Instructions:
Final Report (15 points): Due May 16, 2025 11:59 pm via Brightspace
Submission Requirements:

- One final report (excluding figures, tables and references) should be about **4-5 pages** with 11-point font.
- Do not show Python code in the report. The report only conveys your work and results.
- The report can be typed using any word processor you prefer, and it should not contain Python raw outputs except figures.
- Cite any references used in your project in the reference section.
- Turn in Python code to Blackboard as a separate file. The report will not be graded without code.
- Submit the used data set to Blackboard. All results in the report should be able to be reproduced using Python code and data.

Final Term Project Report should contain the following items:

- Introduction and background
- Motivation of your research question (e.g. Why do you think it is an important question to solve?)
- Dataset description and variable introduction
- Data summary statistics (e.g. summary table, scatterplot matrix, boxplots, histograms)
- Data mining method description (e.g. why do you choose this data analytic approach?)
- Adopt appropriate methods and measures for model evaluation. Use figures or tables to show the results and the model performance.
- Make conclusions following logically from results and findings
- Practical implications (e.g. how do your results apply to the real world? or how the company or society or customers can benefit from your results?)

Proposal:
https://docs.google.com/document/d/1-MJDoze-6IlJwwmCrnUE408qwQAuaT3Shtnhud3dDXM/edit?tab=t.0

CIS 9660 Data Mining
May 16, 2025

Employee Attrition Final Report

Group 2: Jaime Bunay, Siddarth Bhagirath, Jonathan Caussin, Nisargvan Goswami, Hung Hsin Li, Edosa Odia

Introduction and Background

Turnover peaked during the "Great Resignation" in 2022 but has been on the decline since. According to Mercer, "the average voluntary turnover rate in the US from 2023 to 2024 was 13.5%" compared to 24.7% in 2022 and 17.3% in 2023 (Mercer, 2024). While employee attrition rates can be a product of economic conditions and changes in the job market, what are the factors within a company's control that can be used to flag and retain employees despite the many external factors?

Motivation

The goal of our analysis is to uncover factors that predict the most significant levels of employee attrition. We believe that understanding attrition will help both employees and employers. Insights from this data mining analysis can be used to improve job satisfaction for employees, reduce turnover costs for employers, and build a better workplace culture overall. Employee turnover can be a huge issue for companies because of the following costs of hiring a new employee. The average turnover cost per employee is around six to nine months of an employee's salary and in the case of a highly trained employee, even two times the employee's annual salary (SRHR, 2017). Identifying early signs of departure can help companies take proactive measures to retain talent and mitigate costs.

Data Set and Variables

We used the Employee Attrition dataset from Kaggle, which contains 74,498 samples of anonymized employee records. The data includes a binary target variable for attrition (0 = stayed, 1 = left) and several variables that can be used as predictors.

These variables cover four main categories about an employee:

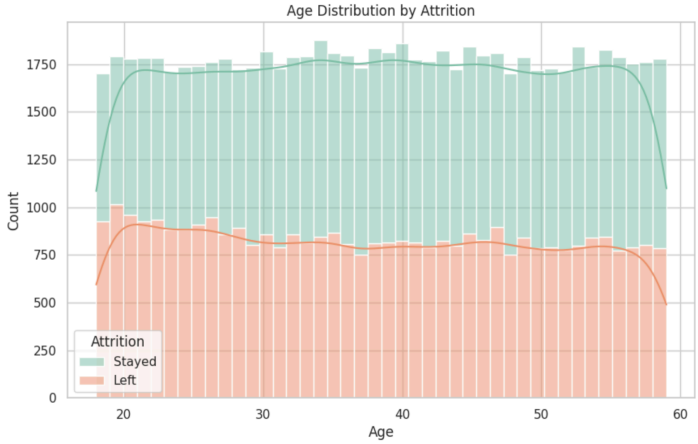
- Demographics: Age, gender, marital status, number of dependents
- Job Characteristics: Role, level, monthly income, years at company, remote work status
- Performance: Satisfaction, performance rating, work-life balance, number of promotions
- Company Features: Company size, company reputation, employee recognition

We will be using this dataset to explore both personal and organizational factors contributing to employee attrition from companies.

Data Summary Statistics

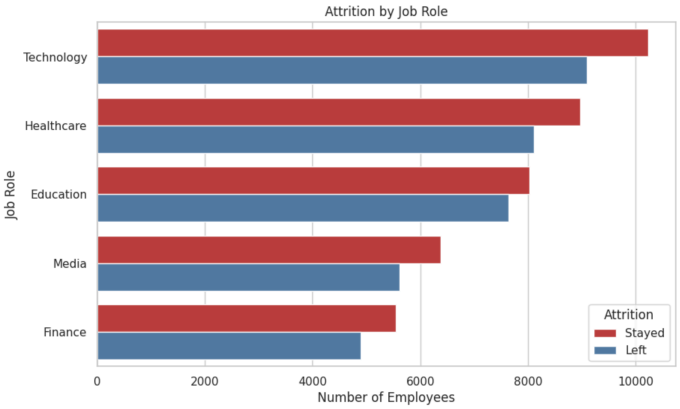
(e.g. summary table, scatterplot matrix, boxplots, histograms)

- Refer to Slides 4-5



Age Distribution by Attrition

A discernible pattern in attrition behavior can be seen in the age distribution histogram. Younger employees with a greater concentration between the ages of 20 and 35 are more likely to have departed the organization. Employees over 40, on the other hand, are more likely to stick with the organization. This implies that younger workers might be more likely to look for better chances, continue their education, or become dissatisfied with their current positions. Stronger employment stability or stronger organizational loyalty among older employees may potentially be the cause of the attrition trend's decline with age.



Work-Life Balance and Attrition

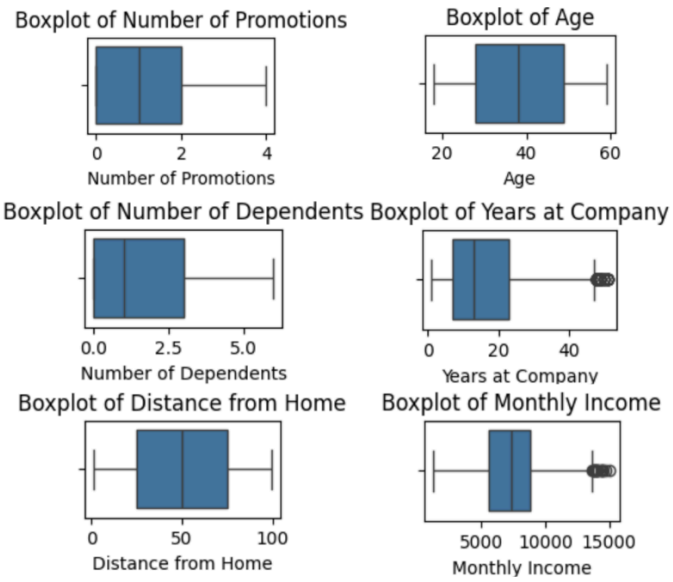
Perceived balance and the probability of leaving the company are clearly correlated, as shown by the bar plot of Work-Life Balance Rating vs. Attrition. Compared to workers who reported "Good" or "Excellent" work-life balance, those who assessed it as "Poor" or "Fair" had a much greater attrition rate. Notably, as the rating deteriorates, attrition rates rise while the number of remaining employees significantly declines.

The significance of preserving a positive work-life balance in employee retention tactics is shown by this trend. The higher departure rates among those who gave worse ratings point to discontent that may be brought about by excessive work hours, rigid scheduling, or high levels of workplace stress. On the other hand, workers who reported a "Good" or "Excellent" balance were more likely to stick around, suggesting that a flexible and encouraging work environment may have a direct impact on worker loyalty and happiness.

Attrition by Job Role

Attrition across various professional activities is further contextualized by the Attrition by Job Role chart. Compared to other roles, technology and healthcare have the greatest total employee numbers and the highest attrition rates, according to the data. This might be a result of these industries' high levels of pressure, increased job market mobility, and need for qualified workers in these fields.

However, attrition rates are more evenly distributed in industries like media, finance, and education, with a smaller difference between those who left and those who stayed. This implies that rather than intrinsic job discontent, attrition in these industries may be more impacted by external market variables or individual career advancement.



Boxplot Analysis

Boxplots were used to visualize key numerical features related to attrition:

- **Promotions:** Most employees had 0–2 promotions. Those who left typically had fewer, suggesting limited advancement may lead to attrition.
- **Age:** Median age is around 35–40. Younger employees showed higher attrition, especially in lower age quartiles.
- **Years at Company:** Most attrition occurs within the first few years, indicating early retention is crucial.
- **Dependents:** Most employees had 0–2 dependents, with no major difference between groups—likely not a strong attrition factor.
- **Distance from Home:** Wide range observed. While not a direct cause, longer commutes may contribute to dissatisfaction.
- **Monthly Income:** Skewed toward lower incomes. Employees with lower pay were more likely to leave, highlighting compensation as a key factor.

Data Mining Methods

Data mining method description (e.g. why do you choose this data analytic approach?)

- Refer to Slide 6

Model Selection: Predictive Modeling of Employee Attrition

To predict whether an employee would stay or leave, we treated **Attrition** as a **binary classification problem** (0 = stayed, 1 = left). We selected several classification models for comparison:

- **Logistic Regression** – A simple, interpretable model used as a baseline.
- **Decision Tree** – Captures non-linear relationships using feature-based splits.
- **Random Forest** – An ensemble of decision trees that improves accuracy and reduces overfitting.
- **Support Vector Machine (SVM)** – Finds the optimal boundary between classes, effective for complex data.

These models were trained on the same dataset to determine which best predicts attrition while balancing accuracy and interpretability.

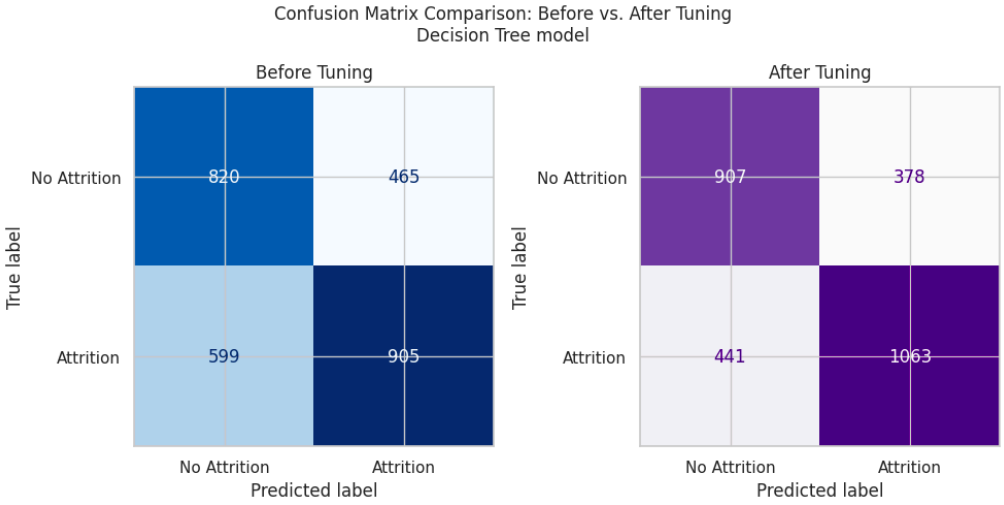
Model Evaluation

Adopt appropriate methods and measures for model evaluation. Use figures or tables to show the results and the model performance.

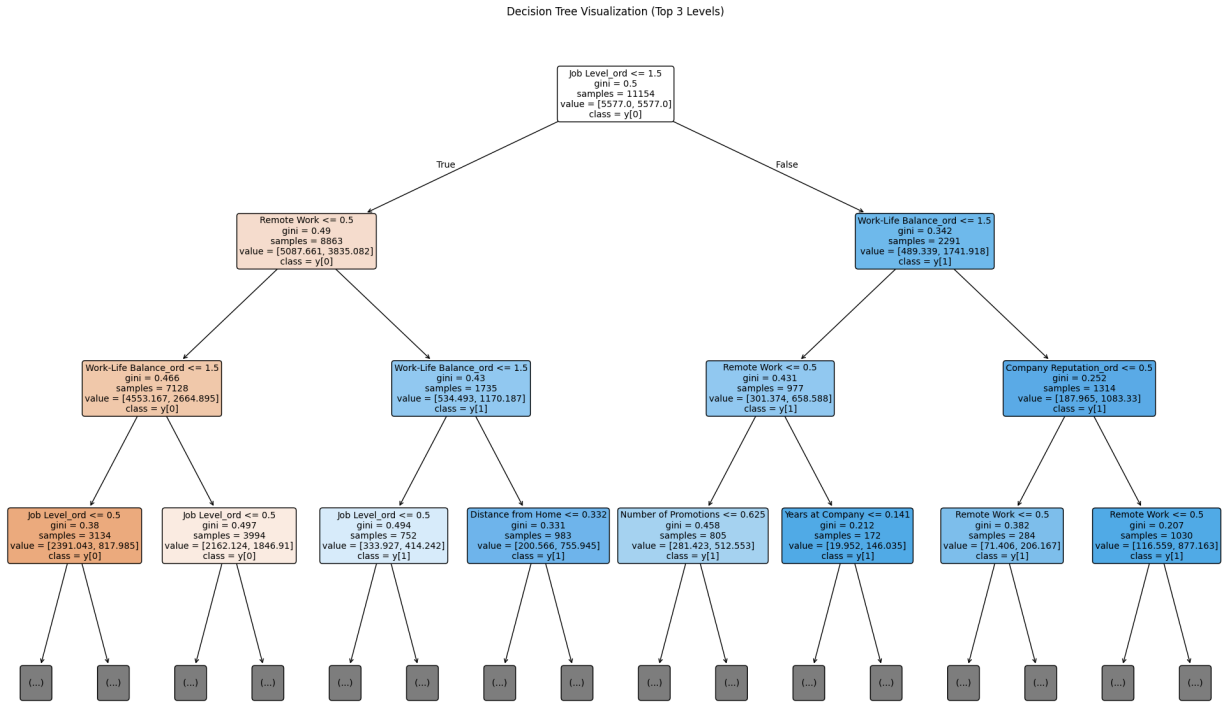
- Refer to slide 7-13

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	71%	74%	71%	72%
Random Forest	71%	76%	69%	72%
Logistic Regression	72%	75%	72%	73%
SVM	71%	77%	67%	72%

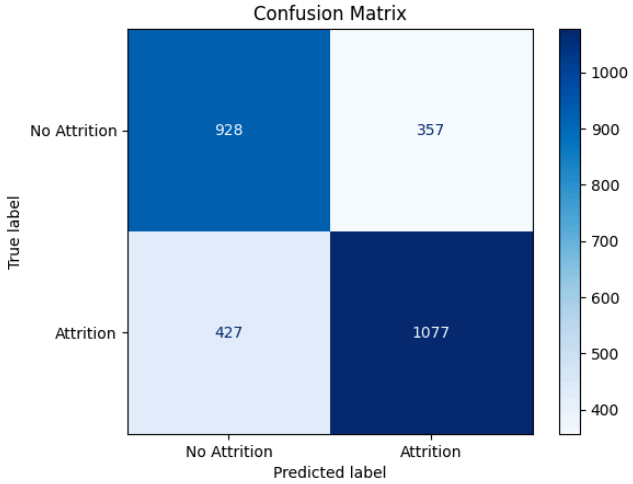
First, our Decision Tree modelling yielded us an overall accuracy of 62%, which is lower than our other models, but still informative. When we look at the breakdown by class, we can see a difference in how well the model identifies each group. For Class 0, which represents employees who stayed, the precision was 0.58, and recall was 0.64, meaning the model correctly identified 64% of the stayers but also misclassified quite a few as likely to leave. For Class 1, or employees who left, we see improved performance with precision at 0.66 and recall at 0.60. This means the model was more effective at identifying actual attrition cases than it was at identifying those who stayed. From the confusion matrix below, we can also see that the model initially predicted 599 correct attrition cases. After tuning, as shown in the matrix on the right, this improved to 1063 correct identifications, while reducing the number of false positives for Class 0. Although Decision Trees are more prone to overfitting, our tuning efforts helped them become more reliable, especially for predicting who will leave, which is the focus of our project. This makes Decision Trees a good interpretive tool for human resources departments to understand data-backed insights into employee attrition.



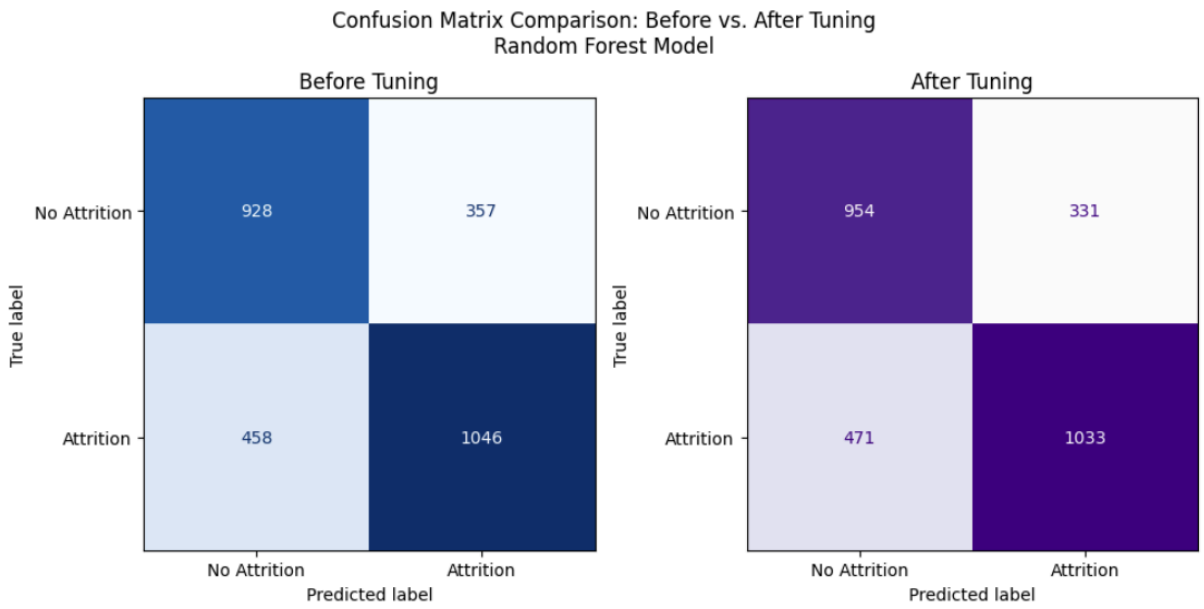
In addition to the decision tree confusion matrix, the visualization is also important to mention. The decision tree modelling visualization most importantly described the hierarchy of attrition within the top three job levels. Node visualization helped us understand that employees at a job level less than or equal to 1.5 nodes are more likely to leave, while employees at a work-life balance level rating of “poor” or “average” are more likely to be at higher risk of leaving.



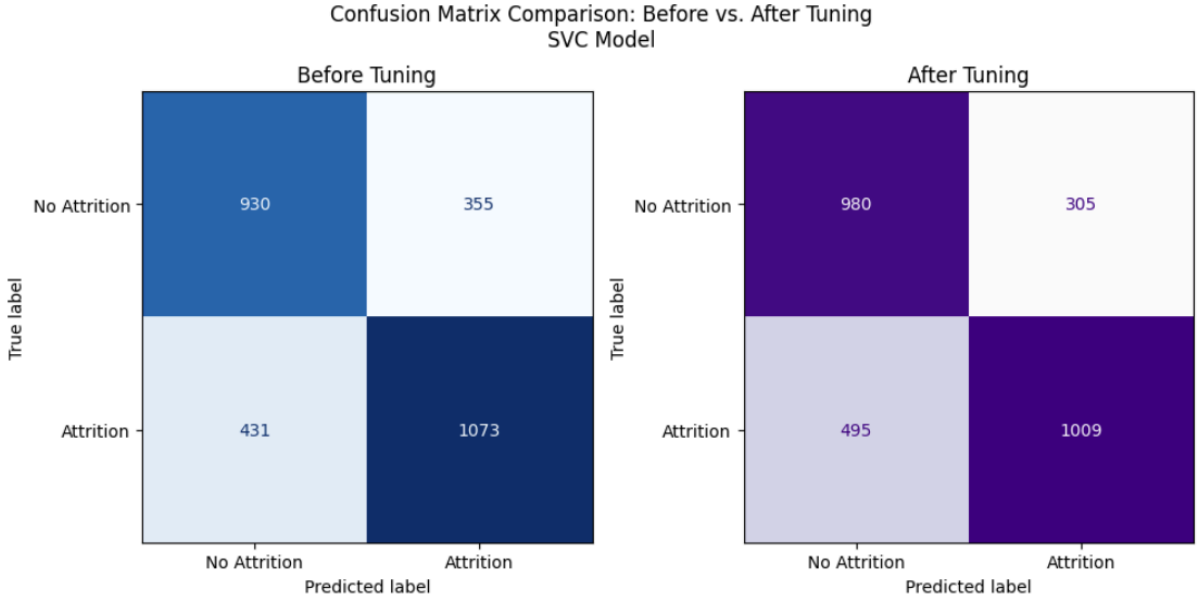
Third, our Logistic Regression model achieved an overall accuracy of 72%, which was the highest among all the models we tested. In this case, Class 0 represents employees who stayed, and Class 1 represents those who left the company. For Class 0, we see a precision of 0.68 and a recall of 0.72, which means the model correctly identified 72% of employees who stayed. For Class 1, the precision rises to 0.75, with an equal recall of 0.72, and an F1 score of 0.73, showing strong performance in predicting attrition. This is also supported by the confusion matrix, where we correctly identified over 1,000 attrition cases , more than any of the other models. When we compare the class-level performance, the model correctly identified 75% of those who left, compared to 68% of those who stayed. This tells us that Logistic Regression is particularly useful when the goal is to accurately detect potential employee turnover.
And because Logistic Regression is also interpretable, with coefficients that can directly show the weight of each feature, this model gives us both strong performance and useful insights for decision-making in the workplace, when for instance HR teams can use this data to not only predict attrition but understand what catalyzes it.



The fourth model we explored was the Random Forest classifier. Unlike the single tree used in the decision tree model, this model leverages bootstrapping to generate multiple decision trees, each is trained on different subsets of the training data. This approach resulted in a more robust model, achieving over 70% accuracy even before hyperparameter tuning. While accuracy showed only minimal improvement after running grid search, it's important to note that cross-validation was limited to 5 folds. However, to truly assess model performance, we need to look beyond accuracy and focus on precision and recall, metrics that are more meaningful when making business decisions, especially related to resource allocation. From the confusion matrix, we can see that the model performs better at correctly classifying class 0 compared to the decision tree and logistic regression models. However, this comes with an increase in false negatives. Whether this trade-off is acceptable depends on the specific business goals.

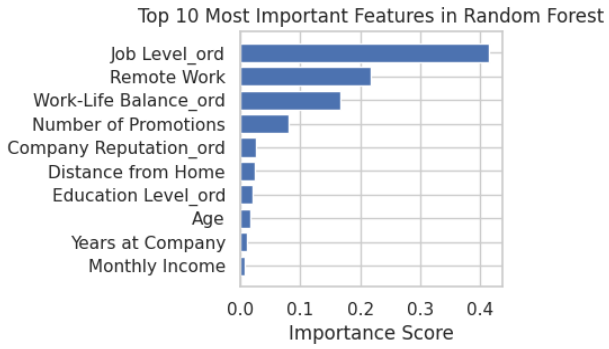


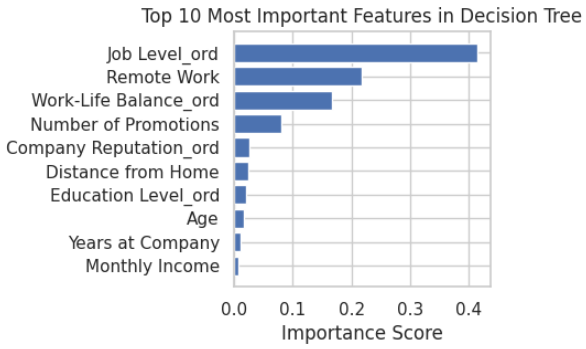
The last model we tested was a Support Vector Machine. Before running a grid search to optimize the model, it achieved an accuracy of 71.28%. However, this dropped slightly after tuning the hyperparameters. Once again, a helpful way to compare model performance is by looking at the confusion matrix. From the results, we can see that out of all the tuned models, the Support Vector Machine is the least effective at correctly identifying class 1. It has the highest number of false negatives, which could be a concern if the goal is to allocate more company resources toward addressing employee attrition.

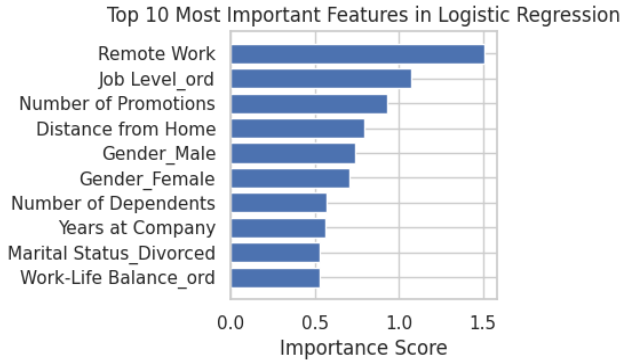


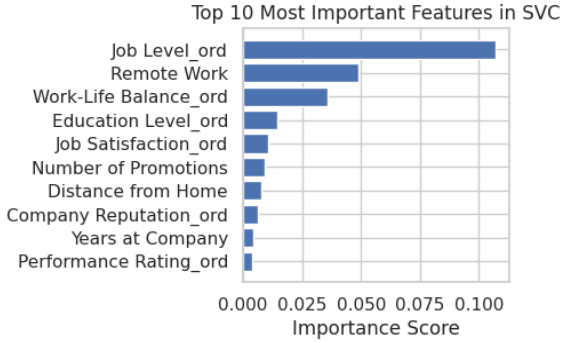
Feature Importances and Testing

Through our four models we were able obtain the most important features that contributed to attrition. Throughout our models: job level, remote work, and work life balance were consistently described as most important. To validate these findings, we retrained each model on data that excluded employees who ultimately left and then tested solely on those withheld cases. Fortunately, each model accurately predicted that the employee would leave the company, proving our models ability to accurately predict attrition.









Conclusion and Practical Implications

As stated previously our analysis has highlighted three critical drivers of employee attrition: job level, remote work options, and work life balance. Entry-level employees are at the

highest risk of leaving while limited options for remote work and long commutes further add to the quit risk. In order to fully address these problems companies should first invest in their entry level employees. Rotational programs, mentorship programs, and transparent promotion tracks seem necessary if companies want to inspire loyalty and bolster engagement among early career hires. Next, companies should integrate remote work or hybrid policies in order to fit employee preferences. If that is not possible, employers should consider commuter stipends, flexible start/end times, and subsidized on-site meals in order to reduce the stress that work naturally brings upon a person. By addressing these areas employers can potentially stop attrition and foster a more satisfied work-force.

References

Mercer. (2024). How much turnover is too much?. Retrieved from https://www.imercer.com/articleinsights/workforce-turnover-trends#:~:text=What%20are]%20the%20current%20turnover,turnover%20rates%20at%20just%209.1%25.

SHRM. (2017). Essential elements of employee retention. Retrieved from https://lrshrm.shrm.org/blog/2017/10/essential-elements-employee-retention.