Employee Attrition:A Classic Problem revisited via Machine Learning

PONNUR SRUJAN
GANGEYEDULA CHARAN SIMHA REDDY
SOURAV PODDAR
POONAM JOSHI

Dr.Snehanshu Saha
Batch No. - 33

March 9, 2018

Problem Statement / Definition

High Turnover Rates

- Employee turnover is a natural part of business in any industry. Excessive turnover decreases the overall efficiency of the company and comes with a high price tag.The monetary cost of such high turnover is enormous. Turnover not only affects the bottom line but also affects the company's morale.
- We are analyzing the problems within the company that are causing the employees to leave the company. We are going to predict if a particular individual is at the risk of leaving the firm or not.If yes, why is he/she leaving and take necessary actions accordingly.
- Make the cost estimates of the turnover before and after our solutions are implemented.Without understanding the negative impacts of turnover,a company may be placing itself in a position that will ultimately lead to their demise.

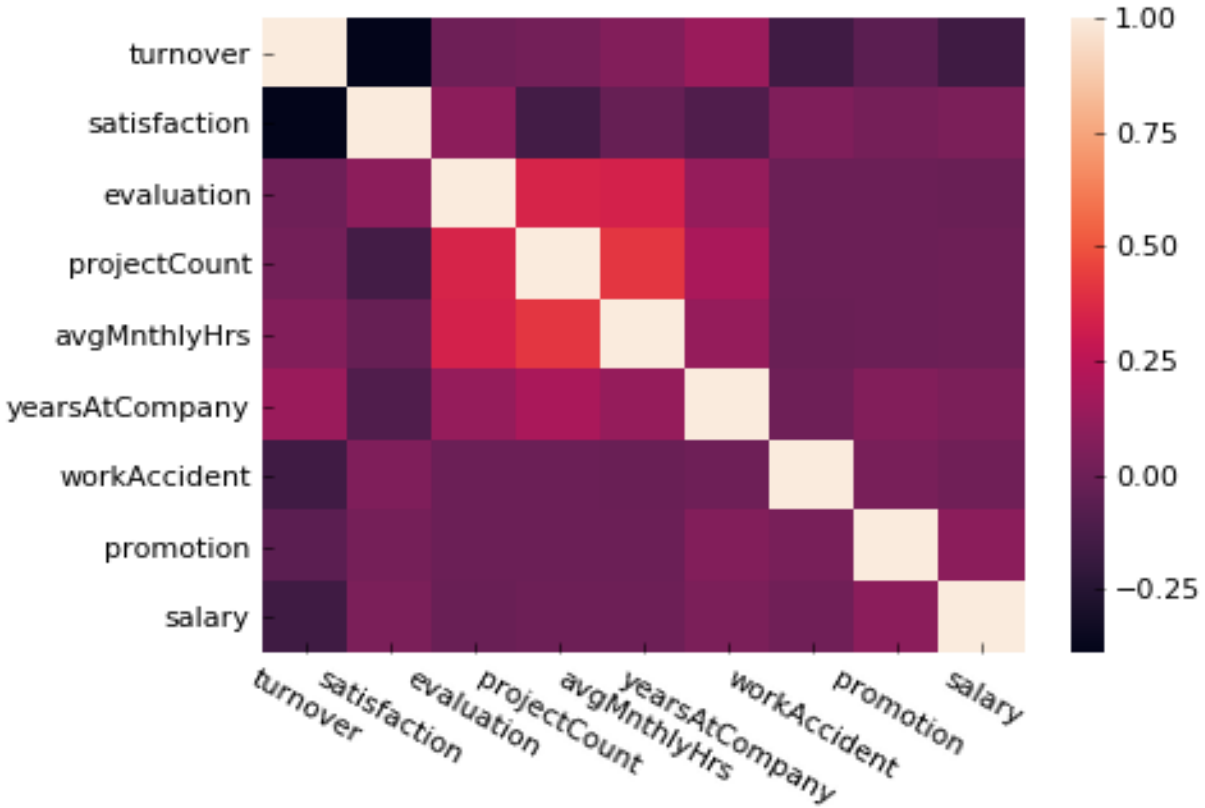
Data Exploration

Statistical Overview:

Here are some important numbers to keep in mind of the dataset:

- There is 14,999 employees and 9 independent variables
- Turnover rate: 24%

Correlation Matrix & HeatMap:



Data Exploration

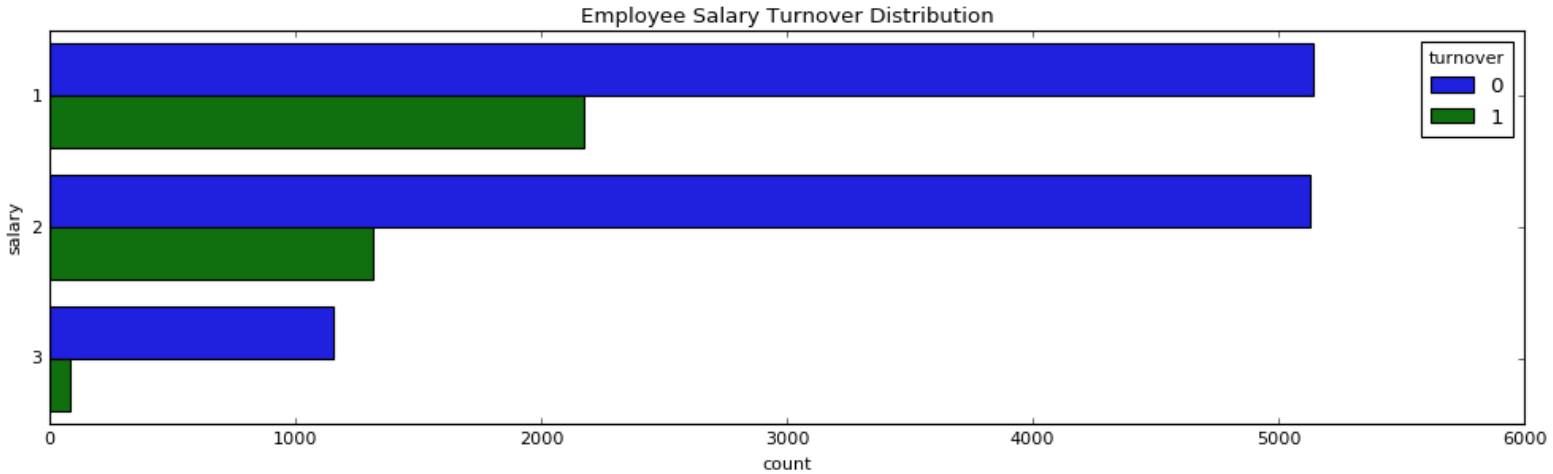
Summary:

- From the heatmap, there is a **positive**(+) correlation between the variables: **projectCount**,**averageMonthlyHours**, and **evaluation**.
- For the **negative** (-) relationships, the most important feature that correlated with our target variable (turnover) is **satisfaction**.

Salary V.S. Turnover

Summary:

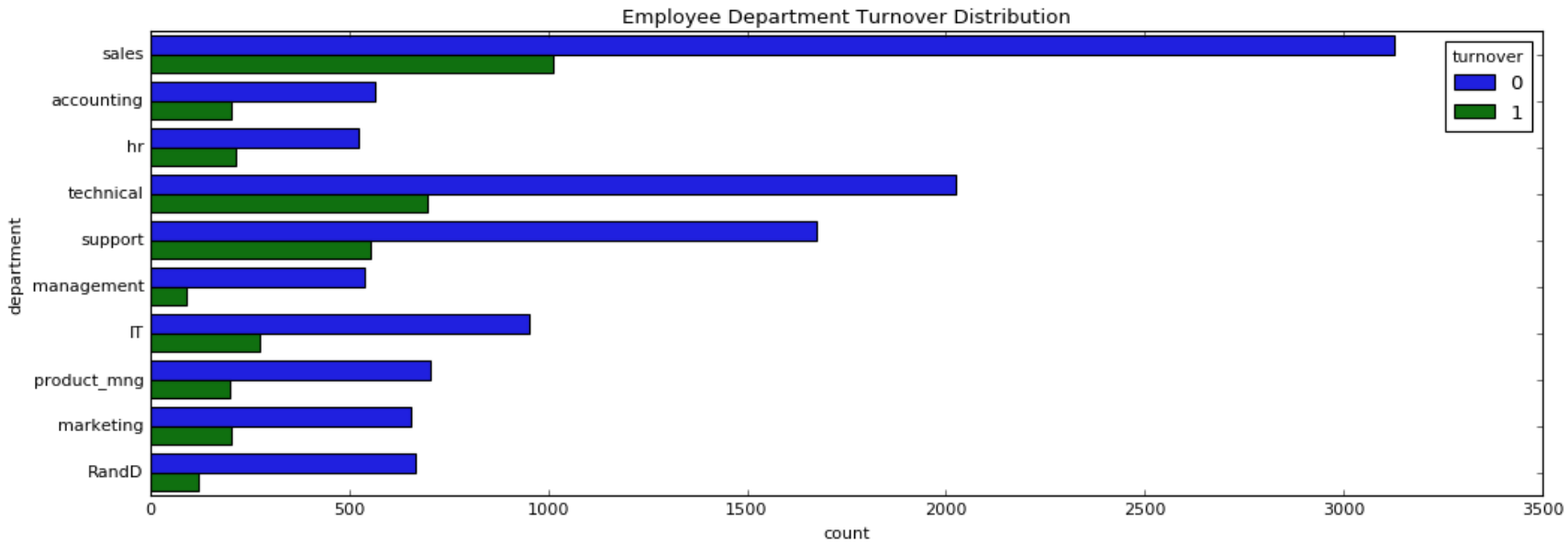
- Majority of employees who left either had low or medium salary
- Only a few employees left with high salary



Data Exploration

Department V.S. Turnover
Summary:

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

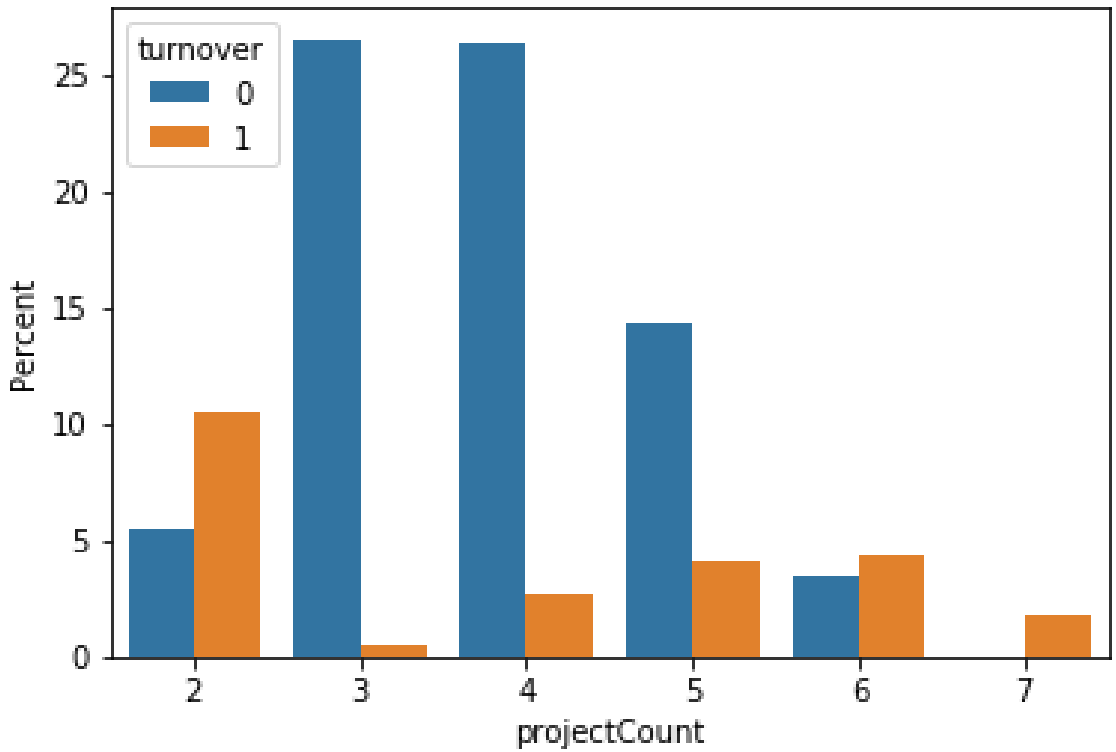


Data Exploration

Project Count V.S. Turnover

Summary:

- More than half of employees with 2,6, and 7 projects left the company
- Majority of the employees who did not leave had 3,4 and 5 projects
- All employees with 7 projects left the company
- There is an increase in turnover as project count increases

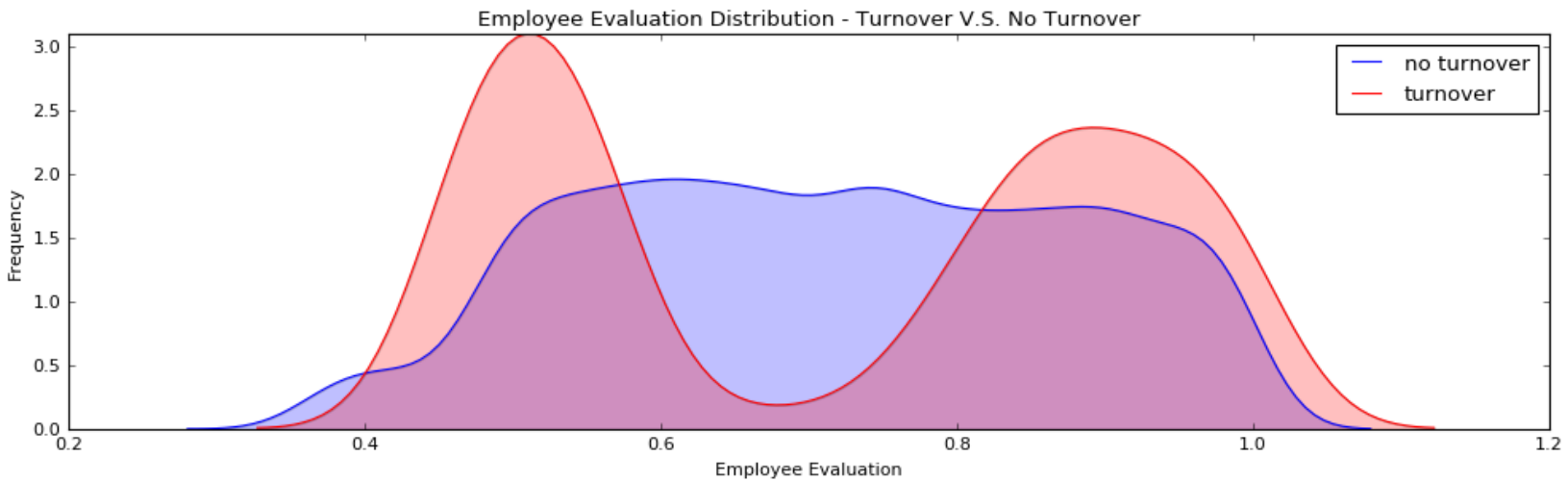


Data Exploration

Turnover V.S. Evaluation

Summary:

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

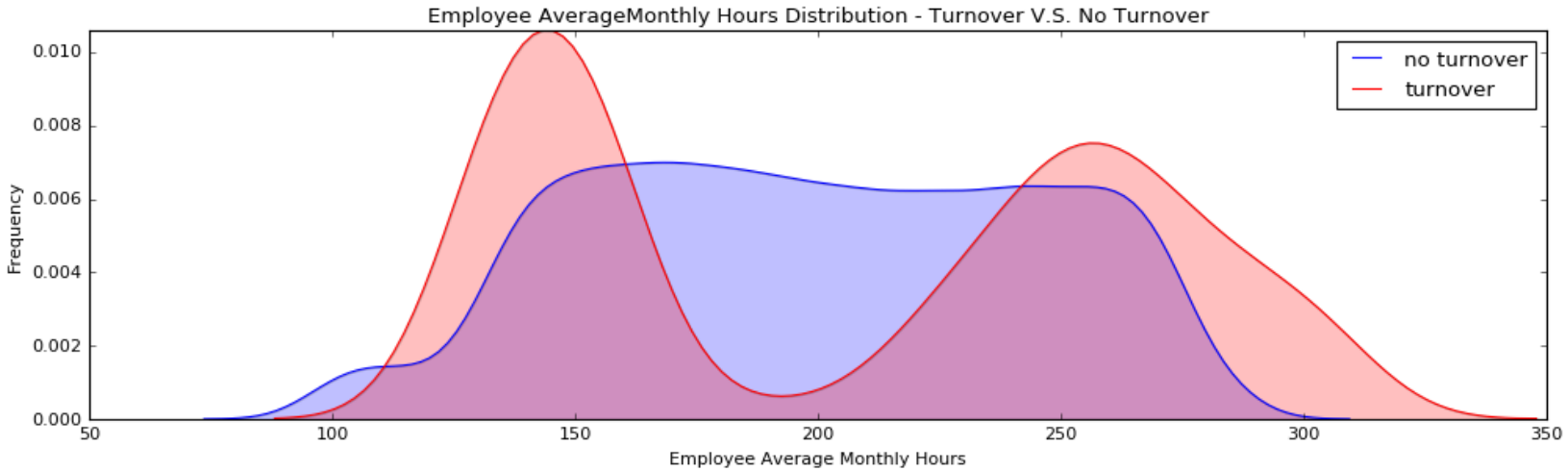


Data Exploration

Turnover V.S. AverageMonthlyHours

Summary:

- Another bi-modal distribution for employees that turnovered
- 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.

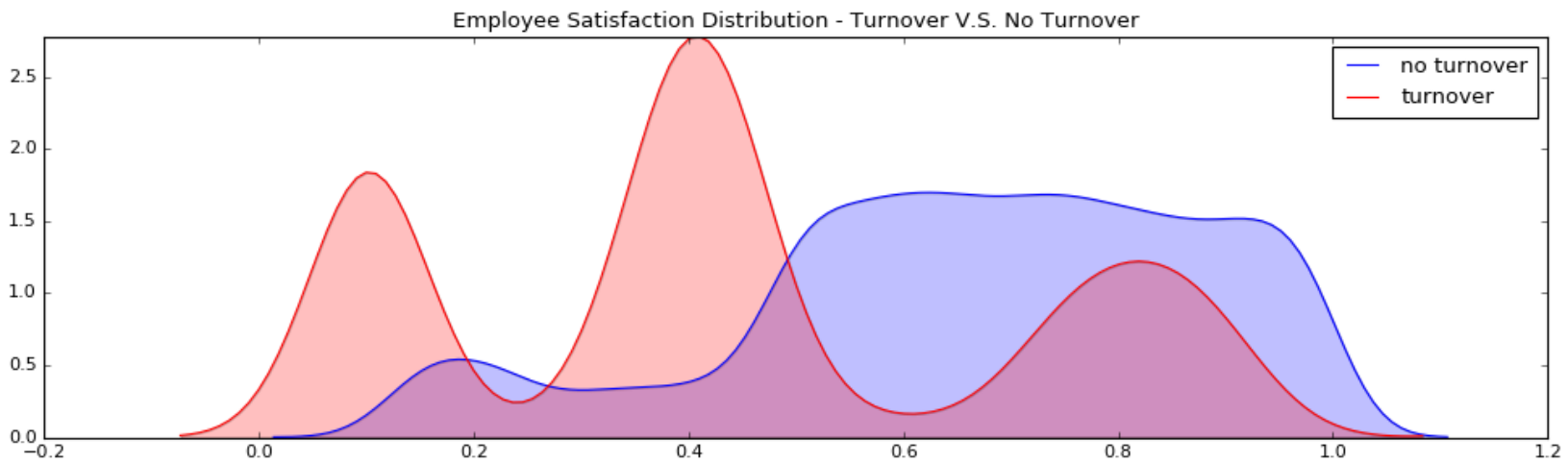


Data Exploration

Turnover V.S. Satisfaction

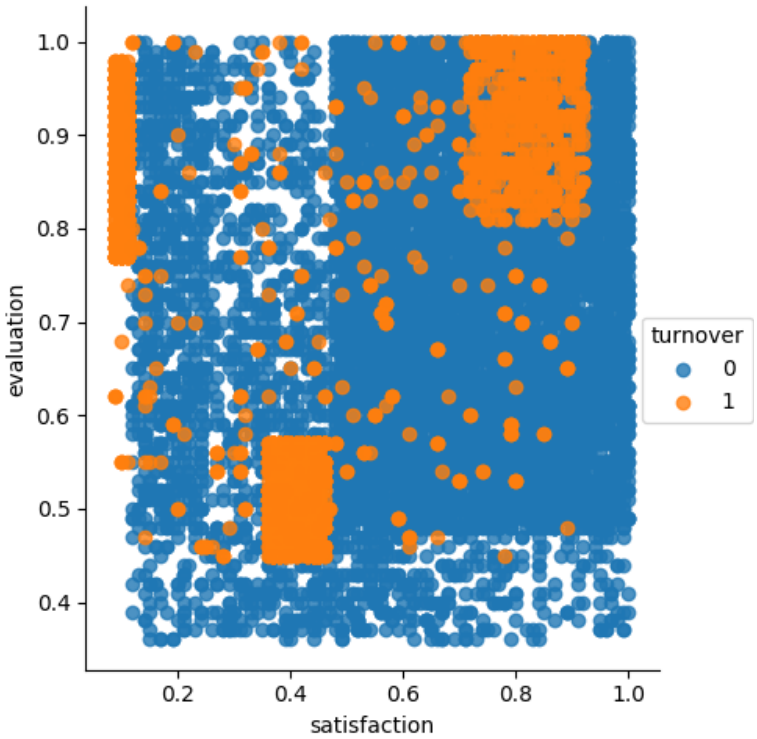
Summary:

- There is a tri-modal distribution for employees that turnovered
- 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

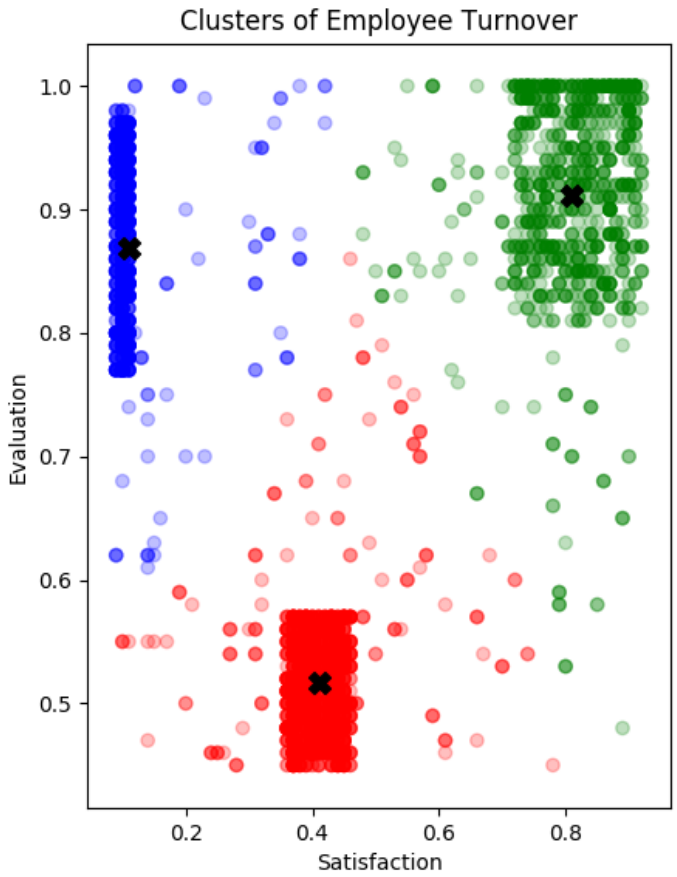


Data Exploration

Clusters in Data



(a) Evaluation V.S Satisfaction



(b) Employee Turnover Cluster

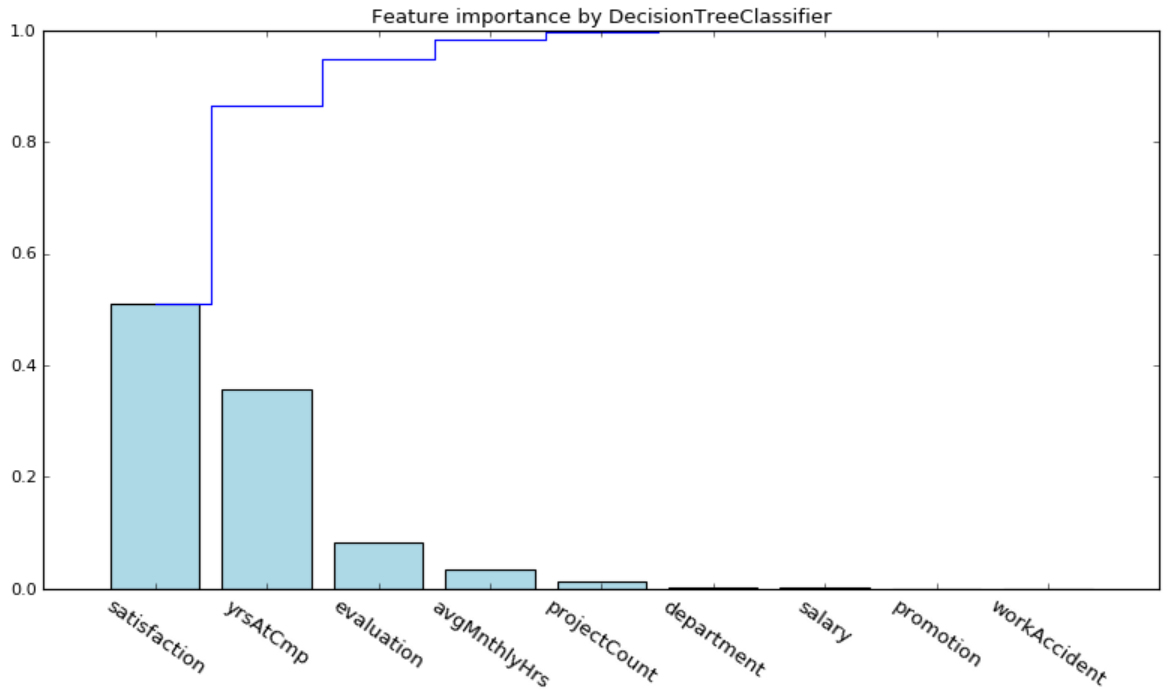
Feature Importance

Summary:

- By using a decision tree classifier, it could rank the features used for the prediction. This is helpful in creating our model for logistic regression because itll be more interpretable to understand what goes into our model when we utilize less features.

Top 3 Features:

- Satisfaction,YearsAtCompany,Evaluation



Overview Of Logistic Regression

- Logistic Regression commonly deals with the issue of how likely an observation is to belong to each group. This model is commonly used to predict the likelihood of an event occurring.
- In contrast to linear regression, the output of logistic regression is transformed with a logit function. This makes the output either 0 or 1. This is a useful model to take advantage of for this problem because we are interested in predicting whether an employee will leave (0) or stay (1).
- Another reason for why logistic regression is the preferred model of choice is because of its interpretability. Logistic regression predicts the outcome of the response variable (turnover) through a set of other explanatory variables, also called predictors.
- Logistic Regression models the probability of success as:

$$\mathbf{logit}[\theta(\mathbf{x})] = \mathbf{log}\left[\frac{\theta(\mathbf{x})}{\mathbf{1}-\theta(\mathbf{x})}\right] = \alpha + \beta_1\mathbf{x_1} + \beta_2\mathbf{x_2} + \cdots + \beta_\mathbf{i}\mathbf{x_i} \quad (1)$$

Modeling

With the elimination of the other variables, well be using the three most important features to create our model: Satisfaction, Evaluation, and YearsAtCompany.
Following overall equation was developed:

- **Employee Turnover Score** =**Satisfaction***(-3.769022) +**Evaluation***(0.207596) + **YearsAtCompany***(0.170145) + 0.181896

```
Optimization terminated successfully.
         Current function value: 0.467233
         Iterations 6
satisfaction      -3.769022
evaluation         0.207596
yearsAtCompany     0.170145
int                0.181896
dtype: float64
```

Test Evaluation

```
 ---Logistic Model---

Logistic Accuracy is 0.77
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
```

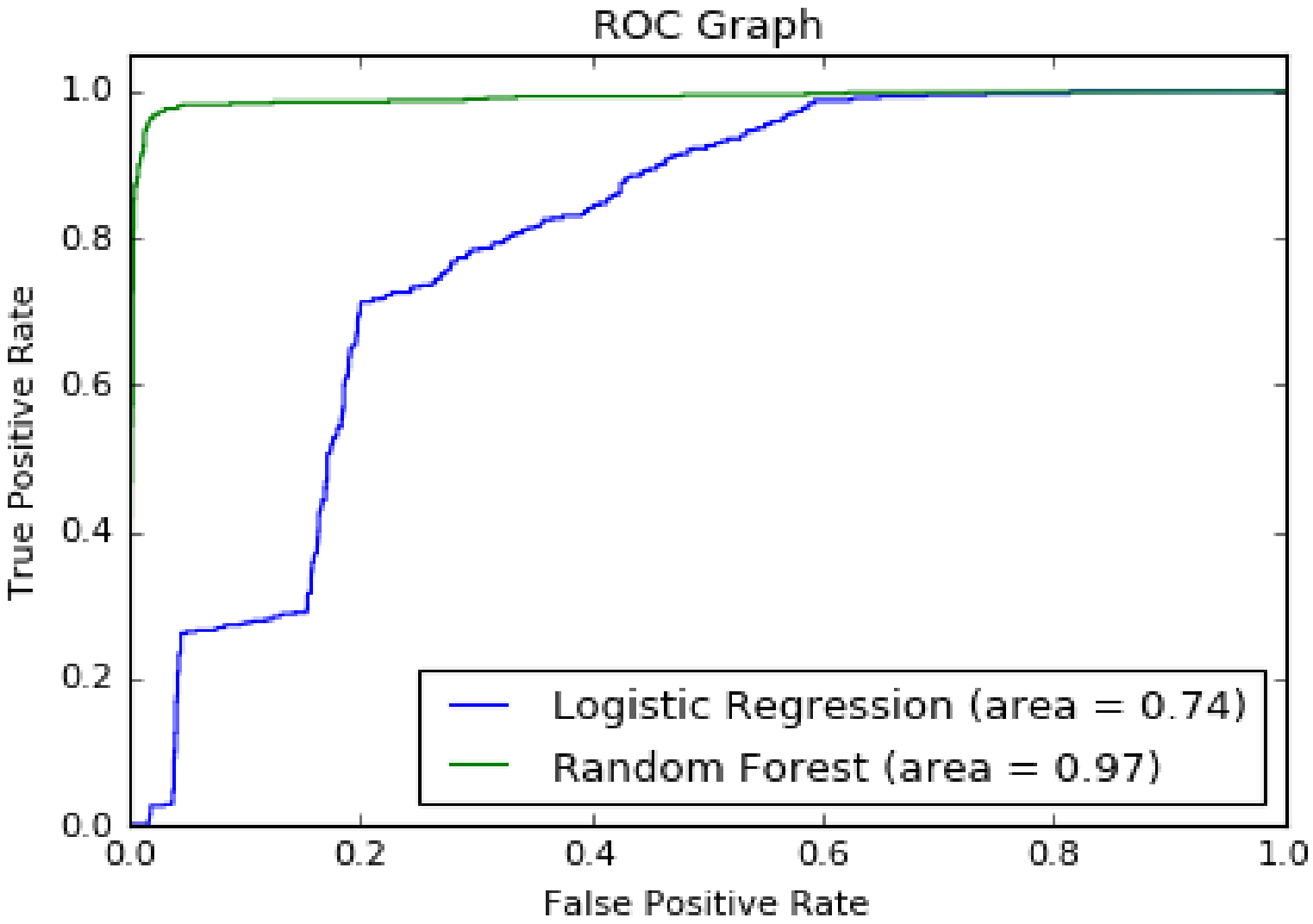
For Example: If we were to use the employee values into the equation:
Satisfaction: 0.7, **Evaluation:** 0.8,**YearsAtCompany:** 3
Employee Turnover Score = (0.7)*(-3.769022) + (0.8)*(0.207596) + (3)*(0.170145) + 0.181896 = 0.14431 = 14 percent
Result:
This employee would have a 14 percent chance of leaving the company.
This information can then be used to form our retention plan.

Random Forest

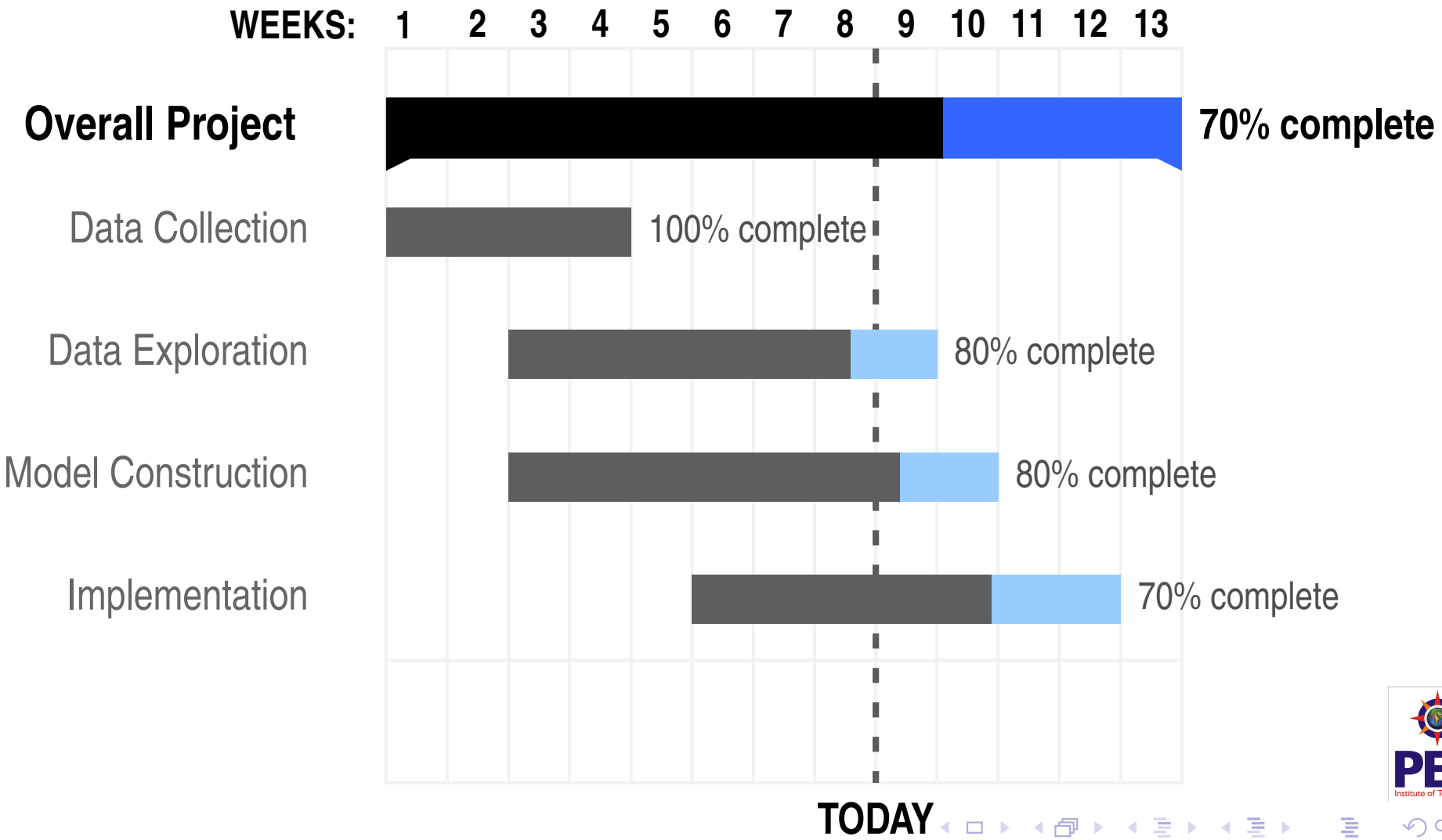
- Random forest takes the average of many decision trees each of which is made from a sample of data.Each tree is weaker than the full decision tree but by combining them we get better overall performance.
- It has methods for balancing error in unbalanced data set.

```
 ---Random Forest Model---
Random Forest-Accuracy is 0.98
Random Forest AUC = 0.97
             precision    recall  f1-score   support

          0       0.99      0.98      0.98      1714
          1       0.95      0.96      0.95       536

avg / total       0.98      0.98      0.98      2250
```

ROC Graph


ROC Graph
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (area = 0.74)
Random Forest (area = 0.97)

Time line of completion of project from Nov 2017-10th April 2018(Gantt Charts).



Expected Outcome/ Results

1.Impact Analysis

The costs of employee attrition ranges from quantifiable numbers to hidden costs.We intend to compute monetary loss for hiring new employees, training and the decreased productivity.

2.Identifying key features

Understanding why an employee is leaving is important to take corrective measures and suppress lurking features. We aim to get to root cause of quiting rather than just predicting if an individual is going to leave or not.

References

Zheng WeiBo1*, Sharan Kaur and Tao Zhi (2010)

A critical review of employee turnover model (1938-2009) and development in perspective of performance

African Journal of Business Management Vol. 4(19), pp. 4146-4158, December Special Review, 2010.

Ladelsky Limor Kessler

The Effect of Oraganizational Culture on IT Employees Turnover Intension in Israle

Anders Frederiksen (2015)

Job Satisfaction and Employee Turnover:A Firm-Level Perspective

IZA Discussion Paper IZA DP No. 9296

Data

- Kaggle HR Data Set

The End