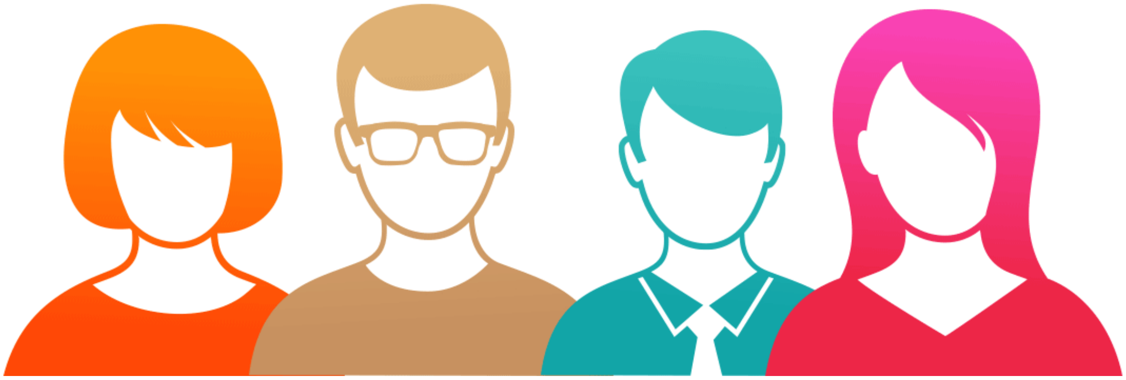
Human Resources Analytics

Predicting Employee Turnover



Business Problem:

A multi-million dollar company is about to go bankrupt and he wants to know why his employees are leaving.

Objective:

- To understand what factors contributed most to employee turnover
- To create a model that predicts the likelihood if a certain employee will leave the company or not
- To create or improve different retention strategies on targeted employees

Approach:

- Obtaining the data is the first approach in solving the problem.
- Scrubbing or cleaning the data is the next step. This includes data imputation of missing or invalid data and fixing column names.
- Exploring the data will follow right after and allow further insight of what our dataset contains. Looking for any outliers or weird data. Understanding the relationship each explanatory variable has with the response variable resides here and we can do this with a correlation matrix.
- Modeling the data will give us our predictive power on whether an employee will leave.
- Interpreting the data is last. With all the results and analysis of the data, what conclusion is made? What factors contributed most to employee turnover? What relationship of variables were found?

The Dataset:

- **Satisfaction:** An employee's level of satisfaction in percentage
- **Evaluation:** An employee's evaluation score in percentage
- **Project Count:** The number of projects the employee has done
- **Average Monthly Hours:** The total monthly hours an employee worked
- **Years At Company:** The number of years an employee was at the company
- **Work Accident:** Whether an employee had an accident or not. Where 0 (zero) means no and 1 (one) means yes
- **Promotion:** Whether an employee had a promotion within the last five years. Where 0 (zero) means no and 1 (one) means yes
- **Department:** The type of department an employee worked under. Which includes sales, accounting, hr, technical, support, management, IT, product management, and marketing.
- **Salary:** The type of salary an employee got, which ranges from low, medium, or high.

Note: *The data was found from the "Human Resources Analytics" dataset provided by Kaggle's website. https://www.kaggle.com/ludobenistant/hr-analytics*

Few observations:

One of the most common problems at work is turnover. Replacing a worker earning about $50,000 cost the company about $10,000 or 20% of that worker’s yearly income according to the Center of American Progress. Replacing a high-level employee can cost multiple of that.

- Cost of off-boarding
- Cost of hiring (advertising, interviewing, hiring)
- Cost of onboarding a new person (training, management time)
- Lost productivity (a new person may take 1-2 years to reach the productivity of an existing person) Source: (https://cnmsocal.org/featured/true-cost-of-employee -turnover/)

Solution:

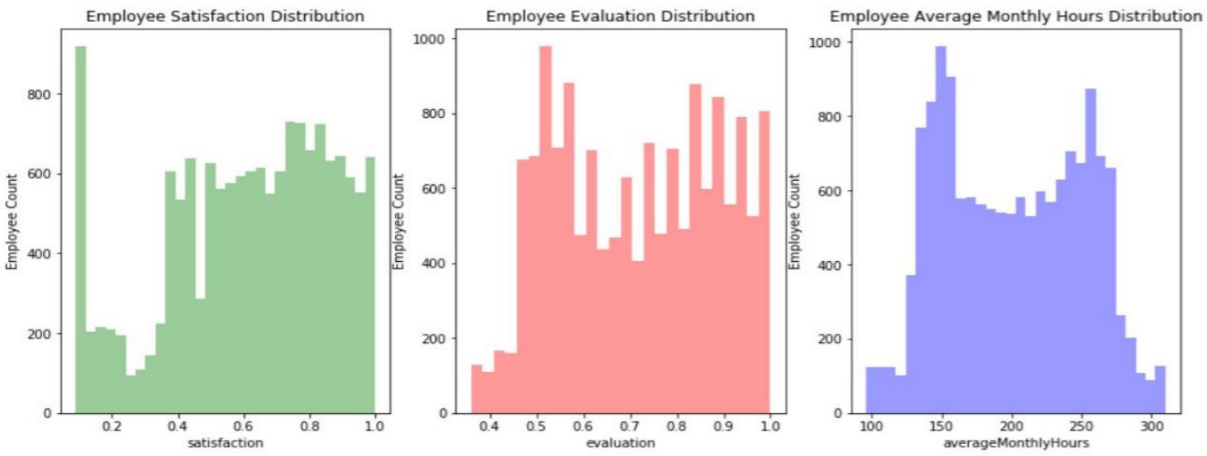
The goal is to create a retention plan! We can help identify who is in need of more support to prevent potential turnover. This model will predict and calculate the likelihood of each employee sticking around in the company.

Exploring the data:

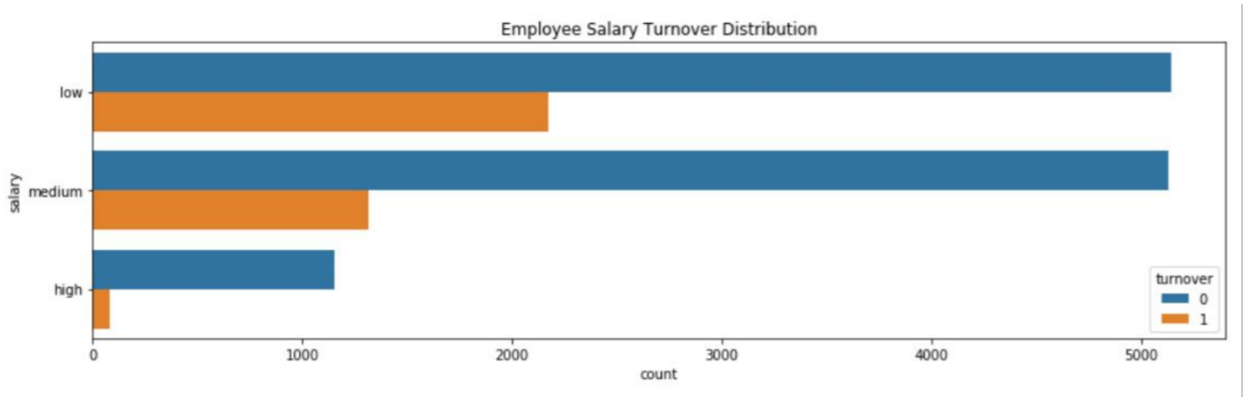
The dataset has-

- About 15,000 employee observations and 10 features
- The company had a turnover rate of about 24%
- Mean satisfaction of employees is 0.61

Satisfaction & Evaluation & Hours Distribution-

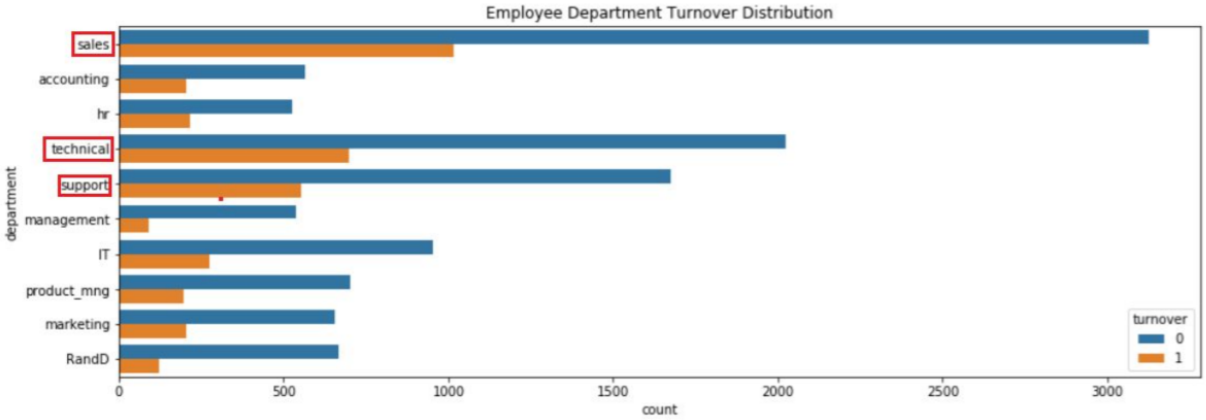


Employee Salary Distribution–



Department Distribution–

- The sales, technical, and support department were the top 3 departments to have employee turnover
- The management department had the smallest amount of turnover

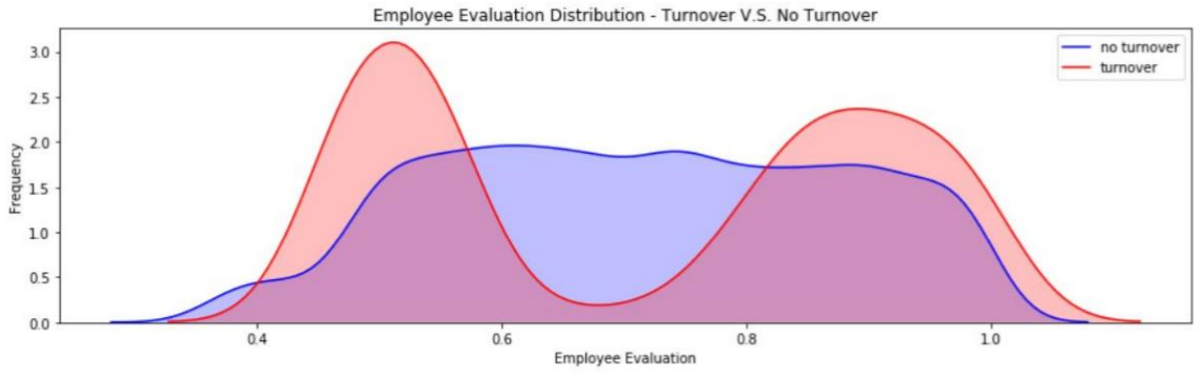


Project Count Distribution-

- More than half of the employees with 2,6, and 7 projects left the company
- All of the employees with 7 projects left the company

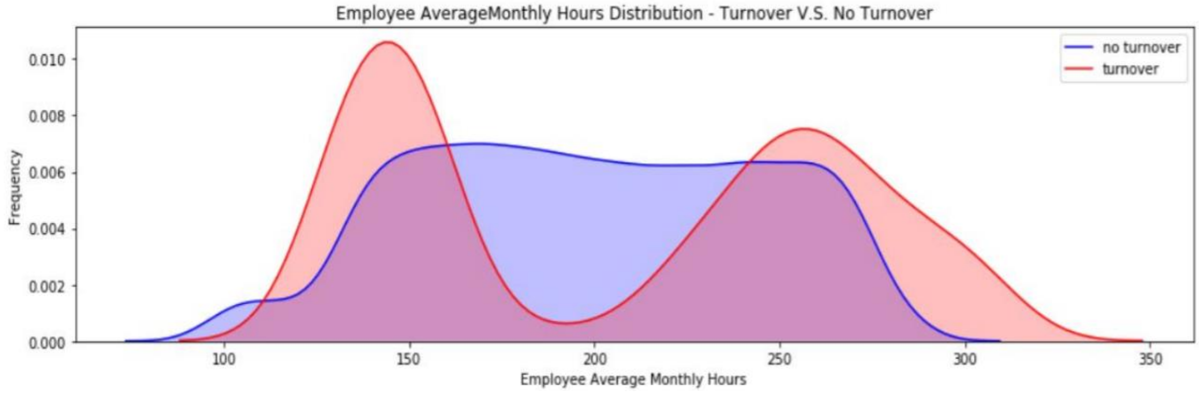
Evaluation Distribution-

- There is a bi-modal distribution for those that had a turnover.
- Employees with low performance tend to leave the company more (0.4~0.6)
- Employees with high performance tend to leave the company more (0.8-1)
- The sweet spot for employees that stayed is within 0.6-0.8 evaluation



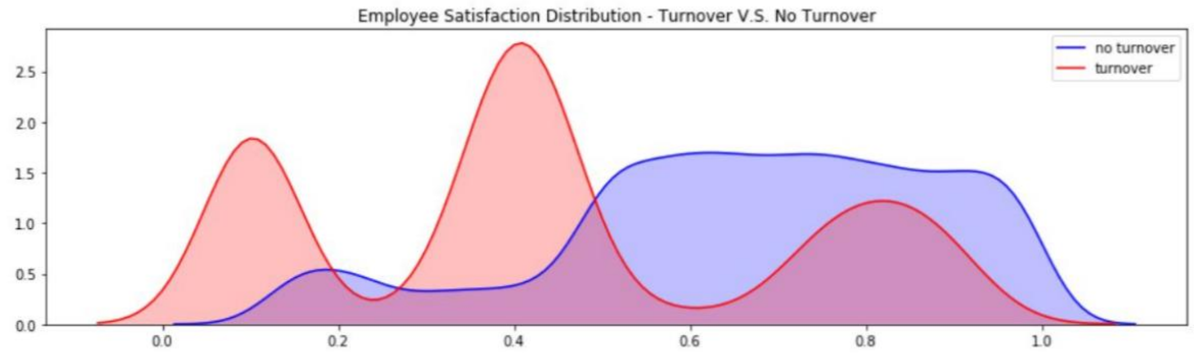
Average Monthly Hours Distribution-

- Employees who had less hours of work (~150hours or less) left the company more
- Employees who had too many hours of work (~250 or more) left the company
- Employees who left generally were underworked or overworked



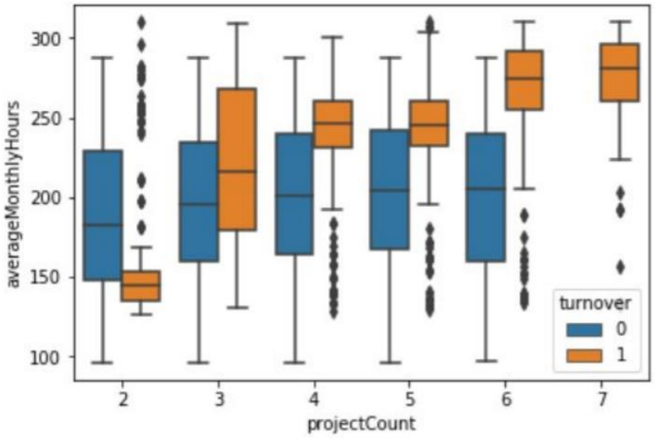
Satisfaction Distribution-

- There is a tri-modal distribution for employees that turnover
- Employees who had really low satisfaction levels (0.2 or less) left the company more
- Employees who had low satisfaction levels (0.3~0.5) left the company more
- Employees who had really high satisfaction levels (0.7 or more) left the company more



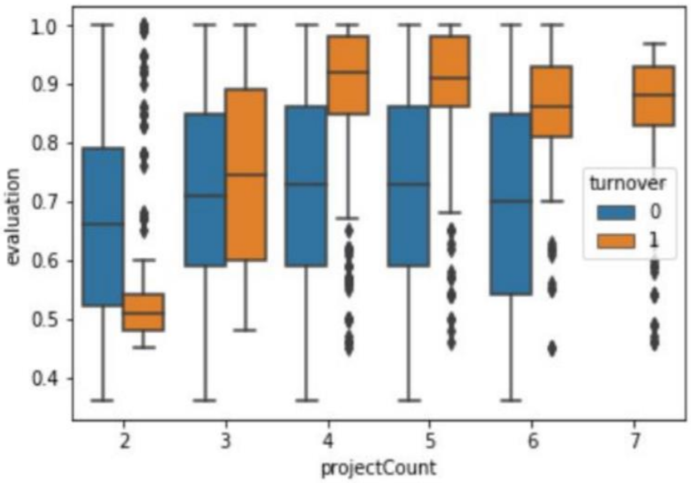
Monthly Hours VS Project Count-

- Employees who had No-Turnover had an even distribution of average monthly hours as the project count increased
- Employees who had Turnover had an INCREASE in average monthly hours as the project count increased



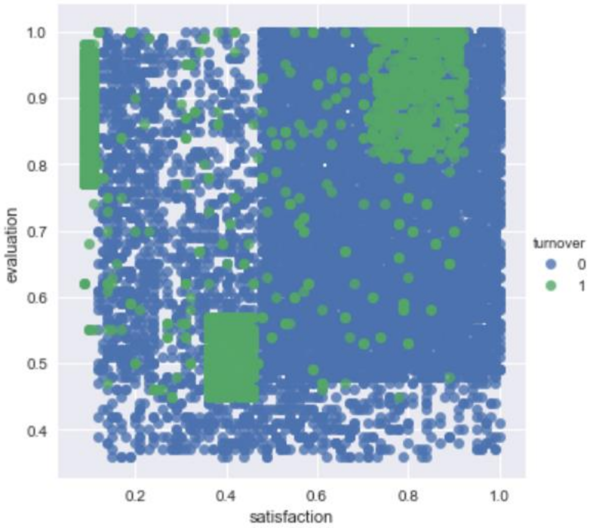
Evaluation VS Project Count-

- There is an INCREASE in evaluation for employees who did more projects within the turnover group
- For the non-turnover group, employees here had a consistent evaluation score despite the increase in project counts

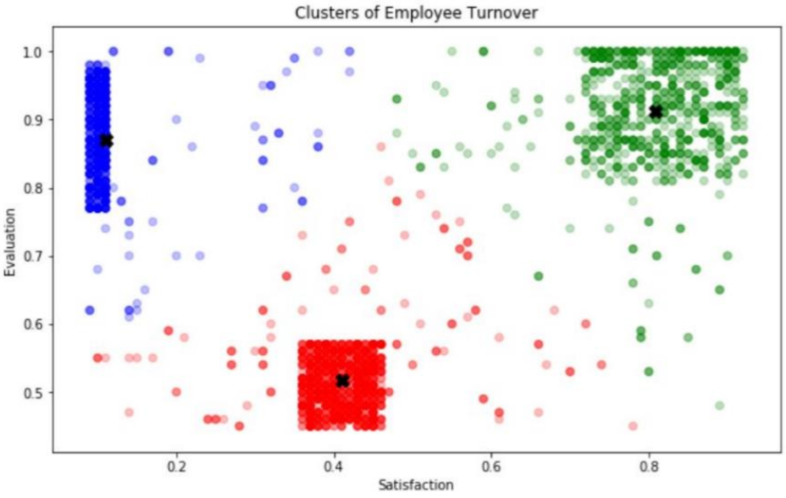


Satisfaction VS Evaluation-

- Cluster 1 (Highly Valued, But Sad) Satisfaction was below 0.2 and evaluations were greater than 0.75
- Cluster 2 (Underperforming) Satisfaction between about 0.35~0.45 and evaluations below ~0.6. This could be seen as employees who were badly evaluated and felt bad at work
- Cluster 3 (Highly Valued, But Happy) Satisfaction above 0.7 and evaluations were greater than 0.8



k-means clustering-



Blue - Overworked Employee

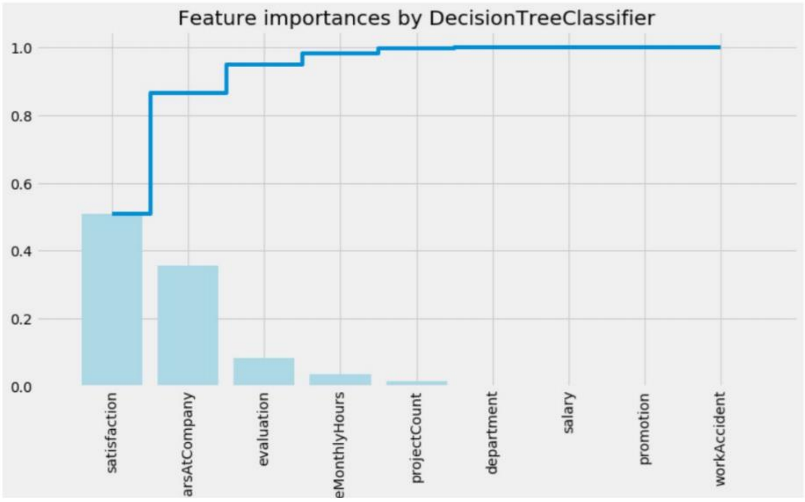
Red - Underperforming Employee

Green - Ideal Employee

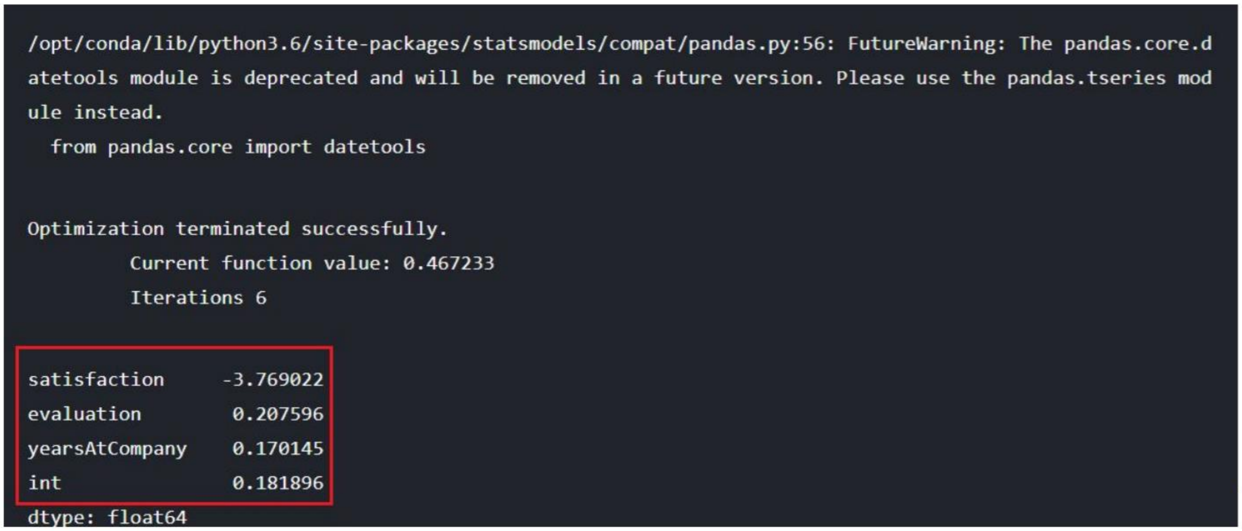
Decision Tree / Feature Importance-

Top 3 Features:

- Satisfaction
- YearsAtCompany
- Evaluation



Logistic Regression-



```
/opt/conda/lib/python3.6/site-packages/statsmodels/compat/pandas.py:56: FutureWarning: The pandas.core.d
atetools module is deprecated and will be removed in a future version. Please use the pandas.tseries mod
ule instead.
  from pandas.core import datetools


Optimization terminated successfully.
         Current function value: 0.467233
         Iterations 6

satisfaction     -3.769022
evaluation        0.207596
yearsAtCompany    0.170145
int               0.181896
dtype: float64
```

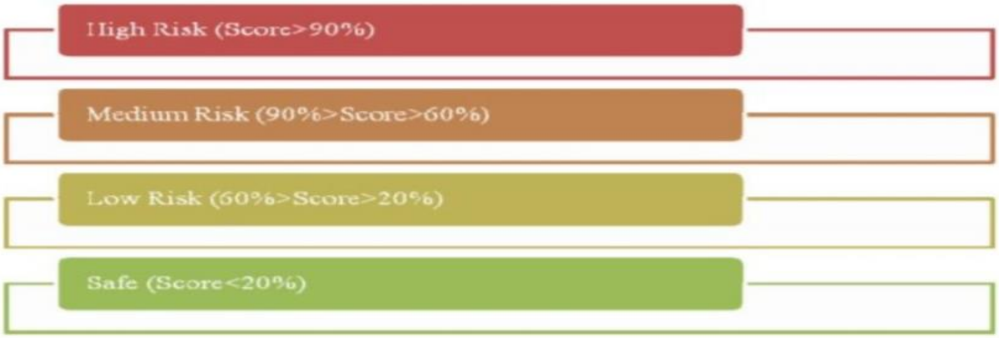
Coefficient:

- Dependent Variable: Employee Turnover Score
- Independent Variables: Satisfaction + Evaluation + YearsAtCompany
- EQUATION: Employee Turnover Score = (-3.769022) Satisfaction + (0.207596) Evaluation + (0.170145) YearsAtCompany + 0.181896
- The values above are the coefficient assigned to each independent variable. The constant 0.181896 represents the effect of all uncontrollable variables

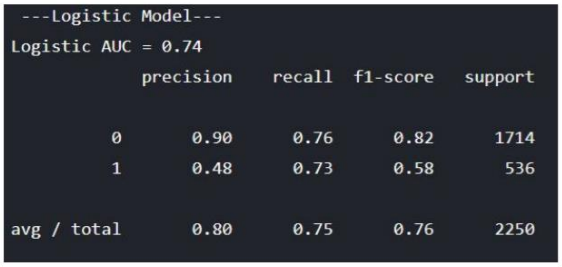
Retention Plan Using Logistic Regression –

- Safe Zone (Green) – Employees within this zone are considered safe
- Low Risk Zone (Yellow) – Employees within this zone are too be taken into consideration of potential turnover. This is more of a long-term track
- Medium Risk Zone (Orange) – Employees within this zone are at risk of turnover. Action should be taken and monitored accordingly
- High Risk Zone (Red) – Employees within this zone are considered to have the highest chance of turnover. Action should be taken immediately

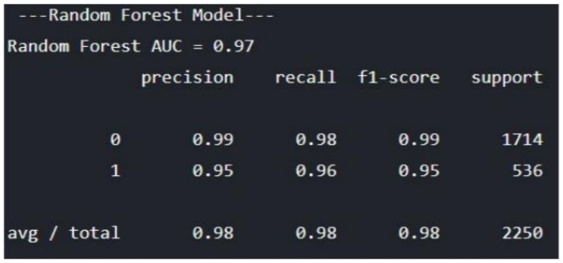
As per our data the employees with 14% turnover score will be in the safe zone.



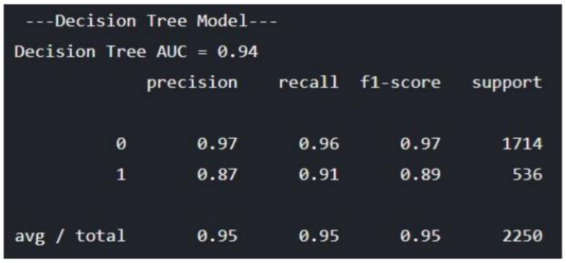
Other Model Evaluations - Confusion Matrix:



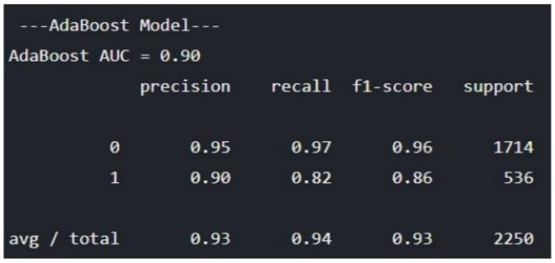
```
 ---Logistic Model---
Logistic AUC = 0.74
                 precision    recall  f1-score   support

          0       0.90      0.76      0.82      1714
          1       0.48      0.73      0.58       536

avg / total       0.80      0.75      0.76      2250
```



```
 ---Random Forest Model---
Random Forest AUC = 0.97
                 precision    recall  f1-score   support

          0       0.99      0.98      0.99      1714
          1       0.95      0.96      0.95       536

avg / total       0.98      0.98      0.98      2250
```



```
 ---Decision Tree Model---
Decision Tree AUC = 0.94
                 precision    recall  f1-score   support

          0       0.97      0.96      0.97      1714
          1       0.87      0.91      0.89       536

avg / total       0.95      0.95      0.95      2250
```



```
 ---AdaBoost Model---
AdaBoost AUC = 0.90
                 precision    recall  f1-score   support

          0       0.95      0.97      0.96      1714
          1       0.90      0.82      0.86       536

avg / total       0.93      0.94      0.93      2250
```

Interpreting the Data-

- Employees generally left when they are underworked (less than 150hr/month or 6hr/day)
- Employees generally left when they are overworked (more than 250hr/month or 10hr/day)
- Employees with either really high or low evaluations should be taken into consideration for high turnover rate
- Employees with low to medium salaries are the bulk of employee turnover
- Employees that had 2,6, or 7 project count was at risk of leaving the company
- Employee satisfaction is the highest indicator for employee turnover.
- Employee that had 4 and 5 yearsAtCompany should be taken into consideration for high turnover rate

- Employee satisfaction, yearsAtCompany, and evaluation were the three biggest factors in determining turnover

<u>Potential Solution-</u>

Binary Classification: Turnover V.S. Non-Turnover

Instance Scoring: Likelihood of employee responding to an offer/incentive to save them from leaving.

Need for Application: Save employees from leaving

In our employee retention problem, rather than simply predicting whether an employee will leave the company within a certain time frame, we would much rather have an estimate of the probability that he/she will leave the company. We would rank employees by their probability of leaving, then allocate a limited incentive budget to the highest probability instances.

Consider employee turnover domain where an employee is given treatment by Human Resources because they think the employee will leave the company within a month, but the employee actually does not. This is a false positive. This mistake could be expensive, inconvenient, and time consuming for both the Human Resources and employee ,but is a good investment for relational growth.

Compare this with the opposite error, where Human Resources does not give treatment/incentives to the employees and they do leave. This is a false negative. This type of error is more detrimental because the company lost an employee, which could lead to great setbacks and more money to rehire. Depending on these errors, different costs are weighed based on the type of employee being treated. For example, if it’s a high-salary employee then would we need a costlier form of treatment? What if it’s a low-salary employee? The cost for each error is different and should be weighed accordingly.

Solution 1:

- We can rank employees by their probability of leaving, then allocate a limited incentive budget to the highest probability instances.
- OR, we can allocate our incentive budget to the instances with the highest expected loss, for which we'll need the probability of turnover

Solution 2: Develop learning programs for managers. Then use analytics to gauge their performance and measure progress. Some advice:

- Be a good coach
- Empower the team and do not micromanage
- Express interest for team member success
- Have clear vision / strategy for team
- Help team with career development

"You don't build a business. You build people, and people build the business." - Zig Ziglar