Springboard Capstone Slidedeck

Understanding and Predicting
Employee Turnover | HR Analytics

Randy Lao

Why?

My motivation:

- Interest Human Behavior and Psychology

On my first job:

- Two people quit within two months (small company)

Became curious...

OBJECTIVE

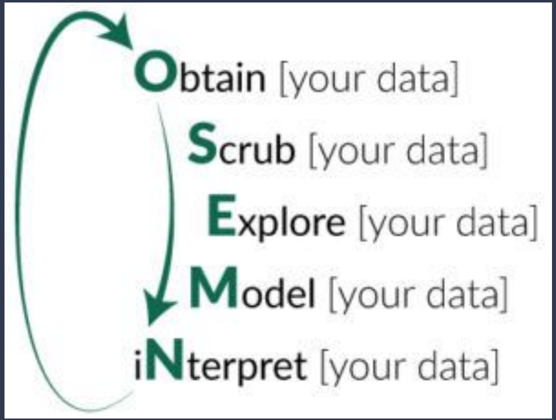
The implementation of this model will allow management to create better decision-making actions.

1. *To **understand** what factors contributed most to employee turnover*

2. *To **create** a model that predicts the likelihood if a certain employee will leave the company or not.*

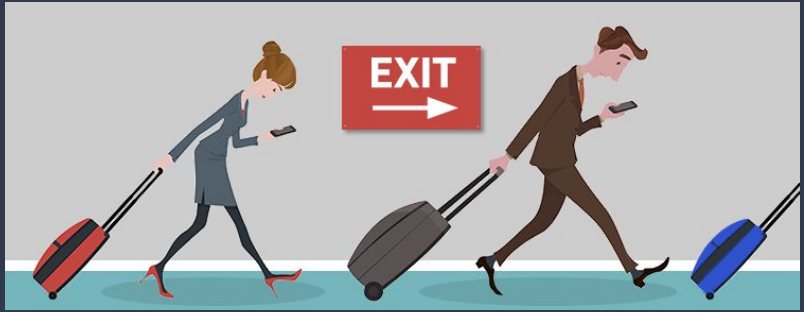
3. *To **create** or **improve** different retention strategies on targeted employees.*

OSEMN Pipeline



1. **O**btaining the data is the first approach in solving the problem.
2. **S**crubbing or cleaning the data. Imputing missing data and converting data to its right format.
3. **E**xploring the data. Understanding our variables and find patterns in our dataset.
4. **M**odeling the data will give us our predictive power on whether an employee will leave.
5. I**N**terpreting the data. What conclusions can we make? What happened?

The Problem



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

Retention Plan

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.

The Dataset

- **Satisfaction**: An employee's level of satisfaction in percentage
- **Evaluation:** An employee's evaluation score in percentage
- **Project Count:** The amount 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.

turnover	satisfaction	evaluation	projectCount	averageMonthlyHours	yearsAtCompany	workAccident	promotion	department
1	0.38	0.53	2	157	3	0	0	sales
1	0.80	0.86	5	262	6	0	0	sales
1	0.11	0.88	7	272	4	0	0	sales
1	0.72	0.87	5	223	5	0	0	sales
1	0.37	0.52	2	159	3	0	0	sales

The Table

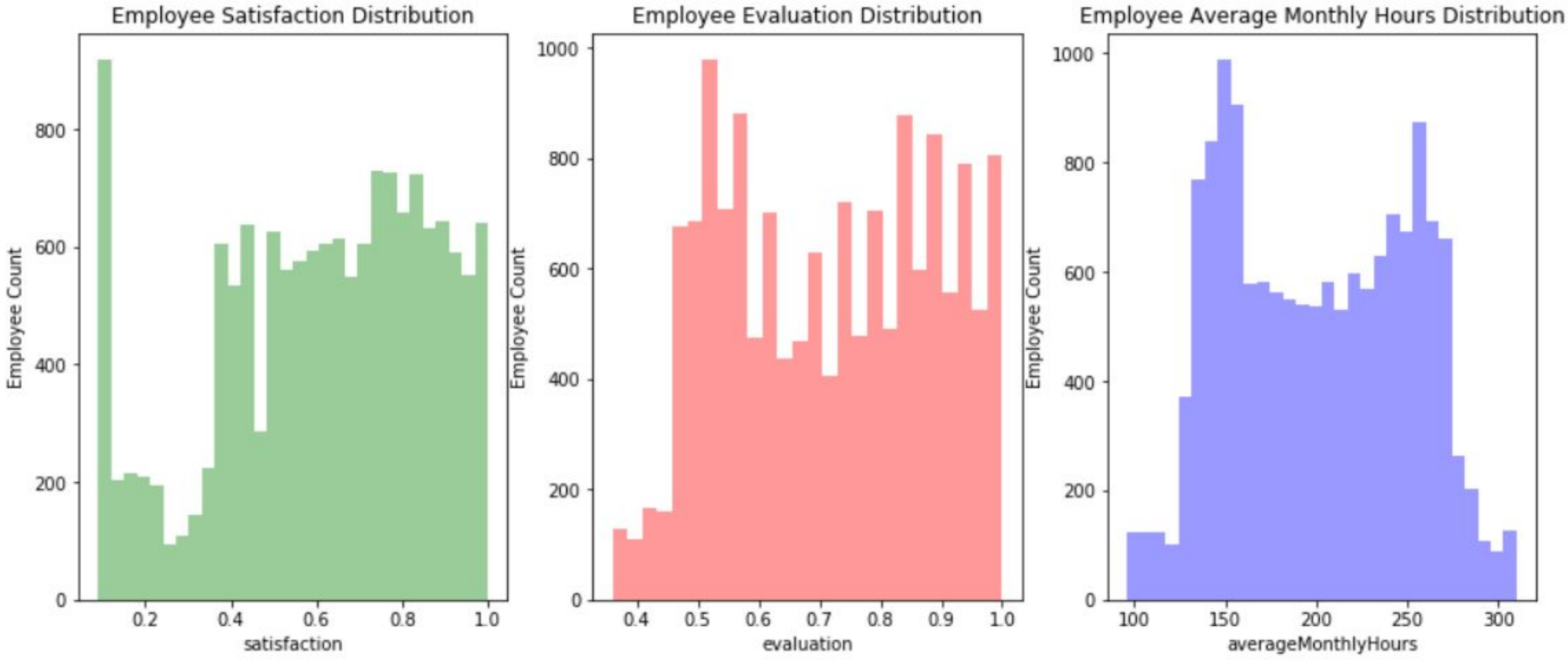
Summary - Turnover VS NoTurnover

The dataset has:

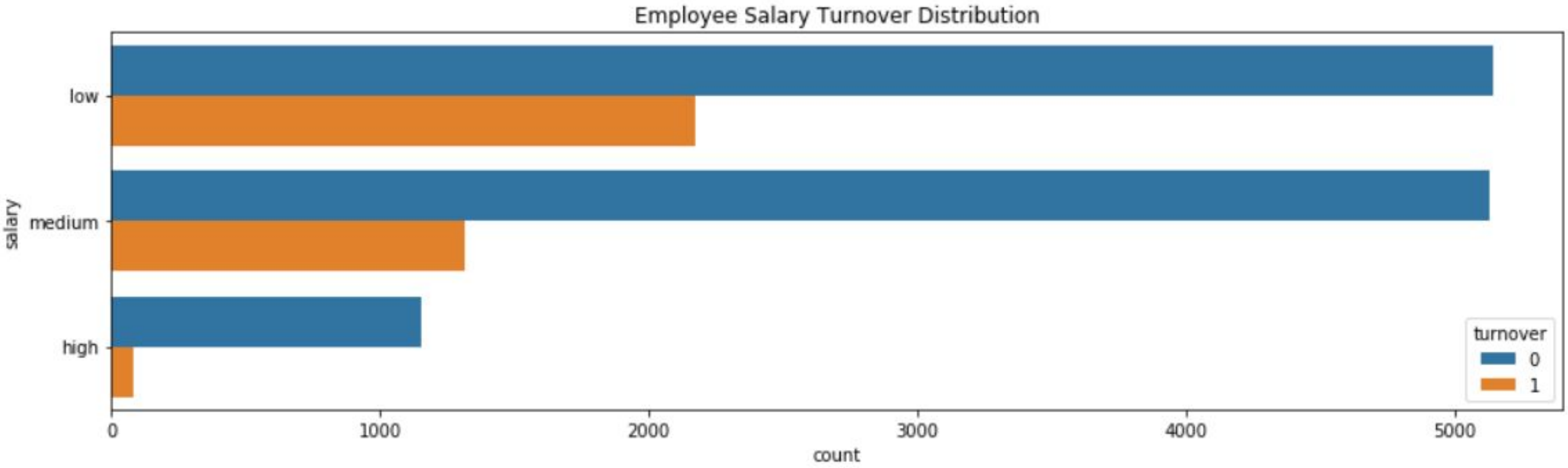
- About **15,000** employee observations and **10** features
- Class Imbalance Problem (Classification)
- The company had a **turnover rate** of about **24%**
- Mean **satisfaction** of employees is **0.61**

	satisfaction	evaluation	projectCount	averageMonthlyHours	yearsAtCompany	workAccident	promotion
turnover							
0	0.666810	0.715473	3.786664	199.060203	3.380032	0.175009	0.026251
1	0.440098	0.718113	3.855503	207.419210	3.876505	0.047326	0.005321

Satisfaction & Evaluation & Hours Distribution

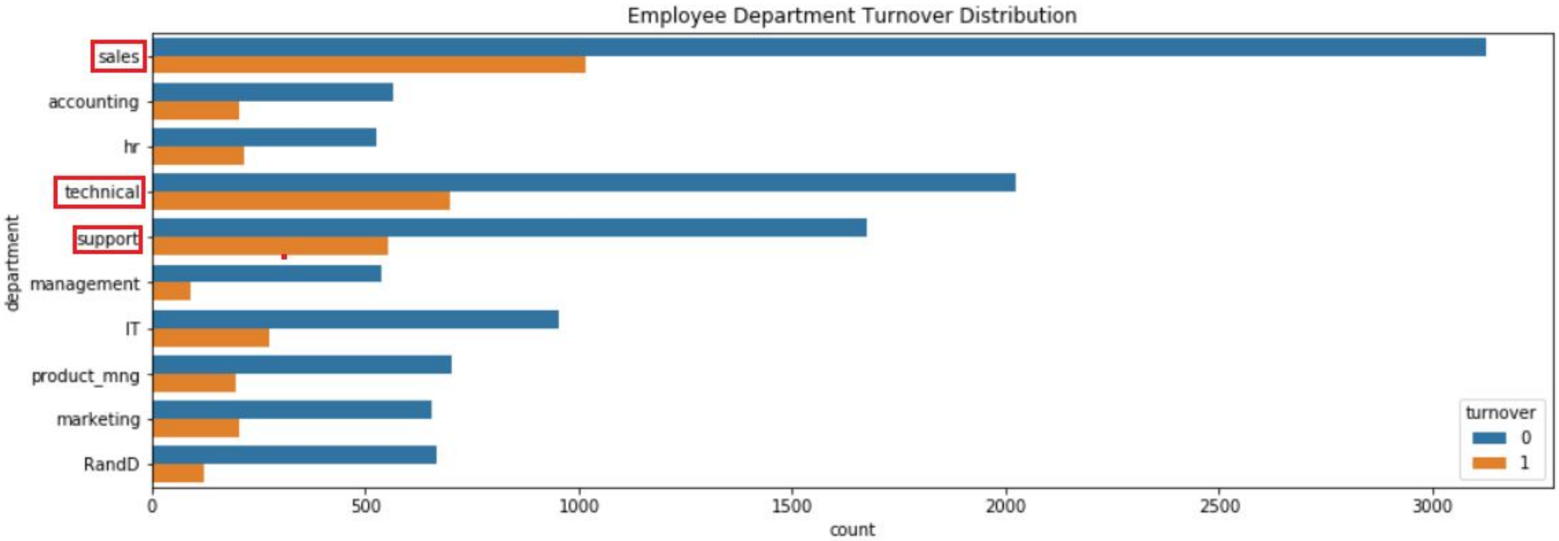


Employee Salary Distribution

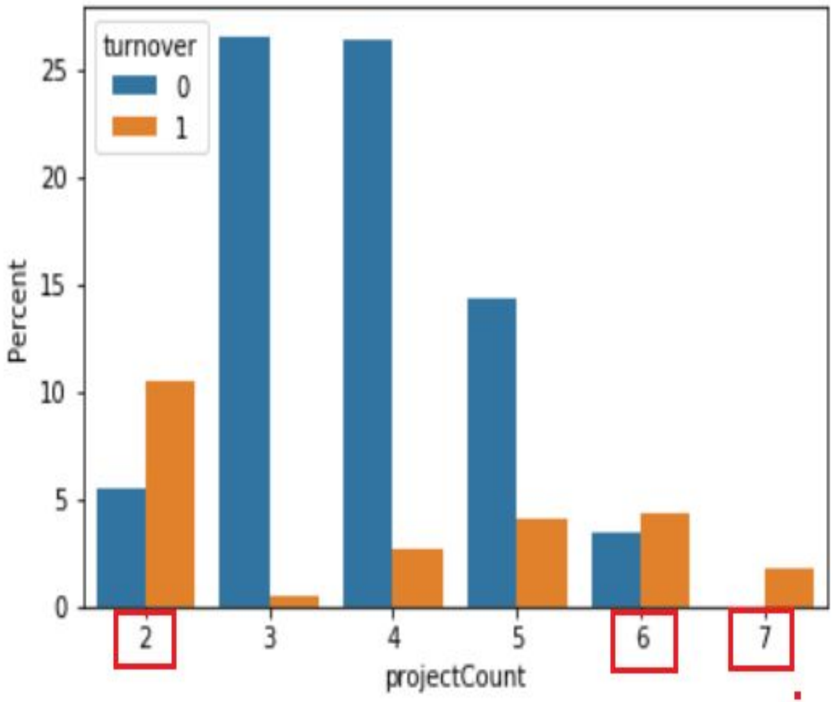


Department Distribution

- 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



Summary

- 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

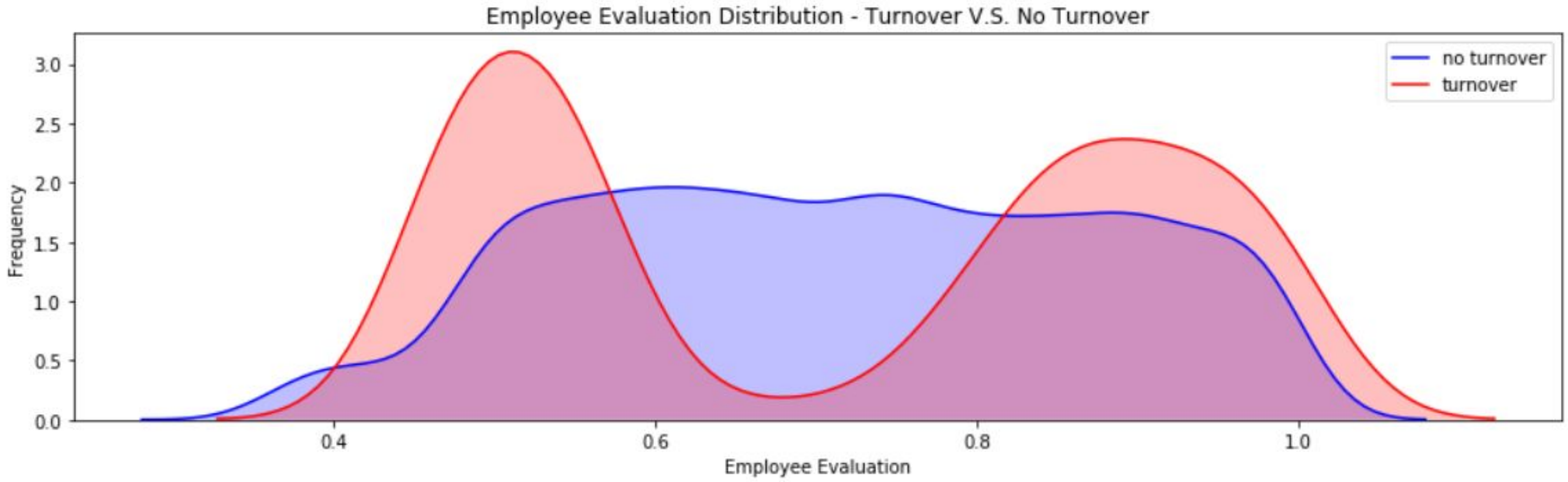
Stop and Think

- Why are employees leaving at the lower/higher spectrum of project counts?
- Does this means that employees with project counts **2 or less** are not worked hard enough or are not highly valued, thus leaving the company?
- Do employees with **6+ projects** are getting **overworked**, thus leaving the company?

Evaluation Distribution

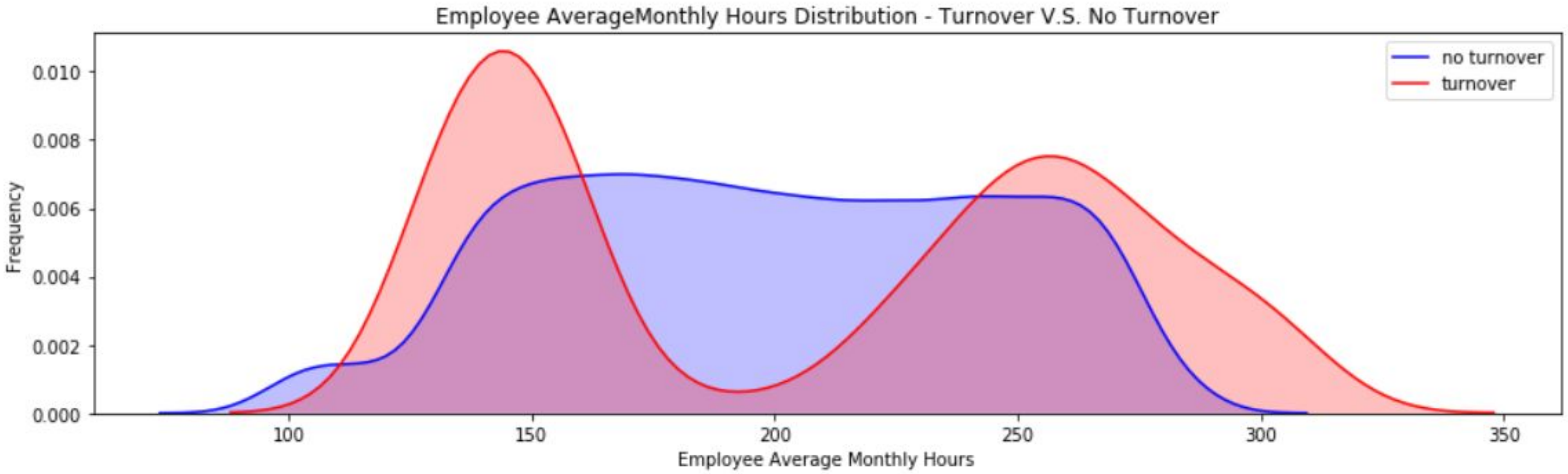
Summary:

- There is a **biomodal** 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



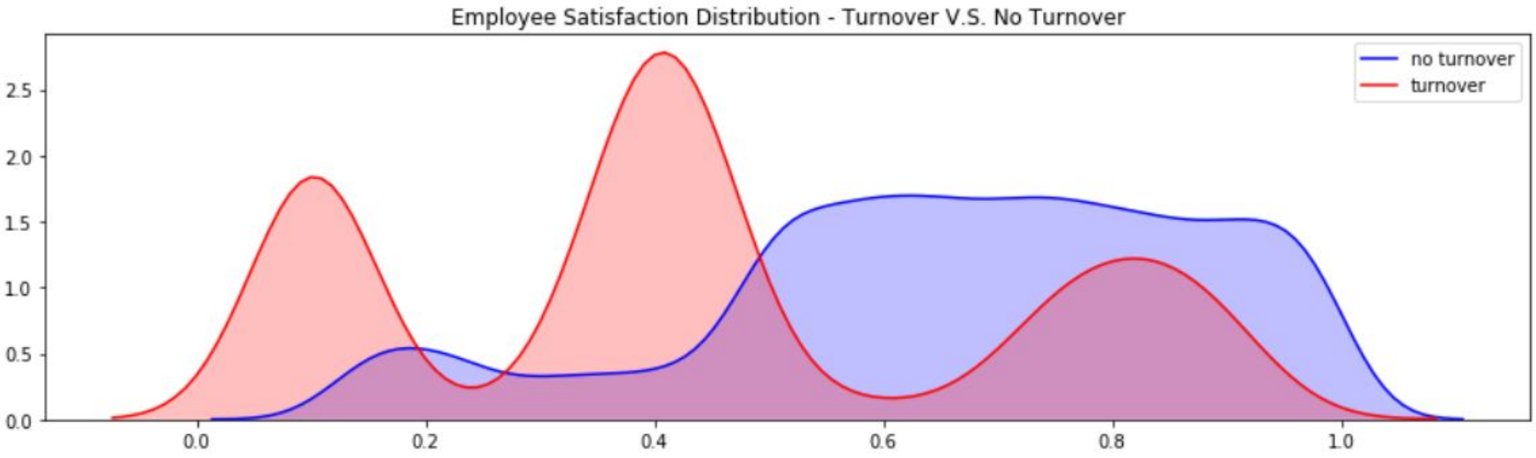
Average Monthly Hours Distribution

- 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**.

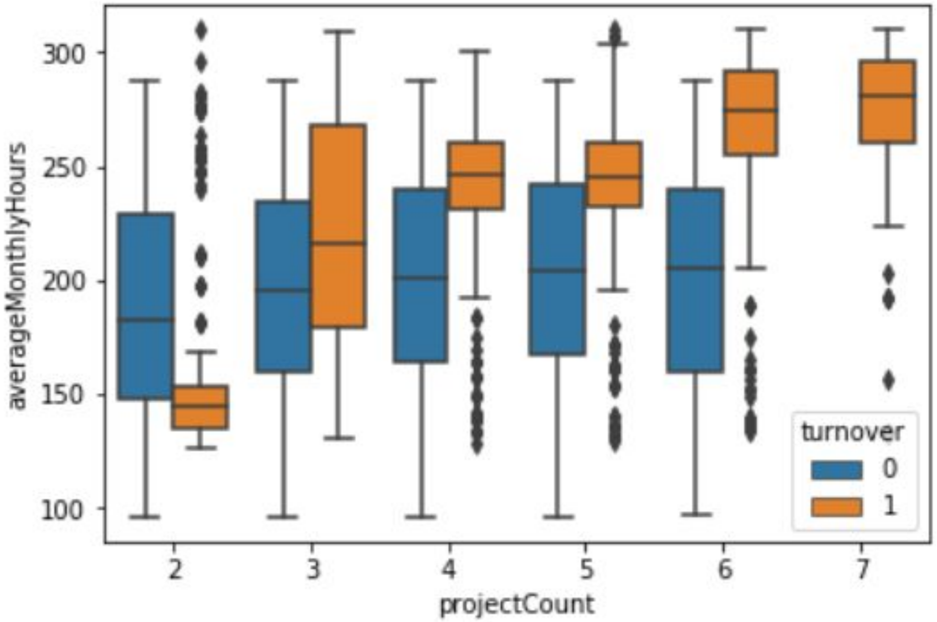


Satisfaction Distribution

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



Monthly Hours VS Project Count



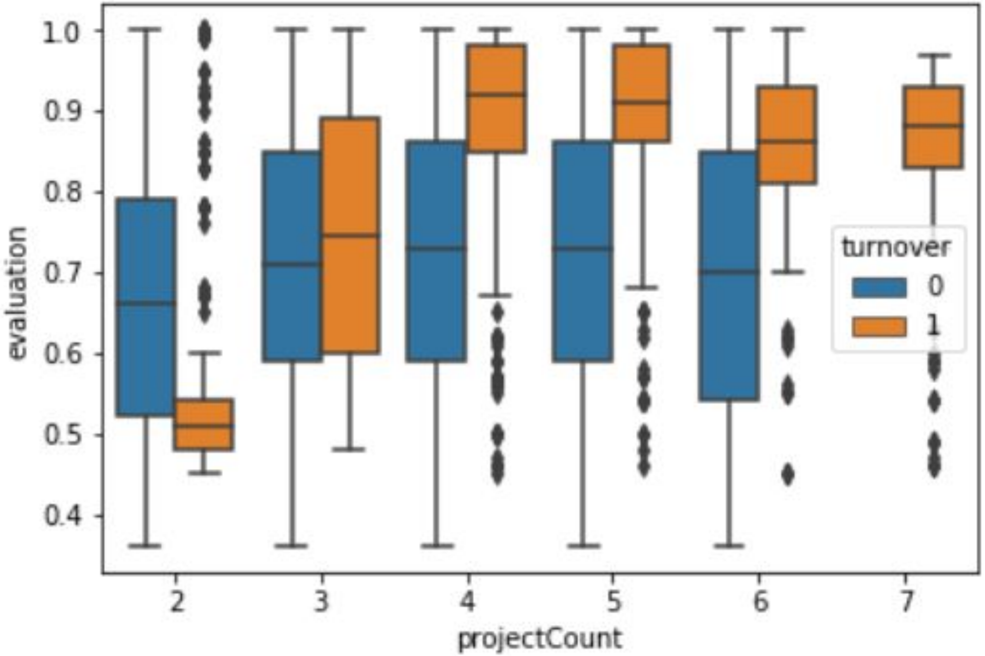
- Employees who had **No-Turnover** had an **even** distribution of average monthly hours as the project count increased

- Employees who had **Turnover** had an **INCREASE** in average monthly hours as the project count increased

Question:

Why is it that employees who left worked more hours than employees who didn't, even with the same project count?

Evaluation VS Project Count

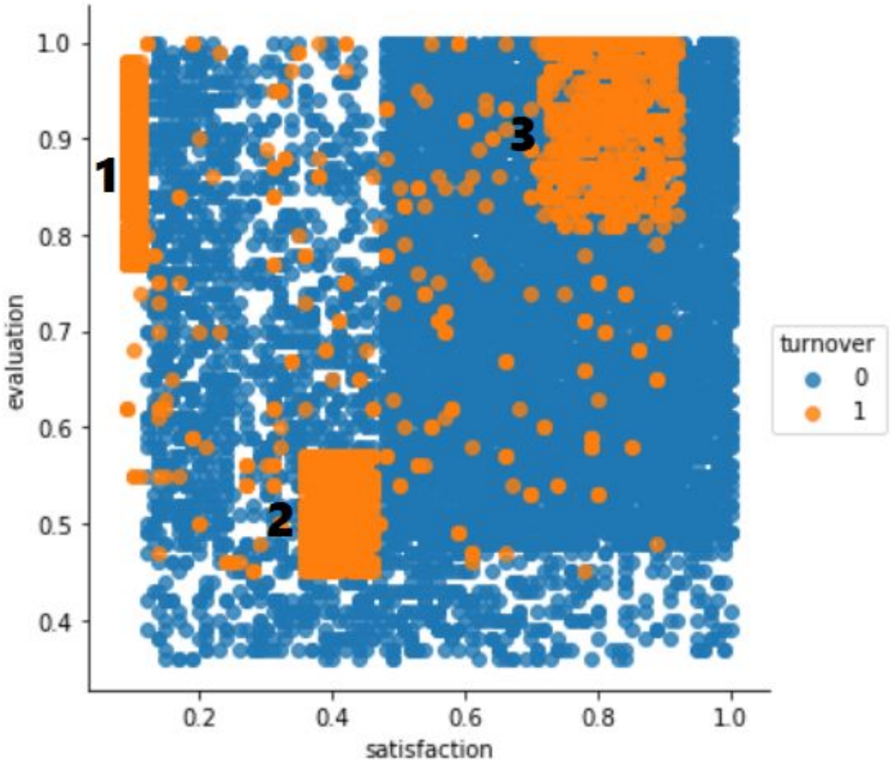


- There is an **INCREASE** in evaluation for employees who did more projects within the **turnover group**.

- For the **non-turnover group**, employees here had a **consistent** evaluation score despite the increase in project counts.

Question:
Why are employees leaving the company more when they are evaluated highly as project count increases?

Satisfaction VS Evaluation

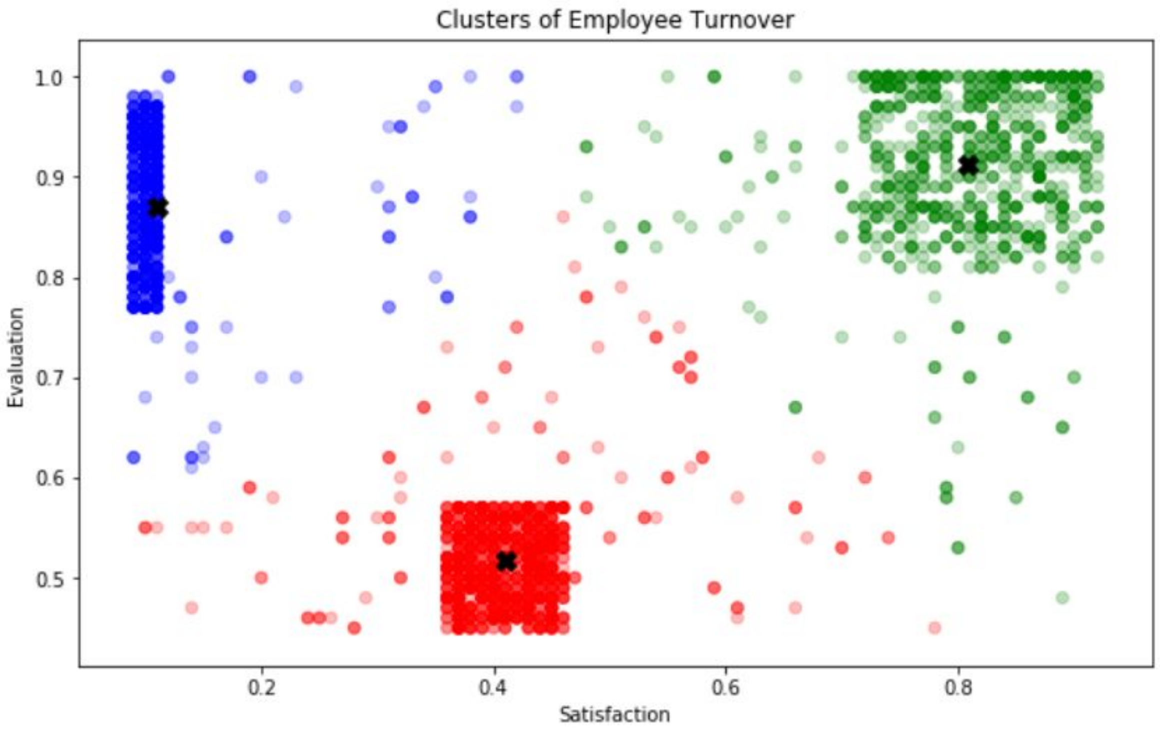


Cluster 1 (Highly Valued, But Sad)
Satisfaction was below **0.2** and evaluations were greater than **0.75**.

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**.

KMeans Clustering

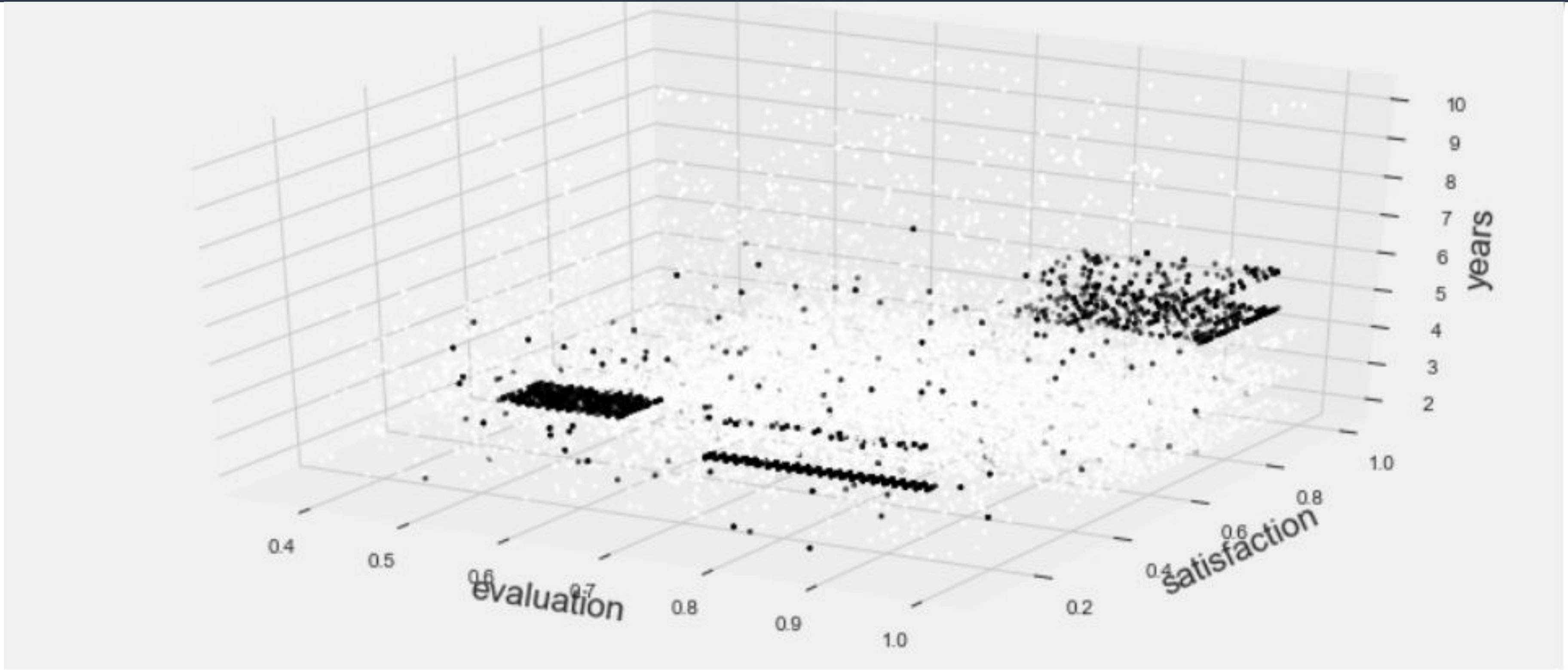


Blue - Overworked Employee

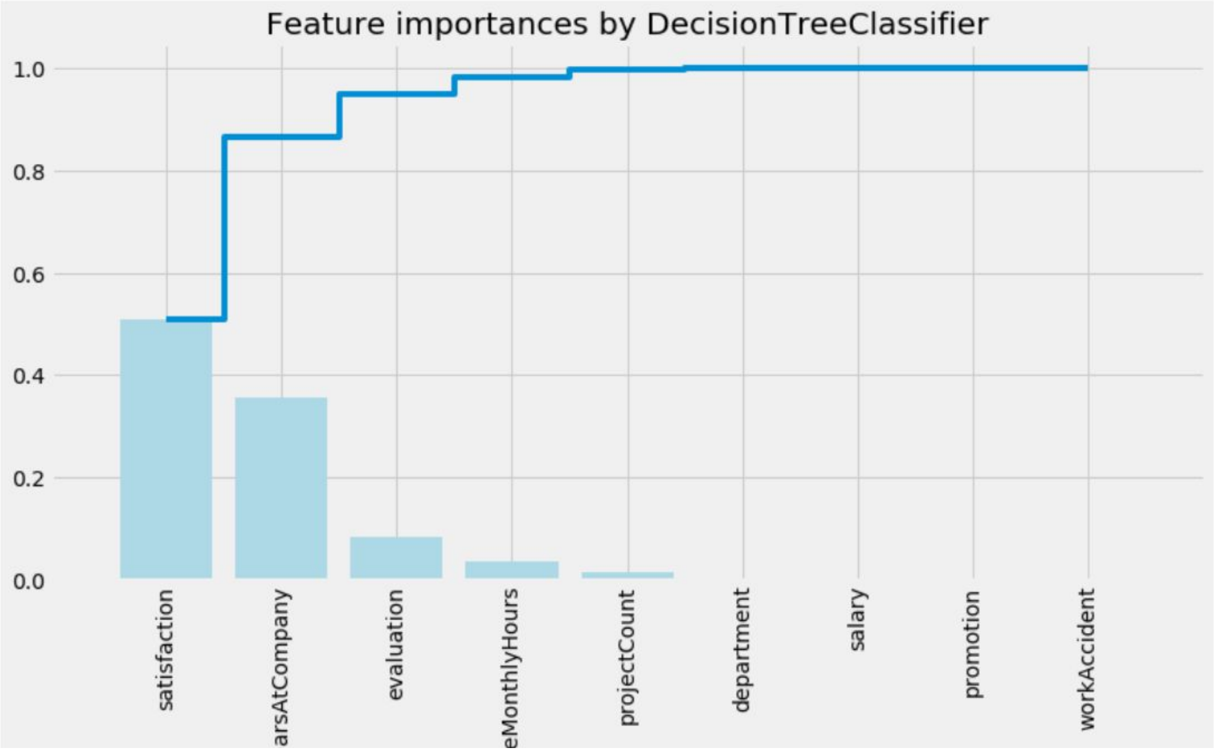
Red - Underperforming Employee

Green - Ideal Employee

3D Cluster (Evaluation + Satisfaction + Years)



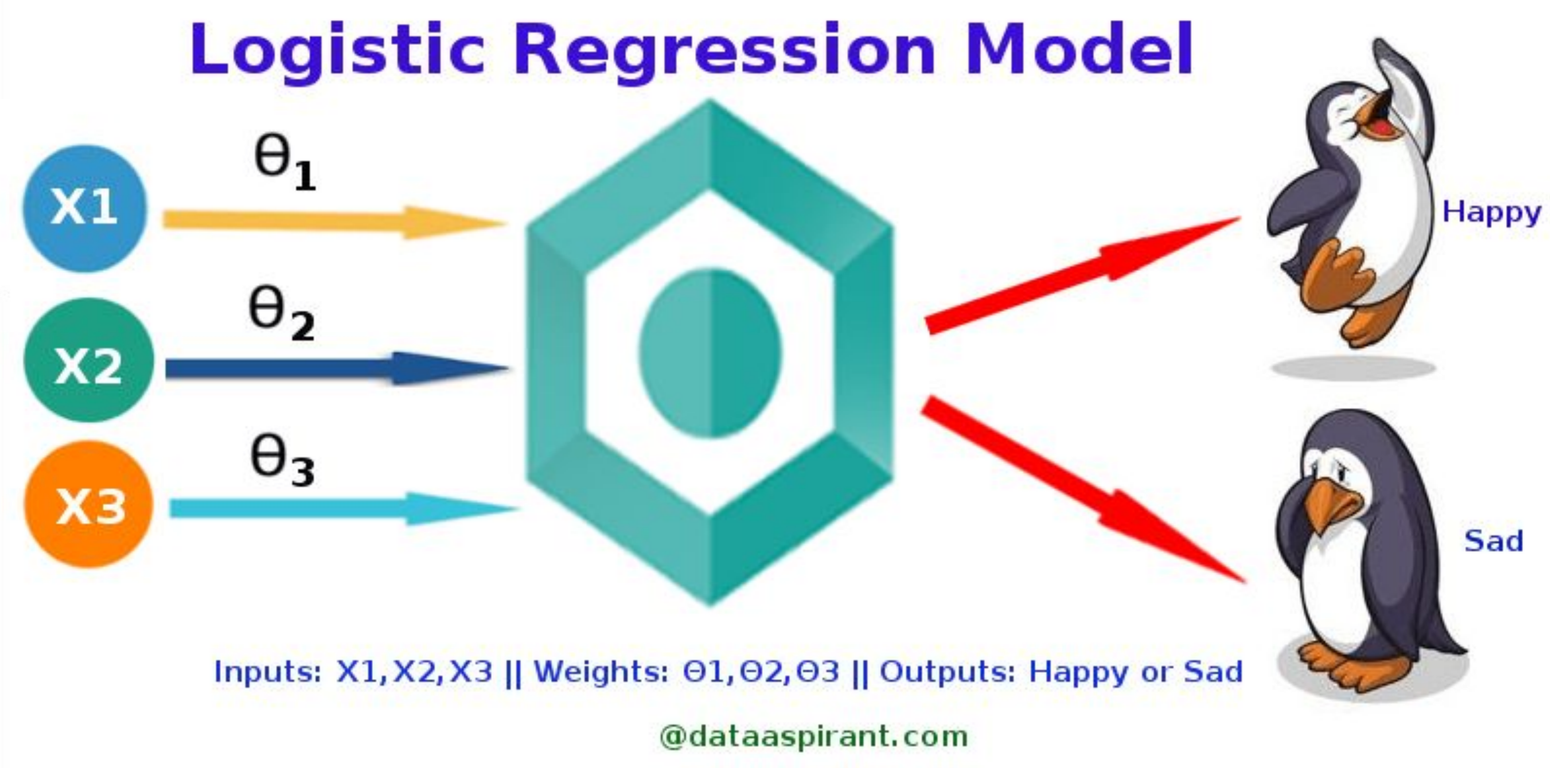
Decision Tree – Feature Importance



Top 3 Features:

1. **Satisfaction**
2. **YearsAtCompany**
3. **Evaluation**

Introduction to Logistic Regression

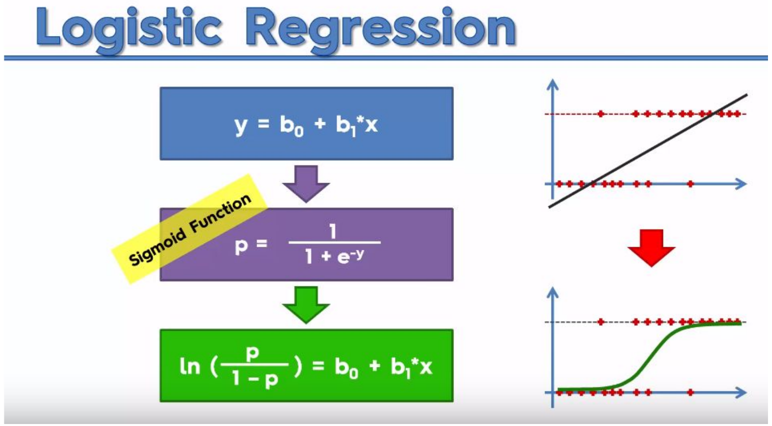


Logistic Regression

$$\operatorname{logit}[\theta(x)] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

The equation above shows the relationship between the dependent/independent:

(θ(x)) - Dependent Variable (Outcome)
(xi) **-** Independent variables or predictor of event
(α) - is the constant of the equation
(β) - is the coefficient of the predictor variables or weights



```
/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
```

Logistic Regression Coefficients

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.

$$\mathrm{logit}\left[\theta(\mathbf{x})\right] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

Hypothetical Example

Intepretation of Score

If you were to use these employee values into the equation:

- **Satisfaction**: 0.7
- **Evaluation**: 0.8
- **YearsAtCompany**: 3

You would get:

Employee Turnover Score = (**0.7**)*(-3.769022)* + *(**0.8**)*(0.207596) + (**3**)*(0.170145) + 0.181896 = 0.14431 = **14%***

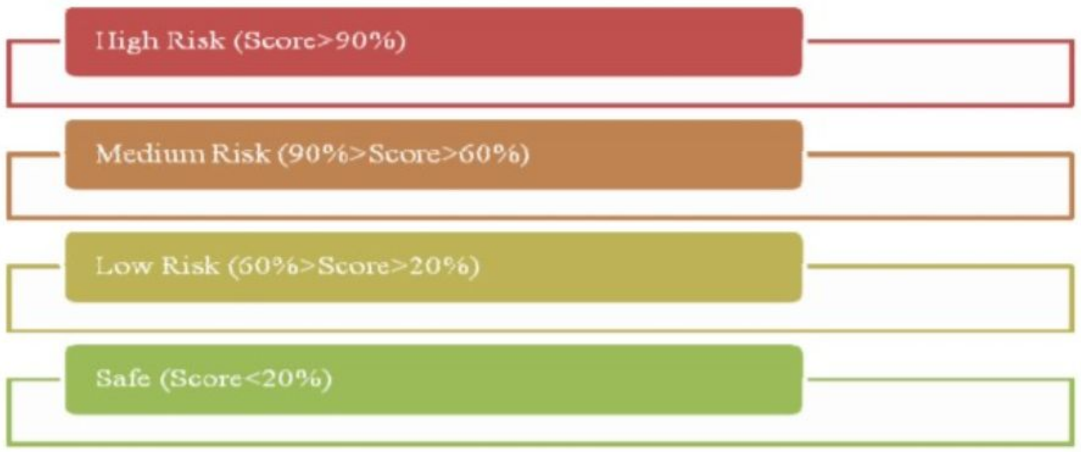
Result: This employee would have a **14%** chance of leaving the company. This information can then be used to form our retention plan.

Retention Plan Using Logistic Regression

Reference: http://rupeshkhare.com/wp-content/uploads/2013/12/Employee-Attrition-Risk-Assessment-using-Logistic-Regression-Analysis.pdf

1. **Safe Zone (Green)** – Employees within this zone are considered safe.
2. **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.
3. **Medium Risk Zone (Orange)** – Employees within this zone are at risk of turnover. Action should be taken and monitored accordingly.
4. **High Risk Zone (Red)** – Employees within this zone are considered to have the highest chance of turnover. Action should be taken immediately.

So with our example above, the employee with a **14%** turnover score will be in the **safe zone**.



Class Imbalance –Evaluation Metric

Precision and Recall / Class Imbalance

This dataset is an example of a class imbalance problem because of the skewed distribution of employees who did and did not leave. More skewed the class means that accuracy breaks down.

In this case, evaluating our model's algorithm based on **accuracy** is the **wrong** thing to measure. We would have to know the different errors that we care about and correct decisions. Accuracy alone does not measure an important concept that needs to be taken into consideration in this type of evaluation: **False Positive** and **False Negative** errors.

False Positives (Type I Error): You predict that the employee will leave, but do not

False Negatives (Type II Error): You predict that the employee will not leave, but does leave

In this problem, what type of errors do we care about more? False Positives or False Negatives?

False Negative V.S. False Positive

The evaluation of our model will be highly dependent on how the organization would want to have its priorities on:

1. **Does a False Positive cost more?** (Incentives to employees, but don't need it)
2. **Does a False Negative cost more?** (No incentive to employees, but they need it)

My opinion: The cost of a false negative where we don't provide support to the employees that need help might outweigh the cost of a false positive where we provide help but don't need it.

So in order to choose the right metric, we have to ask what costs more?
False Positive or False Negatives?

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

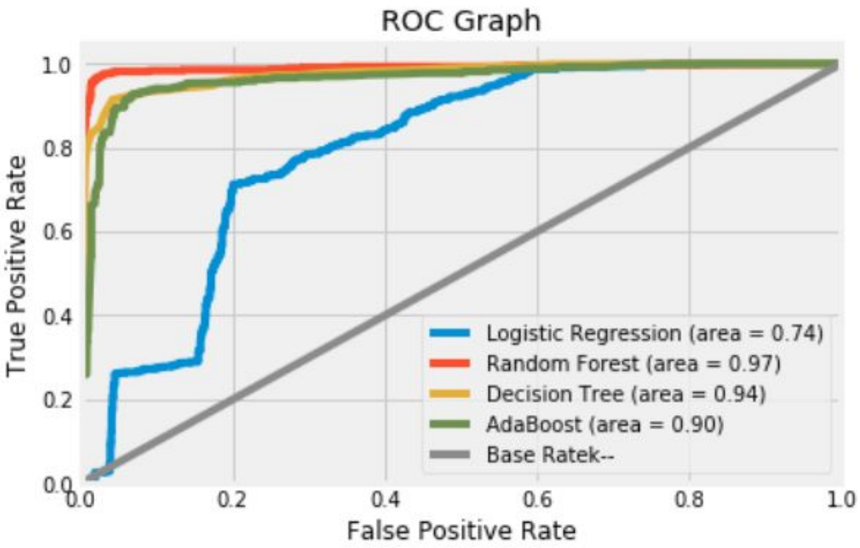
	precision	recall	f1-score	support
0	0.97	0.96	0.97	1714
1	0.87	0.91	0.89	536
avg / total	0.95	0.95	0.95	2250

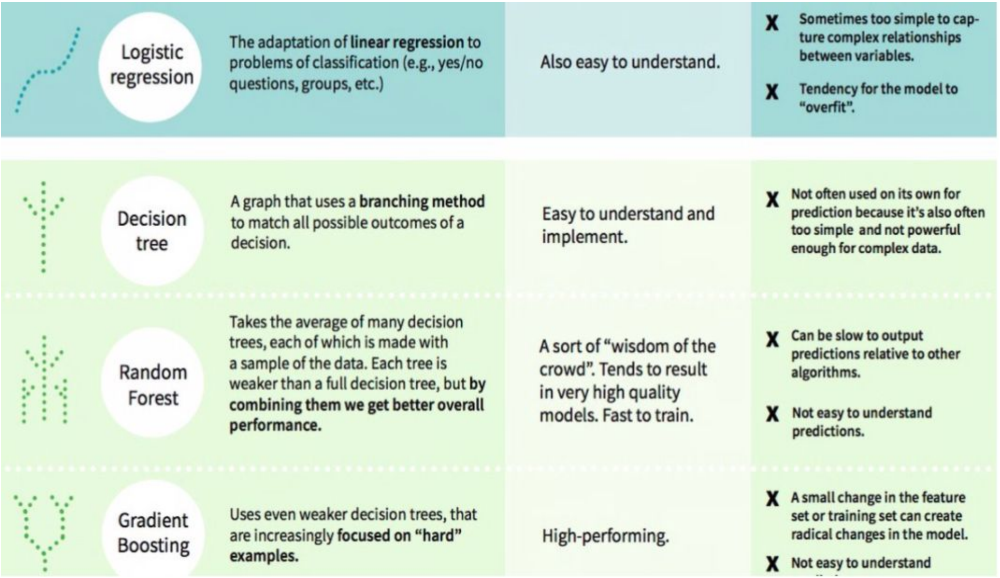
---AdaBoost Model---

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

Model Comparison





Logistic regression	The adaptation of **linear regression** to problems of classification (e.g., yes/no questions, groups, etc.)	Also easy to understand.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
Decision tree	A graph that uses a **branching method** to match all possible outcomes of a decision.	Easy to understand and implement.	X Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.
Random Forest	Takes the average of many decision trees, each of which is made with a sample of the data. Each tree is weaker than a full decision tree, but **by combining them we get better overall performance.**	A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.	X Can be slow to output predictions relative to other algorithms. X Not easy to understand predictions.
Gradient Boosting	Uses even weaker decision trees, that are increasingly **focused on "hard" examples.**	High-performing.	X A small change in the feature set or training set can create radical changes in the model. X Not easy to understand

The ROC Graph allows you to classify your accuracy for your true labels and false labels

Summary

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

The Analysis

Descriptive Analytics - What's happening?
Generally, employees are leaving due to low satisfaction from the amount of hours they work and project counts.

Diagnostic Analytics - Why is it happening?
We'll need to dive in deeper by gathering more information and asking more questions. But from the data, employees are leaving from two ends of the extremes. Low/High Satisfaction and Low/High Years at the company.

Predictive Analytics - What's likely to happen?
Using the logistic regression model, we are able to not only predict whether or not the employee might leave, but we can also get their probability of leaving.

Prescriptive Analytics - What do I need to do?
Using our probability scores for each employee, we can provide further assistance with the help of the Retention Plan.

Zig Ziglar

"You don't build a business -- you build people-- and then people build the business"
- Zig Ziglar

Problem Statement & Solution RECAP

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.

Using the Retention Plan

- Use the retention plan based on the **probability** of an employee leaving.

- Use the retention plan based on the **expected loss** of an employee leaving.

Questions to Ask

1. How would you define high, low, medium performer? Can’t base off Evaluation because it’ll be too biased. Evaluations are inconsistent through departments and highly dependent on relationships.
2. Define if 24% turnover is bad or not.
3. Is this **voluntary** or **involuntary** turnover? Are they contractors, interns, or part-time?

Potential New Features

1. Get employee health benefits data
2. Get employee address and location
3. Get employee marriage status
4. Get employee sex (maybe gender imbalance)
5. Get employee’s manager name
6. Get employee’s department

Do Exit Interviews to get more data!

Future Work

1. This technique to predict employee attrition can be applied to every organization based on employee demographic data.
2. This model should be updated periodically and continuous feedback from employees will definitely help the organization in thriving.
3. Instead of trying to retain everyone, an organization should identify precisely who needs to be kept on board, and how the company can continue to appeal the high potential employees.