering by removing irrelevant features and retaining only the important ones to be used for better performance and interpretability. We also did hyperparameter tuning so that the parameters of the model can be optimized.

2. **Results and discussion**
   In these two datasets, the accuracy of the XGBoost model is 0.82 and 0.84 respectively. This means that in the first dataset, the model correctly predicted 82% of the test samples; in the second dataset, the model correctly predicted 84% of the test samples. The higher the accuracy of the model, the stronger the overall prediction ability of the model, that is, it can well distinguish between resigned employees and non-resigned employees.

   But in the first set of data, the model had a positive class recall of 0.33, which means the model accurately predicted only 33% of employees who actually left the company. The second dataset is also the same situation. So high precision but low recall may mean that the model is more likely to predict that employees will not leave, thereby reducing false positives (i.e., false positives), but at the cost of ignoring many employees who actually leave (i.e., false negatives).

3.4.4 Logistic regression

1. **Principle and implement**
   Logistic regression is used for statistical analysis or machine learning. It can predict the probability of the response variable. The response variable has to be binary. It is different from linear regression which predicts a continuous value. It is used to predict the probability that the binary outcome will occur. it will express the value between 0 and 1. In this experiment, we used logistic regression to predict employee attrition (denoted as Label). First, we used some data from

"WAFn-UseC-HR-Employee-Label" as the training dataset, and then we removed the insignificant features and kept the most significant features. This can help increase the interpretability and the efficiency of the prediction.

2. **Results and discussion**
The recall of the model is 61.54%, the recall is 47.62% and the accuracy of the model is 87.45%. From the recall, we can find that there is improvement in finding all the employees who may want to leave the company. Overall, it is a good model, because it has high accuracy. But it has better predictions on non-leavers than leavers. Therefore, it shows that it has some limitations in sensitivity. To improve this we may introduce more features, creating the decision threshold.

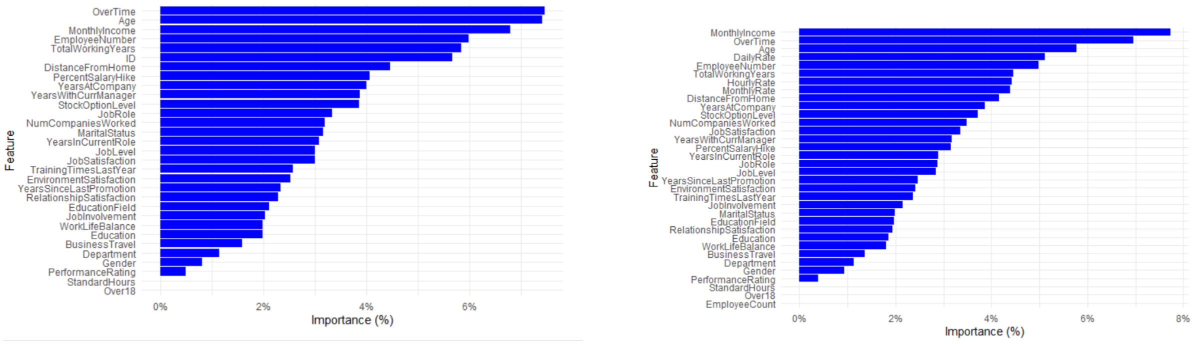
   Model Limitations: It assumes the features have a linear relationship with the response which may not be true in the real world If the data is imbalanced, the result may be affected a lot and lead the researcher to the wrong place.

3.4.5 Random forest

1. **Principle and implement**
In 2001, Breiman propose a Random Forest(RF) algorithm used to predict the result based on the given dataset. [13] RF combines multiple decision tree that are regarded as a specialist and vote for the final result.

   As we mentioned before, the datasets are unbalanced, thus we use oversampling to fit the data. Oversampling increases the number of samples in the minority class to match the majority class. Then use a Random forest to fit the two datasets.

2. **Results and discussion**
The accuracy of the model is 0.966 and 0.972. Moreover, we explore the importance of different variables and the plot is in the following plots. The figure 6a and the figure 6b show the importance of the variable in the first and second datasets. The equation of importance(%) is each variable's importance/sum of variables importance.

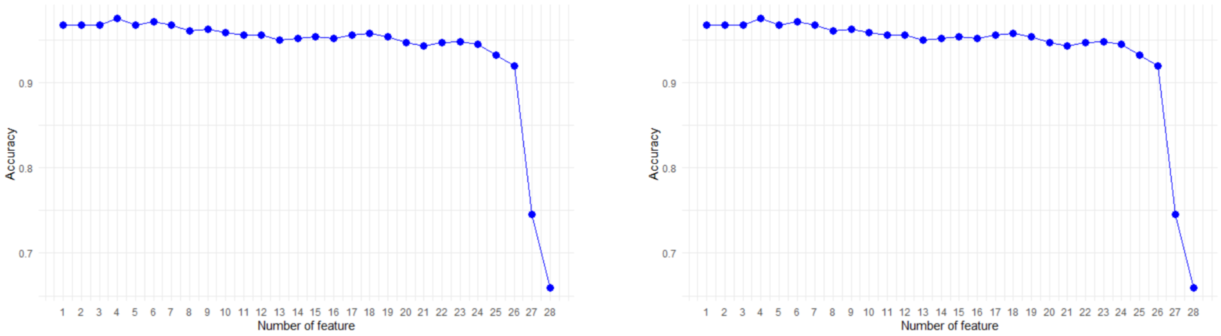


(a) Importance for the first model

(b) Importance for the second model

Figure 6: Comparison of model importances

Moreover, we found some variables shown to be less important, thus we try to remove the feature with the least importance one by one and create a RF. The figure 7a is the accuracy of model for dataset 1 and the figure 7b is for dataset 2. In figure 7a and 7b. the number of feature K means removing K feature with the least K based on the sort of importance and the accuracy is the model without k feature. From the figure 7a shows it could be removed 24 features and the accuracy. From the figure 7b shows it could remove 29 features and the accuracy. These results are caused by imbalanced datasets that could lead to high accuracy, though we use over sample method, the model still learns the same information.



(a) Accuracy of the first model

(b) Accuracy of second model

Figure 7: Comparison of model accuracy

4 Proposal for semester 2

There will be some challenges in next semester. Firstly, the datasets we used are imbalance and lead to high accuracy of the models. Secondly, the model we build is based on kaggle datasets. Thus if we use those models to predict the real dataset that comes from the host, we may have some problems like missing some features in the real dataset and the importance of features may not be the same due to different industries. Therefore, in next semester, we will try different resample methods to solve the imbalance problems of the dataset, and build a final model. Then, we will optimize the model by inputting real data into the model and comparing it with real situations.

Our related work mentions four research questions. In the following working, we will focus on the question ”Effective ways to mitigate employee turnover”. This question is using the Kaggle datasets to build the models and optimise the models by inputting the real data to validate the model. Moreover, by reading the literature, we know the correlation between some features and turnover such as age, and job satisfaction. Thus, in the following step, we know how to adjust the weights of the variable in the models.

4.1 Timeline

This figure 8 is the Timeline and milestones for the semester2. In this figure:

- DI: Solve the data imbalance problems
- DR: Write the draft of our final report
- DM: Develop a final model
- OM: Input the real data and optimize the models
- MR: Modify final report

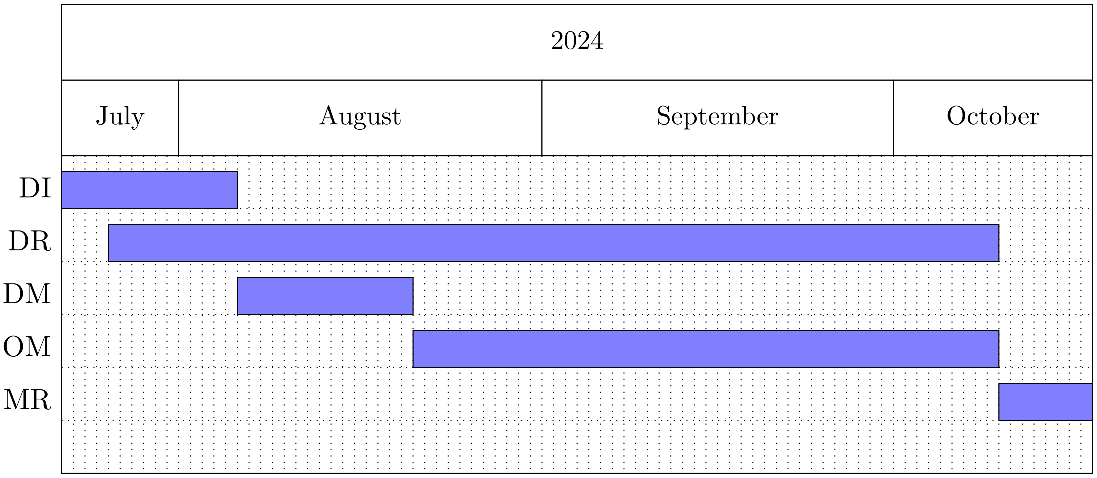


Figure 8: Project Gantt Chart

This role of each group member in semester 2 is following as table 2.

Member	**Role**
Junheng Yu	• Team leader, takes charge of overall project planning, coordination and execution • Optimize Random Forest model when using real data
Cheng Qian	• Preparing agenda and slide for each meeting • Optimize Naive Bayes model when using real data
Pengjiabei Tang	• Handling communications within the host and entire team • Optimize SVM model when using real data
Xuan Ji	• Record and summarize the meeting minutes • Optimize the XGBoost model when using real data
Yufeng Liu	• Handling communications with supervisor • Optimize Logistic Regression model when using real data

Table 1: Team role for semester2

5 Contribution

Junheng Yu	Pengjiabei Tang	Xuan Ji	Yufeng Liu	Cheng Qian
0.2	0.2	0.2	0.2	0.2

Table 2: Contribution for each member

References

[1] Marissa Brouwers and Amelda Paltu. “Toxic leadership: Effects on job satisfaction, commitment, turnover intention and organisational culture within the South African manufacturing industry”. In: *SA Journal of Human Resource Management* 18.1 (2020), pp. 1–11.

[2] Victoria O Akpa, Olalekan U Asikhia, and Ngozi Evangeline Nneji. “Organizational culture and organizational performance: A review of literature”. In: *International Journal of Advances in Engineering and Management* 3.1 (2021), pp. 361–372.

[3] Jina Kim and Hye-Sun Jung. “The effect of employee competency and organizational culture on employees’ perceived stress for better workplace”. In: *International Journal of Environmental Research and Public Health* 19.8 (2022), p. 4428.

[4] Thomas WH Ng and Daniel C Feldman. “Re-examining the relationship between age and voluntary turnover”. In: *Journal of Vocational Behavior* 74.3 (2009), pp. 283–294.

[5] Wasantha Rajapakshe. “Determinant factors of labor turnover–a new perspective”. In: *Journal of Economics, Management and Trade* 27.5 (2021), pp. 19–35.

[6] Saleh Bajaba, Mohammad Tahlil Azim, and Md Aftab Uddin. “Social support and employee turnover intention: The mediating role of work-family conflict”. In: *Revista brasileira de gestão de negócios* 24 (2022), pp. 48–65.

[7] Linjuan Rita Men. “Strategic internal communication: Transformational leadership, communication channels, and employee satisfaction”. In: *Management Communication Quarterly* 28.2 (2014), pp. 264–284. URL: `https://pubmed.ncbi.nlm.nih.gov/22213478/`.

[8] Emerald Publishing Limited. “Factors that impact on employee turnover intentions: How transformational leadership can help”. In: *Development and Learning in Organizations* 36.4 (2022), pp. 41–43. DOI: `10.1108/DLO-01-2022-0016`.

[9] L. Kanchana and R. Jayathilaka. “Factors impacting employee turnover intentions among professionals in Sri Lankan startups”. In: *PLOS ONE* 18.2 (2023), e0281729. DOI: `10.1371/journal.pone.0281729`.

[10] ScienceForWork. *Employee Turnover: How to become a manager that people don’t want to leave.* Retrieved from `https://scienceforwork.com/blog/employee-turnover-manager/`. n.d.

[11] James L Price. “Reflections on the determinants of voluntary turnover”. In: *International Journal of Manpower* 22.7 (2001), pp. 600–624.

[12] Muhammad Nawaz and Faizuniah Pangil. “The relationship between human resource development factors, career growth and turnover intention: The mediating role of organizational commitment”. In: *Management Science Letters* 6.2 (2016), pp. 157–176.

[13] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.