Employee Turnover Analysis

GROUP - 8
JAHNAVI ANGATI
VICTORIA LIU
ANISHA SINGH
HAYOUNG KIM
ALEX KIM

Description of Project Goals:

Employee turnover, also known as employee churn, refers to the rate at which employees leave an organization and need to be replaced. High employee turnover can be costly for businesses, leading to increased recruitment expenses, loss of institutional knowledge, decreased productivity, and reduced employee morale. Our primary objective is to conduct an in-depth analysis of employee turnover, identify factors influencing it, and develop strategies for improved retention. We explored a dataset from a large US company, utilizing ***Logistic Regression***, ***KNN Classifier***, and ***Random Forest*** models to predict turnover and determine the best models.

Importance of the Problem:

Employee turnover poses significant costs and challenges for every business as follows:

Productivity and Efficiency: High turnover disrupts team dynamics and workflows, leading to decreased productivity. It creates instability and stress among the remaining workforce.
Loss of Knowledge and Expertise: Departing employees take valuable knowledge and skills. This loss can hinder the organization's ability to innovate, solve problems, and deliver high-quality products or services.
Team Cohesion: Consistent turnover disrupts team cohesion, impacting effective collaboration. It may take time for new employees to integrate fully into the team and work seamlessly with their colleagues.
Reputation and Employer Branding: High turnover harms a company's reputation and makes it difficult to attract top talent.

Exploratory Data Analysis:

- *No. of Observations:* **9540**
- *Numerical Variables:* **5** (review, projects, tenure, satisfaction, average working hours)
- *Categorical Variables:* **5** (department, bonus, promoted, salary, left)

The target variable *'Left'*, which represents whether people leave the company or not, has an overall leave rate of around **30%**. In this context, the baseline accuracy for predictive models we developed later for this binary classification problem would be **70%**.

By examining the data structure, we found that there are no missing values in the data set. Based on the histogram and descriptive statistics (Tables 1 and 2), the variables in our dataset appear to have a

distribution that approximates normality. This observation is reassuring as it indicates that our data is not heavily skewed and does not exhibit significant outlier problems. By checking the correlation matrix (table 3), ‘review’ has the highest correlation with ‘left’. Employees who received higher review scores tend to have a higher churn rate (table 3). We may conclude that exceptional employees usually don't stay loyal to one company. However, employees' satisfaction score doesn’t show a trend that 'employees with lower satisfaction would highly likely to quit'. Even people with a relatively high satisfaction score have a higher intention to quit (table 3). In terms of average working hours, we found out that people who have relatively low working hours and relatively high working hours are more likely to quit (table 6). We may presume from the result that 'people who are thinking of quitting would put a minimum amount of time on working' or 'overworking is one of the obvious reasons for employees to quit'.

To dive into categorical variables, the dataset includes 10 departments and 3 salary levels within 70% of people in the median type. 80% of employees didn’t receive a bonus and only 3% of employees got promoted. By checking the influence of different levels of categorical variables on churn rate, the churn rate doesn’t show a variation in employees from different departments, salary types and whether they got a bonus (table 4). However, we found out that employees who do not get promoted have a higher churn rate. (table 4). We could presume that promotion is a significant consideration for employee turnover.

Solution and Insights:

1. Logistic Regression:

The first model that we have chosen to fit our dataset is the logistic regression model. Since our output i.e., left column or target variable indicates if an employee has left the organization or not based on various features has two possible classes i.e., yes or no, we decided to go with this model.

We performed K-fold cross-validation on the logistic regression (Table 5) to assess the model’s performance and avoid any potential issues of overfitting. We have calculated and plotted the graph between mean accuracy and the different number of folds(K). The best accuracy that we got was 73.46 at K=2.

From the Classification Report Table (Table 6), we got a test accuracy of about 73.92 and a precision of about 70. However, the model did poorly on recall with a value of just 22.5.

From the confusion matrix of our model (Table 7) where we got a test accuracy of about 73.9. We can see that the training accuracy of 73.6 is very close to the test accuracy. Therefore, we have not underfitted or overfitted our model.

We also plotted the feature importance graph of our logistic regression model (Table 8). From the graph we saw that the top four

features which have higher importance are performance reviews,average hours per month, tenure and satisfaction.

Before running the model we thought that salary would be an important feature in making predictions however after plotting the graph we saw that is definitely not the case.

2. KNN Classifier:

We chose KNN Classifier for the second model. To improve the performance of KNN, we excluded the department variable which has a lack of discriminative values for the target variable. Before fitting the data, we scaled it using StandardScaler() for distance-based modeling. Using GridSearchCV, we found the best k value to be 15 (table 9).

As a result, k=15 achieved 86% of train accuracy and 84% of test accuracy, a 15% improvement from the baseline accuracy. However, the recall rate was low at 65% possibly due to the imbalanced data. SMOTE helped increase recall to 83%, but overfitting occurred (table 10).

Using permutation importance, we found that average monthly working hours and satisfaction scores were the top 2 features for KNN (table 11). However, plotting the KNN graph with these two features showed 4 colors instead of the expected 2 (table 12). Even employing 2D PCA using all predictors didn't show significant improvement (table 13), leading us to believe that the top two features, average monthly working hours, and satisfaction score, worked relatively well for KNN Classifier.

We concluded that predicting with many features is definitely not suitable for KNN. Even so, KNN might have performed better with a dataset containing more discriminative values based on different classes.

3. Random Forest:

For this final model, we tried a number of trees ranging from 100 to 1000, and a max tree depth ranging from 10 to 50. The k-fold cross-validation accuracy of every unique combination of these parameters was then used to determine the optimal set to train our final model on. As shown in Table 14, all of the combinations with a max tree depth of 10 had the lowest accuracy, holding the number of trees constant (e.g. iterations 0, 5, 10, etc.). Looking at how the number of trees affected accuracy (in batches of 5 iterations, with iterations 0-4 all having n=100), we observed a general positive trend up until n=1000, where accuracy seemed to plateau. This culminated in iteration 18, with parameters of 750 trees at a max depth of 30, having the highest overall accuracy.

Once the optimal parameters were identified in the previous step, we next fit our final model to these parameters. Running this model against the test data resulted in an accuracy of 87.59%, precision of 80.86%, and recall of 73.13%. As seen from these results as well as in the confusion matrix (Table 15), precision was a bit higher than recall. Potentially, if the business were more concerned with capturing every true positive even if that were to result in more false positives, we could lower the cutoff threshold from 0.5. This would result in higher recall and fewer

false negatives (people not expected to leave who do), but more true and false positives.

4. Variable Importance:

The variable importance plot for the Random Forest model (Table 16) had 'satisfaction' as the most significant, followed closely by 'review' and 'avg_hrs_month', and lastly 'tenure' to round out the top four. This slightly differed from the permutation importance that was observed with the KNN model - the same top four variables in but in a different order.

Suggestions:

Based on our comprehensive analysis, we propose the following suggestions for the top 4 variables we have identified.

Satisfaction: The highest churn rate was observed not only in employees with low satisfaction scores (30-50 out of 100) but also in the group with high satisfaction scores (70-90 out of 100). Assuming high satisfaction scores indicate loyalty would be naive, as some employees may not have provided honest responses in the survey. A more comprehensive understanding may require exploring various aspects beyond direct survey inquiries, including everyday observations.

Review: The unexpected positive correlation between turnover and review scores, along with the high churn rate among outstanding employees with high ratings (90 or above), calls for a focus on providing appropriate promotions and bonuses to retain top talent. For underperforming employees, organizational-level education and management are crucial, while high-performing employees may benefit from exclusive rewards and enhanced opportunities for motivation.

Average Working Hours of Month: The unexpected high turnover in the group with the least monthly average working hours suggests considering industry standards, as it may indicate potential overworking compared to norms. Alternatively, the second-highest turnover in the group with the most working hours suggests that working hours indeed have a significant impact on employee retention. Therefore, exploring company-wide measures to promote a culture of efficiency and time management, preventing burnout, and improving overall well-being might be essential to address this issue.

Tenure: There were significant turnover rates observed in the 2-3 years and 7-8 years tenure groups, which could be related to promotions and active job-seeking periods. Since tenure is closely tied to individual career building, it may seem somewhat out of the company's control. Nevertheless, discerning some patterns, developing retention strategies for long-tenured employees, such as offering growth opportunities, mentorships, and fostering a positive work environment, would be essential.

References:

Table 1:

```
*checking missing values*
department       0
promoted         0
review           0
projects         0
salary           0
tenure           0
satisfaction     0
bonus            0
avg_hrs_month    0
left             0
dtype: int64
*checking distribution*
                count        mean       std        min         15%  \
promoted       9540.0    0.030294  0.171403    0.00000    0.000000
review         9540.0    0.651826  0.085307    0.31000    0.565130
projects       9540.0    3.274843  0.579136    2.00000    3.000000
tenure         9540.0    6.556184  1.415432    2.00000    5.000000
satisfaction   9540.0    0.504645  0.158555    0.00000    0.335051
bonus          9540.0    0.212055  0.408785    0.00000    0.000000
avg_hrs_month  9540.0  184.661571  4.144831  171.37406  180.172303

                     25%         35%         50%         65%         75%  \
promoted        0.000000    0.000000    0.000000    0.000000    0.000000
review          0.592884    0.616170    0.647456    0.681370    0.708379
projects        3.000000    3.000000    3.000000    3.000000    4.000000
tenure          5.000000    6.000000    7.000000    7.000000    8.000000
satisfaction    0.386801    0.433915    0.500786    0.570781    0.622607
bonus           0.000000    0.000000    0.000000    0.000000    0.000000
avg_hrs_month  181.472085  182.634610  184.628796  186.776679  187.728708

                     90%         max
promoted        0.000000    1.000000
review          0.765393    1.000000
projects        4.000000    5.000000
tenure          8.000000   12.000000
satisfaction    0.714491    1.000000
bonus           1.000000    1.000000
avg_hrs_month  189.772293  200.861656
```

Table 2:

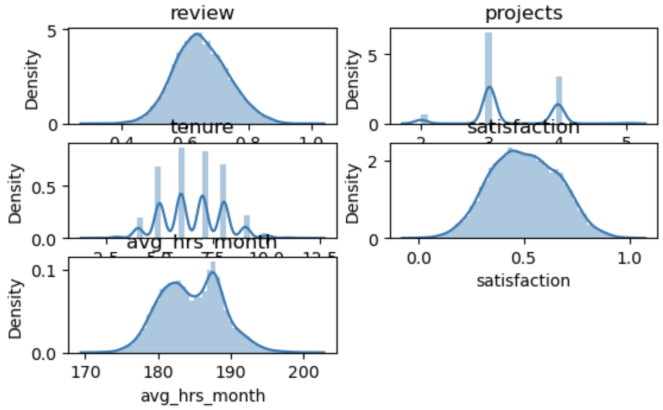
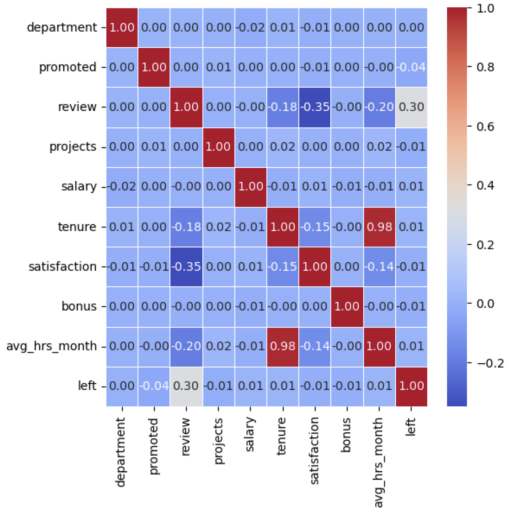


Table 3:



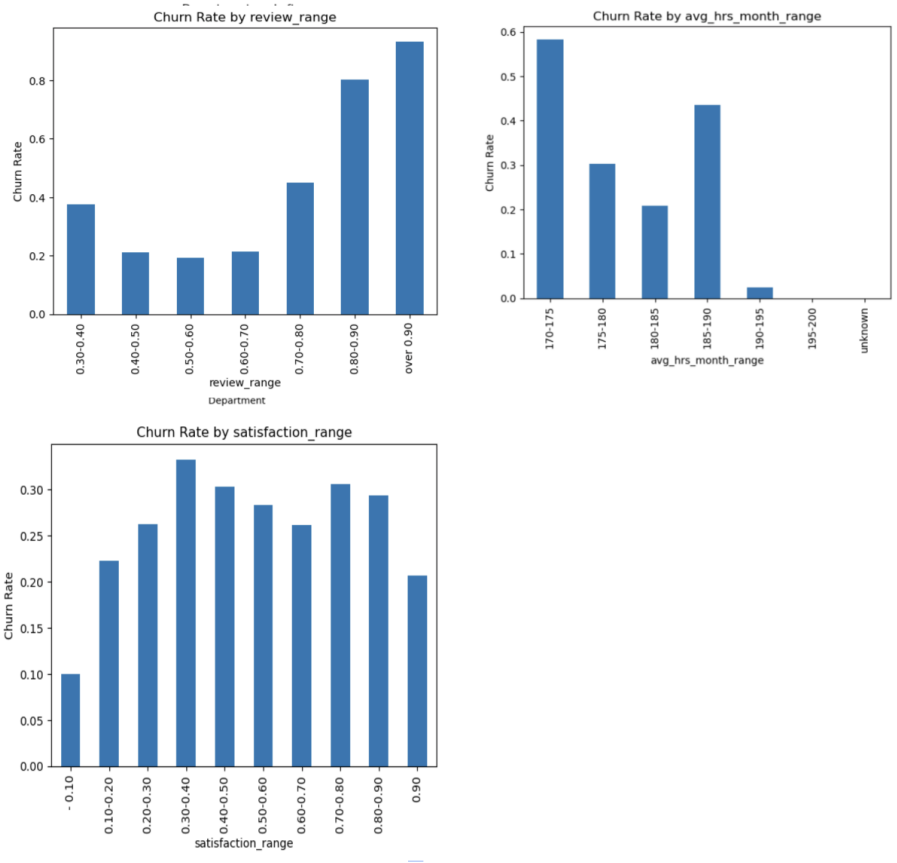

Churn Rate by review_range
Churn Rate
review_range
0.30-0.40
0.40-0.50
0.50-0.60
0.60-0.70
0.70-0.80
0.80-0.90
over 0.90
Churn Rate by avg_hrs_month_range
Churn Rate
avg_hrs_month_range
170-175
175-180
180-185
185-190
190-195
195-200
unknown
Churn Rate by satisfaction_range
Churn Rate
satisfaction_range
- 0.10
0.10-0.20
0.20-0.30
0.30-0.40
0.40-0.50
0.50-0.60
0.60-0.70
0.70-0.80
0.80-0.90
0.90

Table4:

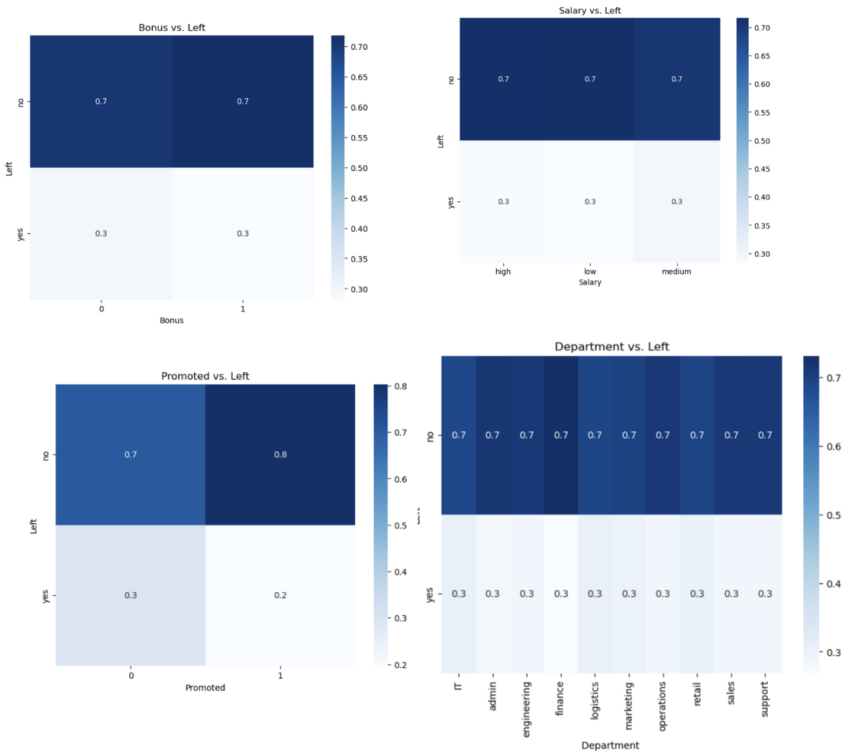

Bonus vs. Left
Left
Bonus
Salary vs. Left
Left
Salary
high
low
medium
Promoted vs. Left
Left
Promoted
Department vs. Left
Department
IT
admin
engineering
finance
logistics
marketing
operations
retail
sales
support

Table 5:

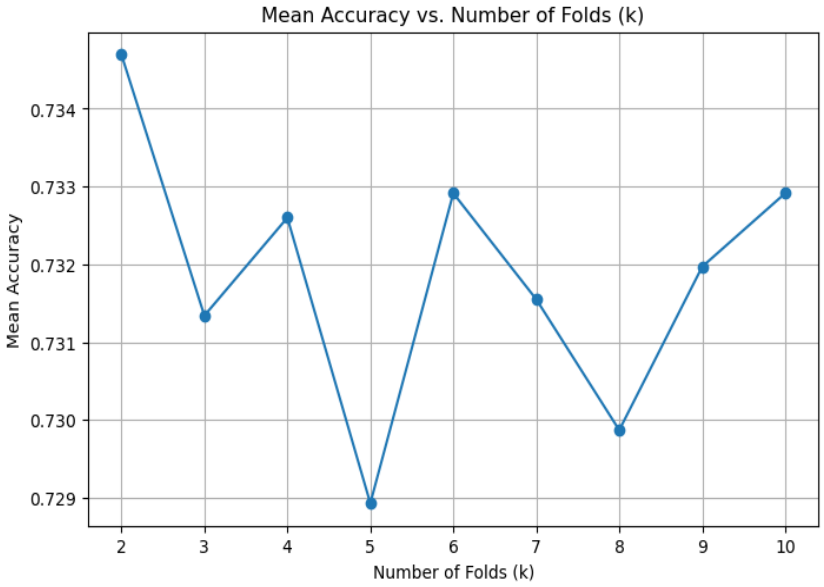


Table 6:

Table 7:

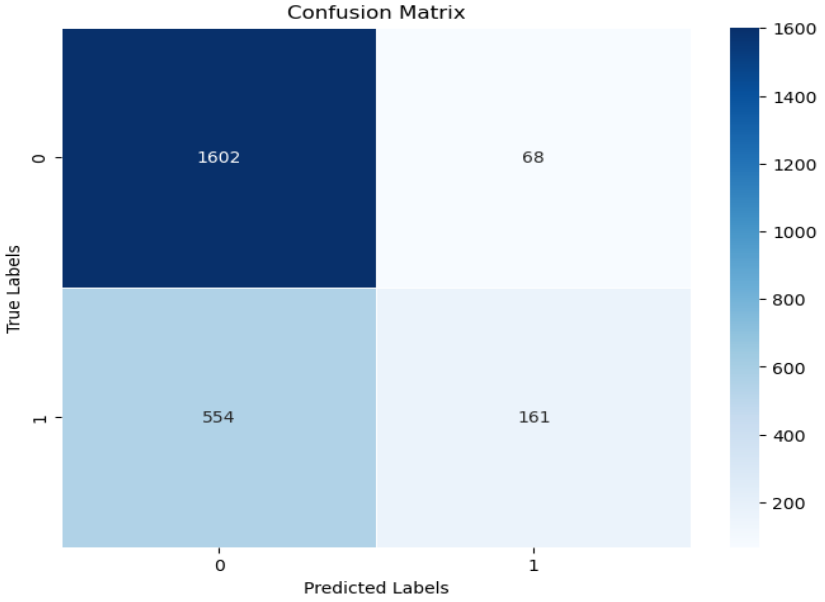


Table 8:

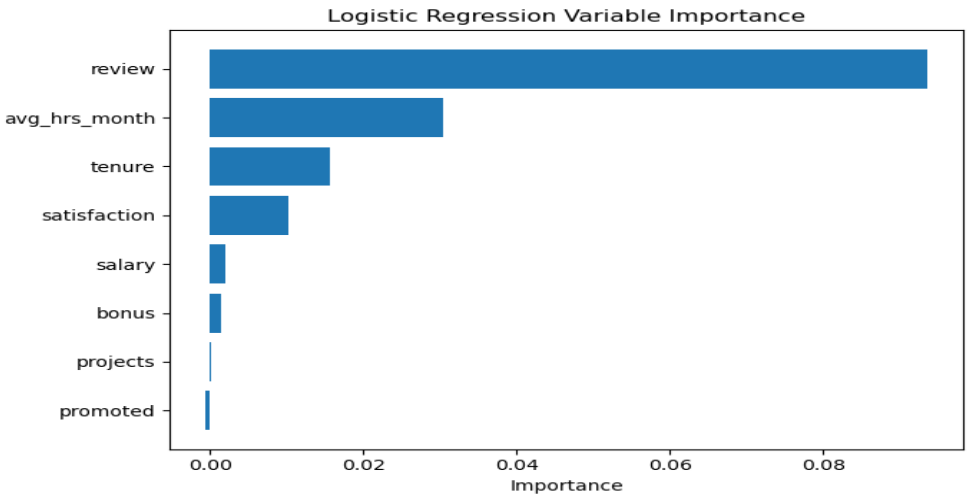
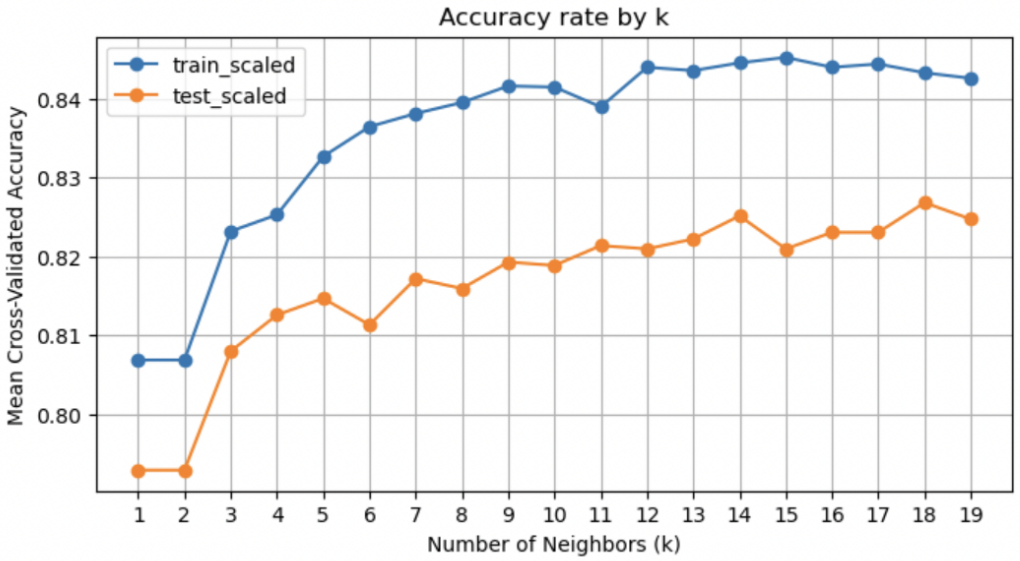


Table 9:




```
Train_scaled has best k = 15 while test_scaled's best k = 18
Overall accuracy rate is higher when k = 15
```

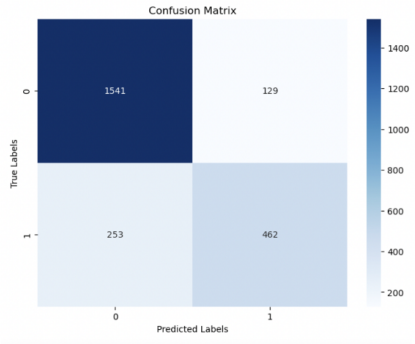
Table 10:

KNN Classifier

Results

** Test data*

Accuracy	Recall	Precision
84 (train: 86)	65	78



Balancing data (SMOTE)

Overfitting!!

** Test data*

Accuracy	Recall	Precision
100 (train: 79)	83	61

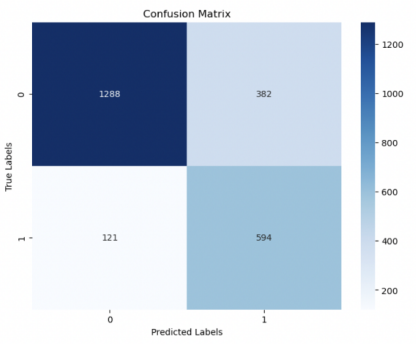


Table 11:

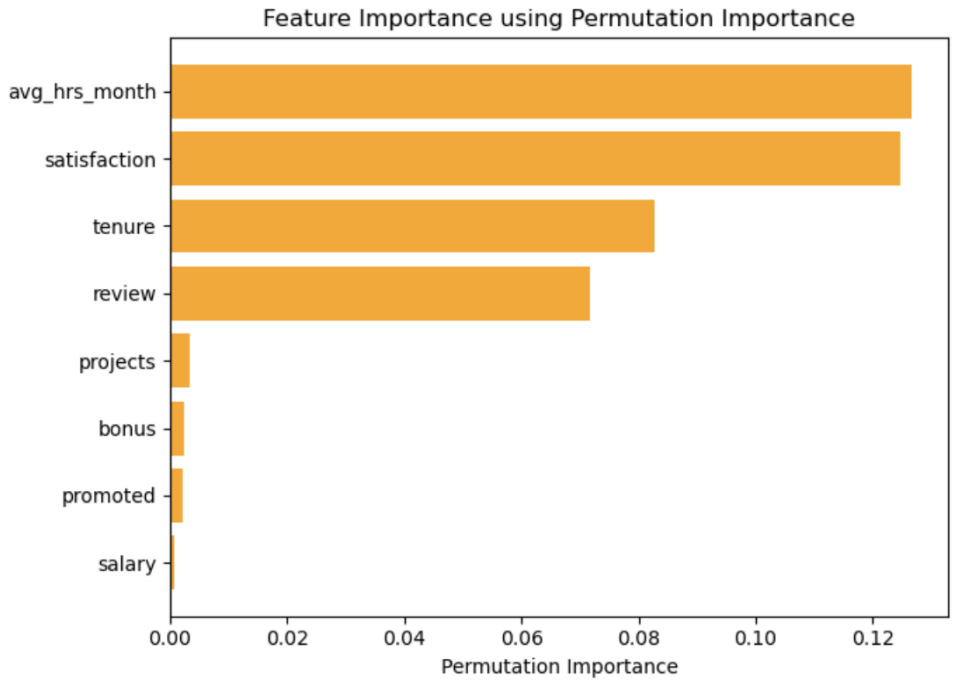


Table 12:

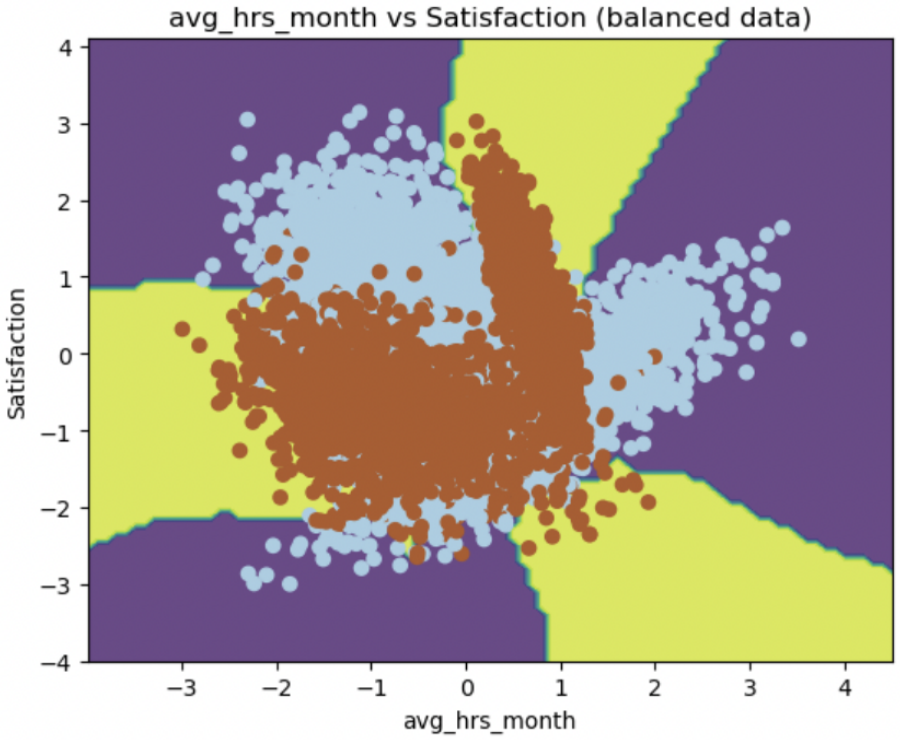


Table 13:

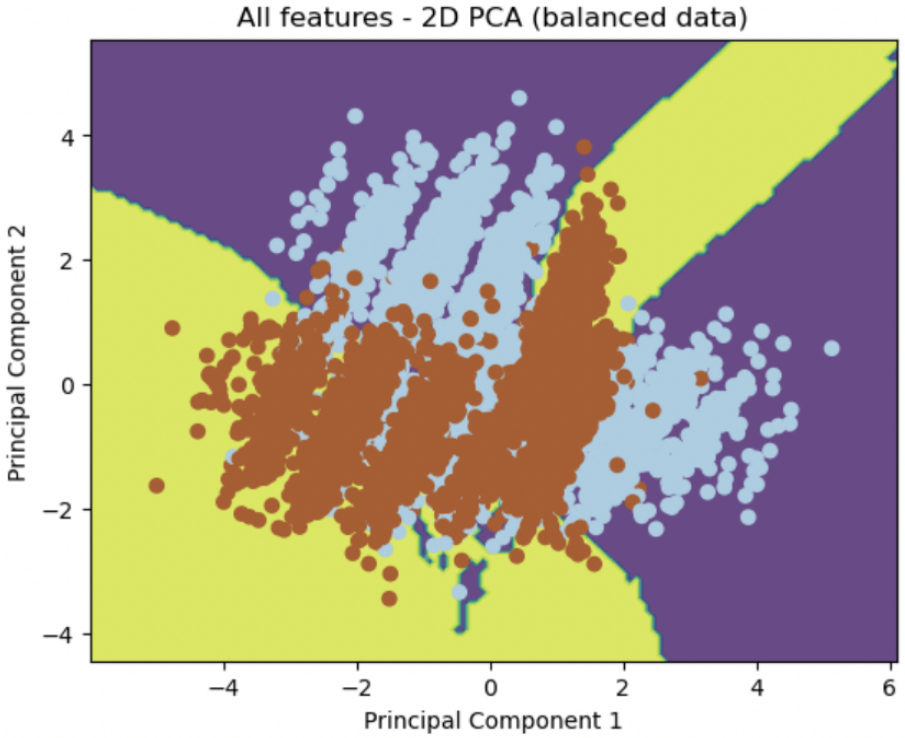


Table 14:

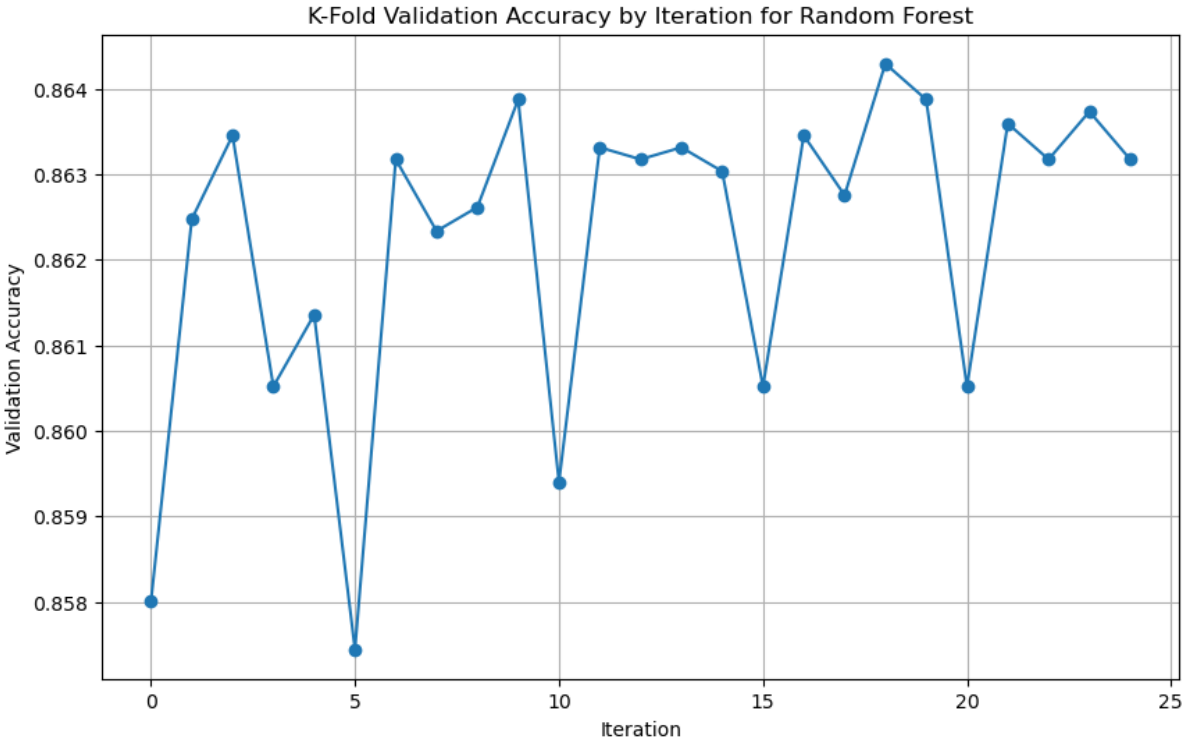


Table 15:

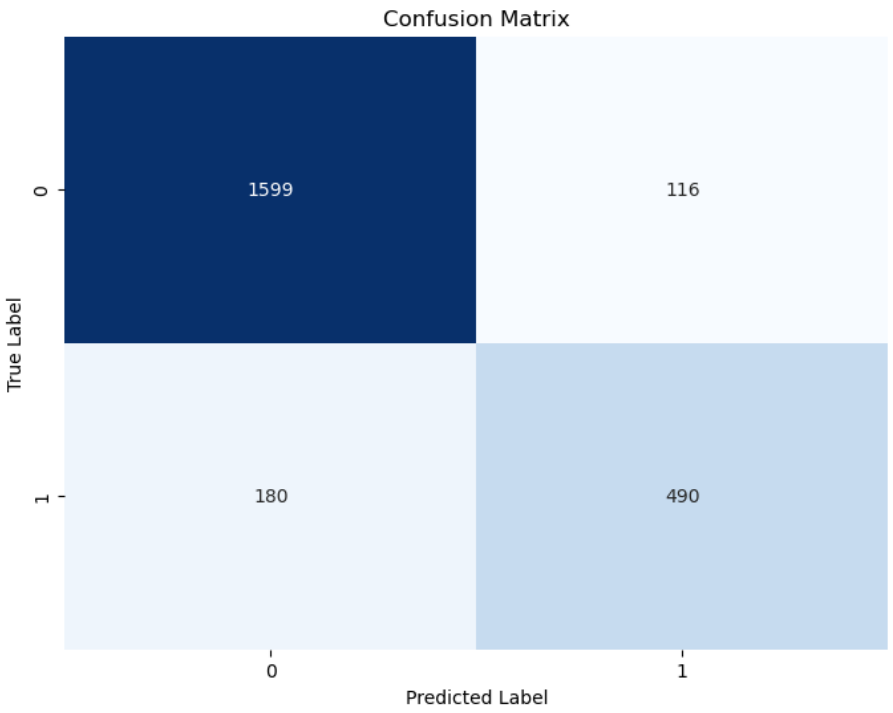
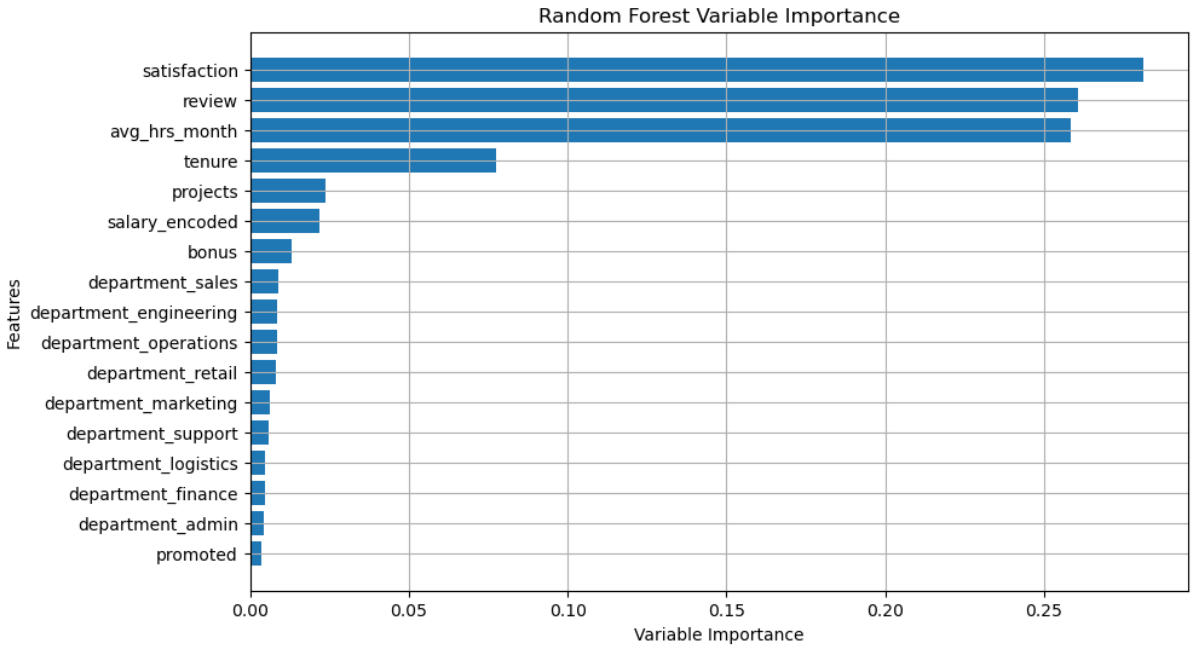

Confusion Matrix
1599
116
180
490
True Label
0
1
0
1
Predicted Label

Table 16:


Random Forest Variable Importance
satisfaction
review
avg_hrs_month
tenure
projects
salary_encoded
bonus
department_sales
department_engineering
department_operations
department_retail
department_marketing
department_support
department_logistics
department_finance
department_admin
promoted
Features
0.00
0.05
0.10
0.15
0.20
0.25
Variable Importance