**Slide 1:**

Hello Everyone. We are ECDS group 3 consisting of Marenda, SunJin and Raksha and this is our SC1015 mini project where we will be predicting employee promotion using machine learning. We will be using this source taken from kaggle.

**Slide 2:**

Promotion is a key milestone in an employee’s career and it affects one’s mrolace, salary and career path. Although thousands of employees are eligible for each promotion cycle in large companies, 67% of employees in 2024 viewed performance evaluations as unfair and say they do not trust the fairness of the process. This creates mistrust in the system.

**Slide 3:**

This brings us to question: how can HR teams fairly and efficiently identify high-potential employees and how can employees understand where they stand and how they can improve towards promotion?

**Slide 4:**

This leads to our problem definition: how do different variables affect whether an employee is promoted in different departments?

**Slide 5:**Here, as seen in the slide, our data set consists of employees from 9 different departments and we assumed that different departments may have different criteria for promotion. Hence, we decided to analyse the top 3 departments, which are Sales & Marketing, Operations and Technology.

We explored the different variables provided by our dataset and this slide here shows the variables, split into categorical and continuous variables.

**Slide 6:**

We then conducted exploratory data analysis using bar chart, mosaic plots and cramer’s v values for categorical data and box plots, correlation coefficient values and heatmap for continuous data.

**Slide 7:**

We performed EDA for each of the 3 earlier mentioned departments individually. For this video, we will focus on the sales & marketing department. The process for the other 2 departments was the same and more information is found in our jupyter notebook.

**Slide 8:**

Among the categorical variables, only previous\_year\_rating and awards\_won? showed a meaningful relationship with is\_promoted after we analysed the bar chart and mosaic plots.

**Slide 9:**

Other variables such as age or education for example did not show any noticable relationship with is promoted.

**Slide 10:**

To further confirm this relation, we analysed the cramer’s V values, and found that there are 3 variables – namely previous year rating, awards won and region which have a moderate association with being promoted. Hence we will use these categorical variables to train our model.

**Slide 11:**

After observing the box plots, among continuous variables, **only avg\_training\_score** had a visible relationship with is\_promoted. Employees with higher training scores were more likely to be promoted.

Other features like age or number of training didn’t show any clear patterns in the box plots or violin plots as seen in our notebook.

**Slide 12:**

This is also visible from the value of the correlation coefficient here. Only avg training score has decent to moderate correlation with being promoted and we will use this continuous variable to train the model.

After conducting EDA on all 3 departments, we found that the key variables are the same across all 3 departments.

**Slide 13:**

This leads us to the refined problem definition: How do factors like previous\_year\_rating, awards\_won?, region, and avg\_training\_score impact promotions across departments like Sales, Operations, and Tech?

**Slide 14:**

Before modelling, we cleaned our data thoroughly. First, we used SMOTE-NC, an oversampling technique, to balance the target variable, since only about 8.5% of employees were promoted.

We dropped columns like employee\_id and age that didn’t contribute to prediction, removed rows where education was missing and replaced NA values in previous\_year\_rating with 0 — since these are likely first-year workers.

And lastly, we converted categorical variables to strings and encoded them numerically to be used in our models later.

**Slide 15**

After cleaning the data, we move on to machine learning. We used a total of 3 models - decision tree, random forest, and cat boost - to perform our machine learning algorithms. Here, we will be focusing on the sales and marketing department.

**Slide 16**

In addition, we have also included a technique called hyperparameter tuning, which is the process of finding the optimal combination of hyperparameters that control the behavior of the algorithm to improve model performance. The results can be found in the notebook.

**Slide 17**

The first type of model is the decision tree. Looking at the train and test data set, we can see that the classification accuracy for both are quite high with a discrepancy of about 0.116. Every time we ran the notebook, we observed that the classification accuracy for both the train and test data set ranged from approximately 0.8-0.95.

**Slide 18**

Next we implemented the random forest model. We found that our classification accuracy is still quite high and similar to the decision tree model and there are very minimal changes in our false positive rates for both data sets.

**Slide 19**

To test for efficiency in performing machine learning, we also tested the Cat Boost model. CatBoost can automatically handle categorical data without any encoding needed, making it much more convenient. However, the classification accuracy for both the test and train dataset decreased to about 0.76 to 0.79.

**Slide 20**

Through this project, we explored different statistical and machine learning models to understand what factors most influence our prediction target—employee promotions.

Creating and analyzing the 3 models helped us learn how each algorithm works, what their strengths and weaknesses are, and how they handle different types of data—especially when categorical variables are involved.

From our analysis, we identified key variables that impact promotion likelihood. This insight can be valuable for employees.

For example, they can better understand which parts of their resume or profile they should strengthen—whether it’s their performance ratings, certifications, or years of experience.

They can also use these insights to adjust their performance or focus areas, ultimately increasing their chances of being promoted within the company.

**Slide 21**

If we look at the confusion matrix from the Decision Tree on the test data, we see a high number of true negatives and false positives — 3119 out of a total of about 3368 non-promoted predictions, which is roughly 91.3%

Digging deeper, we found that within our test set, for employees with a previous year rating of 0, about 91.8% of them were not promoted.

These percentages are very close, which tells us that the model likely relied heavily on the previous\_year\_rating variable when making predictions. In other words, that feature provided a very clear, straightforward signal about whether someone would be promoted or not.

**Slide 22**

Comparing the results from the 3 machine learning models, we found that decision trees and random forests provide similar statistics where they have a higher classification accuracy (generally over 90%) as compared to Catboost (average of 79%).

**Slide 23**

We are in favour of using Decision Trees due to its lower time taken for fitting the model and hyperparameter tuning, making it more ideal especially if one wishes to create separate models for each department. However, although Catboost has a lower classification accuracy and F1-score, it has an advantage of efficiency in that it does not require any encoding beforehand.

It can also be seen that for a vast majority of the Decision Tree and Random Forest models, hyperparameter tuning was able to improve metrics or at least maintain similar metrics such as classification accuracy and F1-Score.

**Slide 24**

As shown by the EDA, the general predictors seem to be the same across departments, and thus it may be possible to create these machine learning models for all departments.

Metrics for the Sales & Marketing models seem to be the best out of the other 3 departments, likely due to the larger number of training data for the department.

Therefore, since the predictors are the same across all departments, aggregating all the departments into 1 dataset may provide better accuracy and prediction.

**Slide 25**

Thank you for listening to our presentation!