DDSANYALTICS Talent Management Initiative

INITIAL ANALYSIS & REPORT OUT #1

- IDENTIFY & DESCRIBE AT-RISK ATTRITION DEMOGRA
- PREDICTIVE MODELING POTENTIAL

Project Overview

Description & Deliverables

Description

Talent management is defined as the iterative process of developing and retaining employees. It may include workforce planning, employee training programs, identifying high-potential employees and reducing/preventing voluntary employee turnover (attrition). To gain a competitive edge over its competition, DDSAnalytics is planning to leverage data science for talent management. The executive leadership has identified predicting employee turnover as its first application of data science for talent management. Before the business green lights the project, they have tasked your data science team to conduct an analysis of existing employee data.

Deliverables

- Identify the top factors which contribute tot turnover. Clearly document and defend the analysis.
- Discuss any other material insights, trends, or observations gleaned from the dataset.
- Construct models which predict attrition and monthly income

Executive Summary

DDSAnalytics has concluded an initial analysis of employee attribute data which demonstrates the ability to predict, using a naïve Bayes classification model, an individual employee's voluntary attrition potential, as well as their monthly income using a multiple linear regression model. Many of these explanatory attributes may already exist in employee files while the others could be easily, and cost-effectively, be collected.

The analysis found these independent variables to be most impactful and their specific demographic values to be most at-risk for attrition:

(1)	Does the employee work overtime?	Most At-Risk: "Yes"
(2)	Employee's total years with the Company	Most At-Risk: 0 – 10
(3)	Employee's martial status	Most At-Risk: "Single"
(4)	Employee's tenure in their current role	Most At-Risk: 0 - 4
(5)	Monthly Income	Most At-Risk: $0 - $5,811
(6)	Department in which employee works	Most At-Risk: "Sales"
(7)	Role held by the employee	Most At-Risk: "Sales Rep"
(8)	Age of employee	Most At-Risk: 18 – 28
(9)	Employee's tenure with their current manager	Most At-Risk: 0 - 4

Executive Summary

Continued

DDSAnalytics has also found that a relatively simple multiple linear regression (MLR) model can effectively describe, and predict, an employee's monthly income.

The MLR model below achieves a statistically significant (p-value << 0.05) solution with an Adjusted R^2 = 91%.

$$MonthlyIncome = \beta_0 + \beta_1 Distance + \beta_2 JobLevel + \beta_3 PercSalaryHike + \beta_4 TotalWorkingYears + \beta_5 YearsWithCurrentManager$$

MODEL PARAMETERIZATION

Residuals:

Min	1Q	Median	3Q	Max
-5759	-872	16	740	4035

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
β_0 (Intercept)	-1707.30	227.30	-7.51	1.5e-13
β_1 DistanceFromHome	-15.57	5.74	-2.71	0.0068
β_2 JobLevel	3723.77	68.43	54.41	< 2e-16
β_3 PercentSalaryHike	9.57	12.72	0.75	0.4519
β_4 TotalWorkingYears	68.12	10.41	6.54	1.0e-10
β_5 YearsWithCurrManager	-60.04	14.70	-4.09	4.8e-05

Residual standard error: 1370 on 864 degrees of freedom
Multiple R-squared: 0.911, Adjusted R-squared: 0.911
F-statistic: 1.78e+03 on 5 and 864 DF, p-value: <2e-16

Overview

Dataset Description

The dataset provided to the project team captures general demographic information for each employee (example: age, gender, marital status) company-specific information (example: total years with the company, department, role), along with employee-reported satisfaction scores for several areas (example: Work Life Balance, Environment Satisfaction).

Total rows: 870

Total gross/net columns: 36/32

- ID/EmployeeCount/Over18/StandardHours [Columns with singular values removed from the analysis]

Character/Categorical Fields: 8

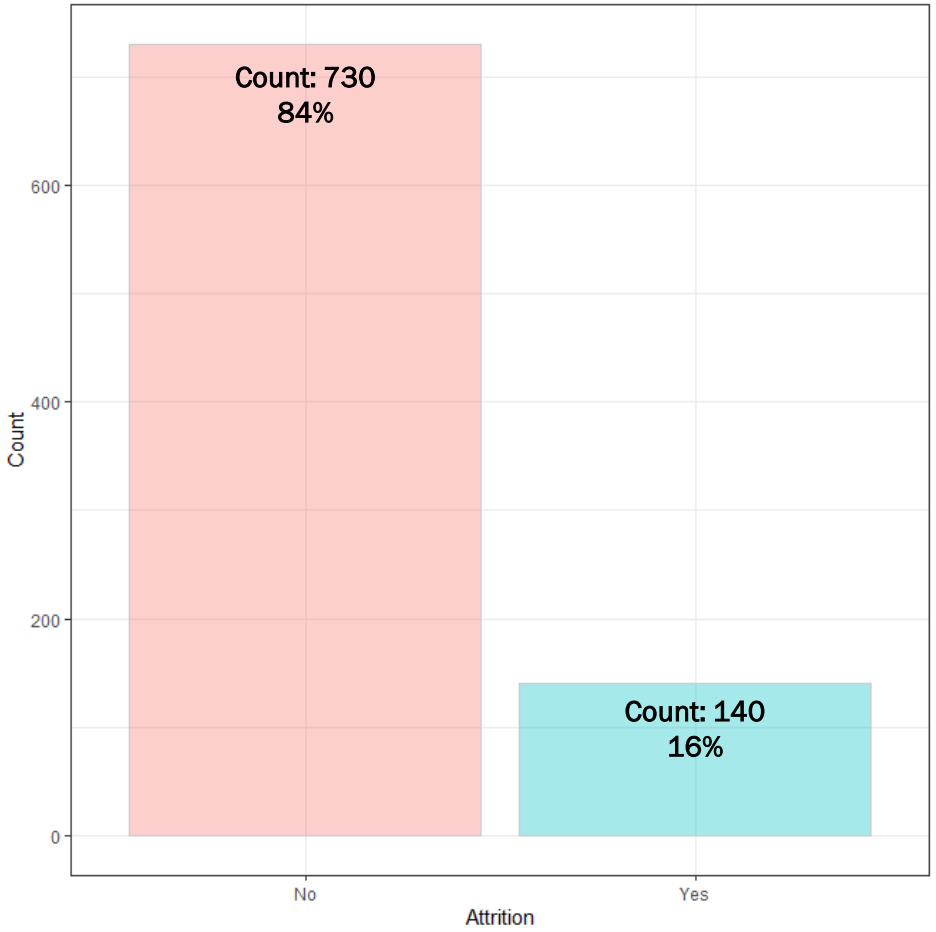
Numeric Fields: 24

- 10 integer fields behave more as categorical variables

Mercifully, the dataset did not contain any missing or null values

Total Employee Base

Total Attrition Rate



Observations

US companies had an average turnover rate of 22% in 2018, with 15% attributed to voluntary turnover.[1]

- The majority (81%) of employees who left voluntarily did so for a better job opportunity.[2]

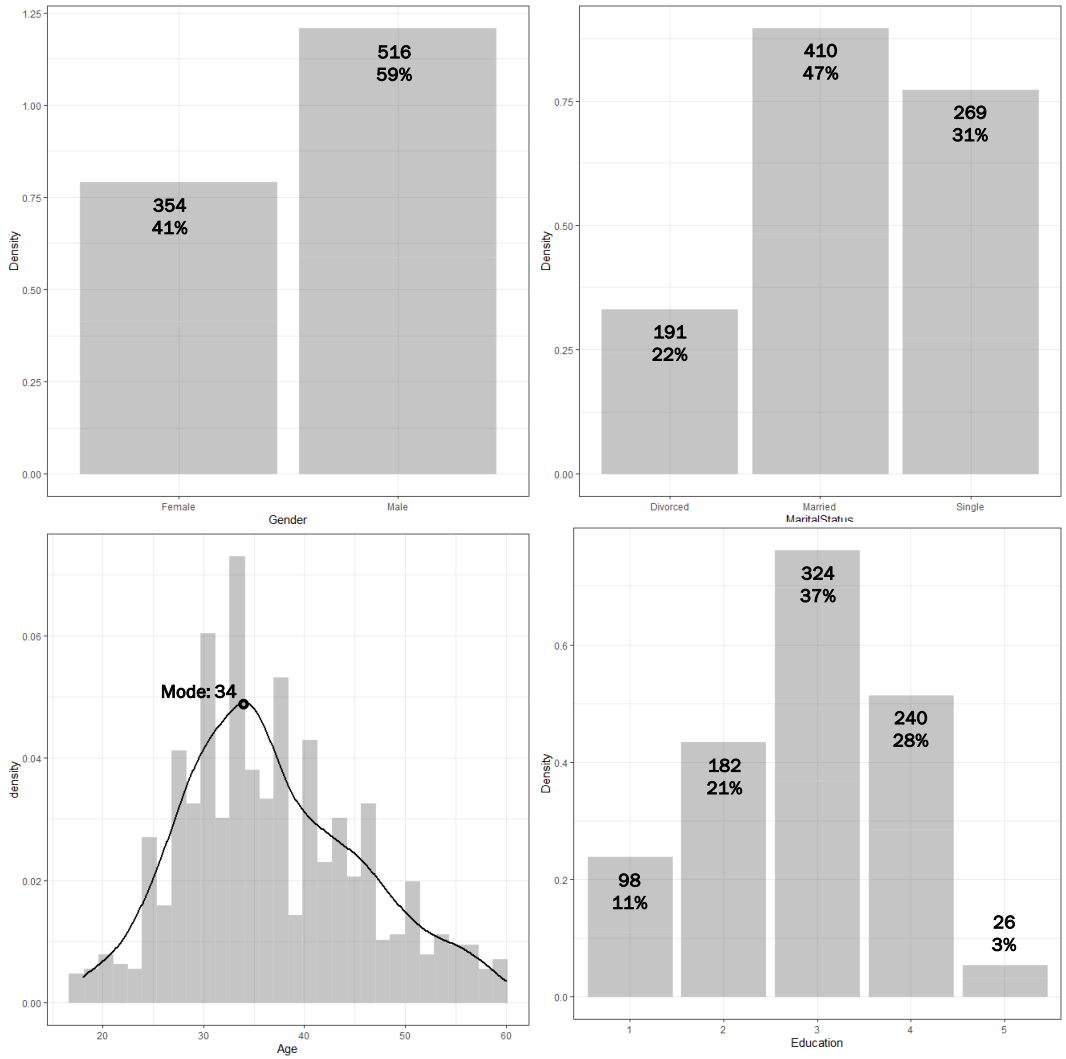
In 2019, the total quits rates for all industries was 27.9%, which has steadily increased since 2015, when it was 23.7%.[3]

The Company has an attrition rate (16%) largely in line with the broader US workforce.

[1,2] Mercer, "North American Employee Turnover: Trends and Effects."
[3] "Job Openings and Labor Turnover Survey News Release," U.S Bureau of Labor Statistics press release, March 17, 2020.

Company Employee-Base

General Demographics



Observations

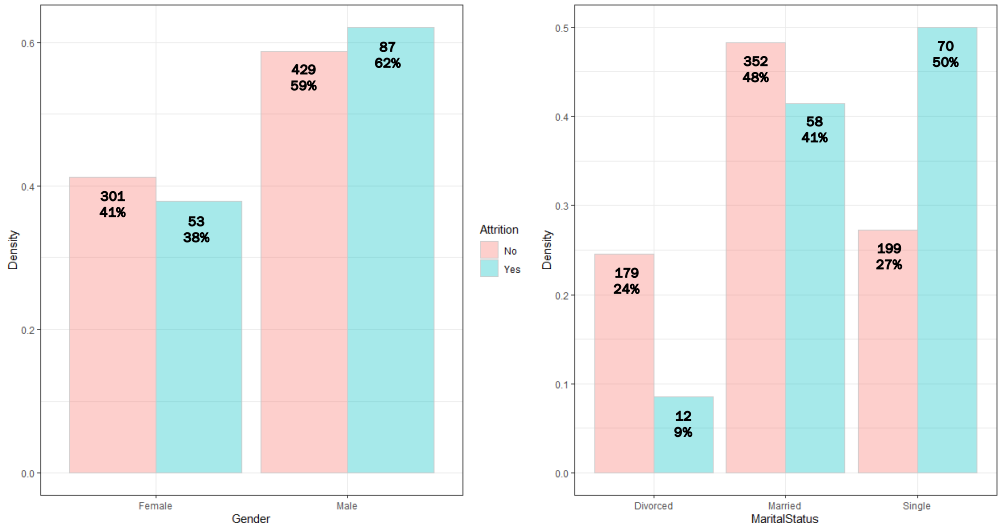
Generally, The Company has an employee-base dominated by individuals in their early to mid 30's, married and single , moderately-to-well educated and overweighted towards men.

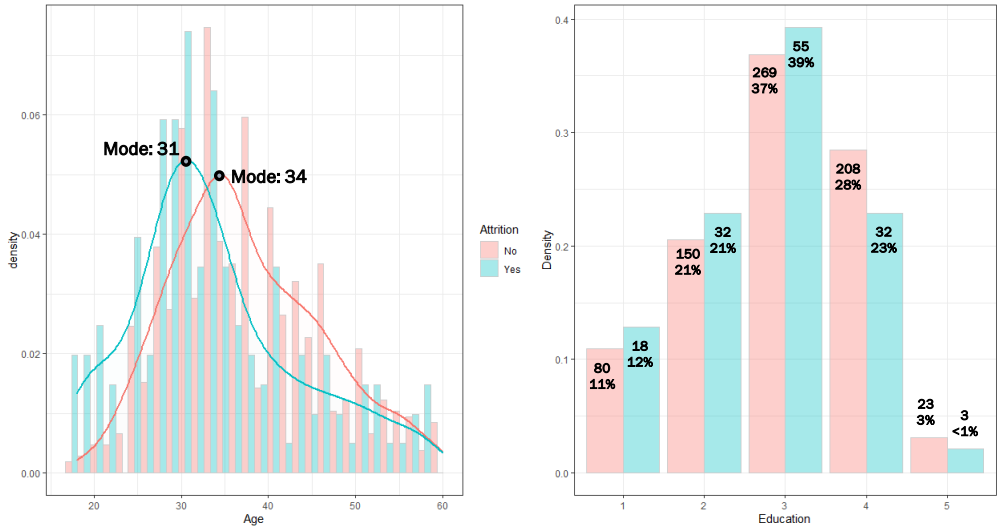
Employee base profiles such as these are naturally prone to attrition risk as the weights lie on demographics known for being in their most ambitious career advancement mindsets.

The advantage of these demographics are that they have enough experience to be delivering truly value-adding input along with a potential higher energy level than the older colleagues.

Company Employee-Base

General Demographics by Attrition Populations





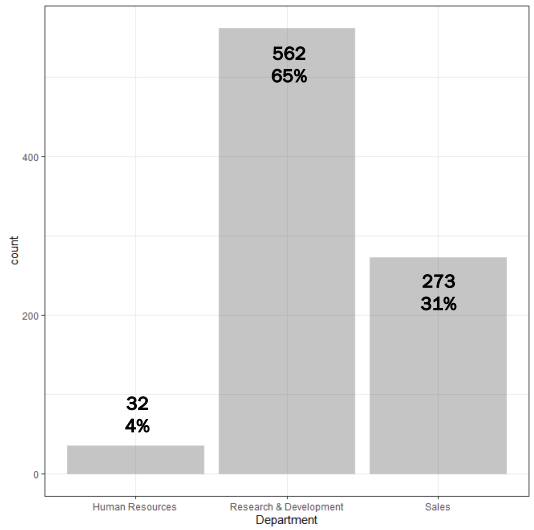
Observations

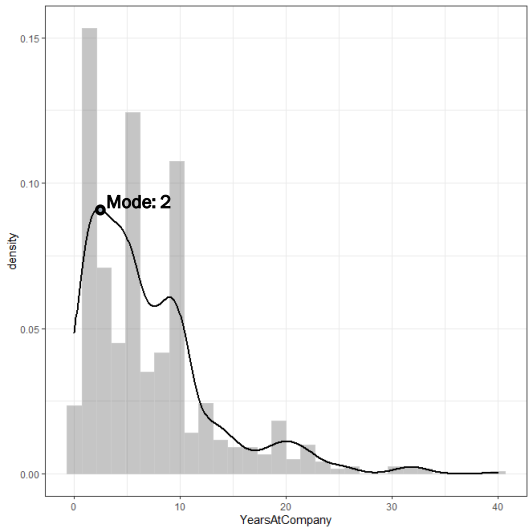
Consider the independent densities (relative proportion of data weighting) of the same demographic variables for those employee populations which has left versus those which have stayed.

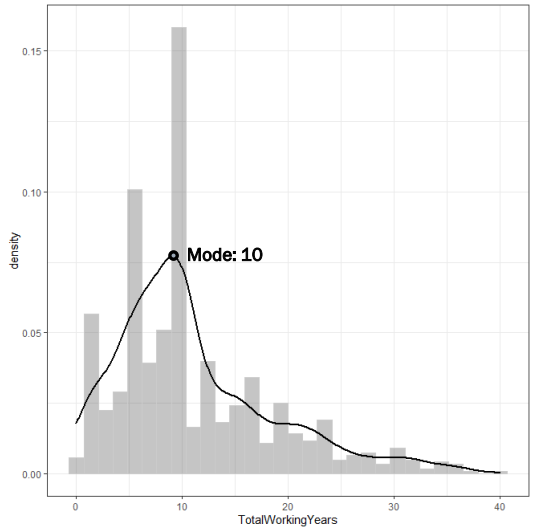
- Males have tended to leave at a higher proportion than Females
- Single employees dominate the population which has left
- Those employees which have left tended to be much younger (28-33 years)
- Those employees what have left tended to possess lower Education scores.

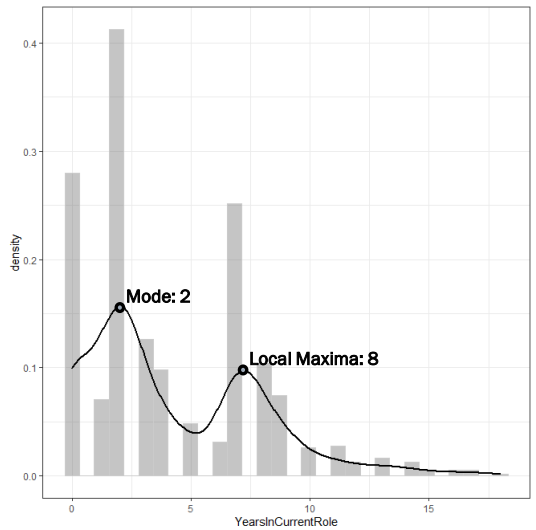
Company Employee-Base

Company Internal Demographics









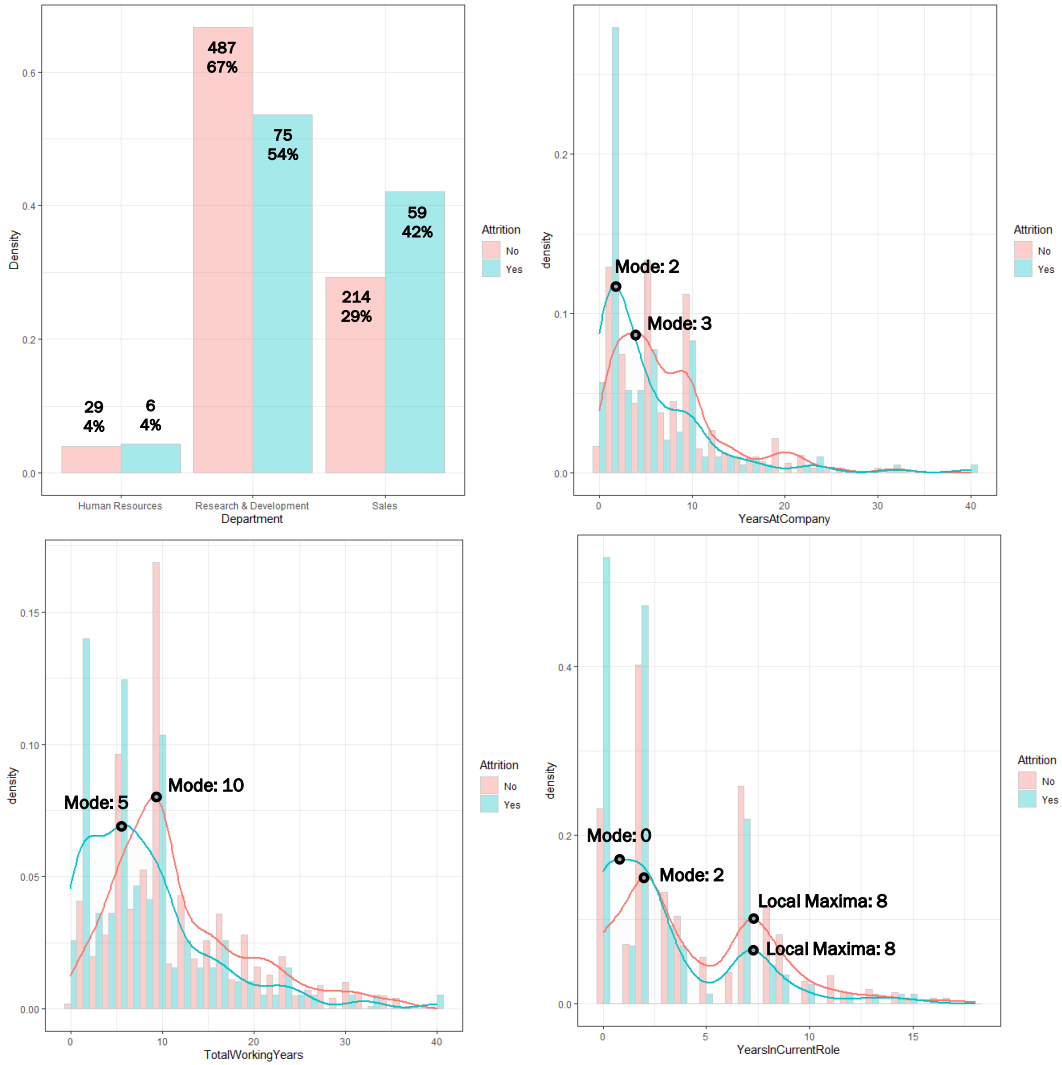
Observations

Now consider more Company internal demographic density trends.

- Employee headcount by department is dominated by the R&D group.
- Bulk of employee base has 2-12 years total working experience and 0-5 years with The Company.
- When considering the number of years an employee has spent in their current roles, two peaks emerge. The larger peak at year 2 is likely related to those employees newer to the company who still reside in the role to which they were hired.

Company Employee-Base

Company Internal Demographics by Attrition Populations



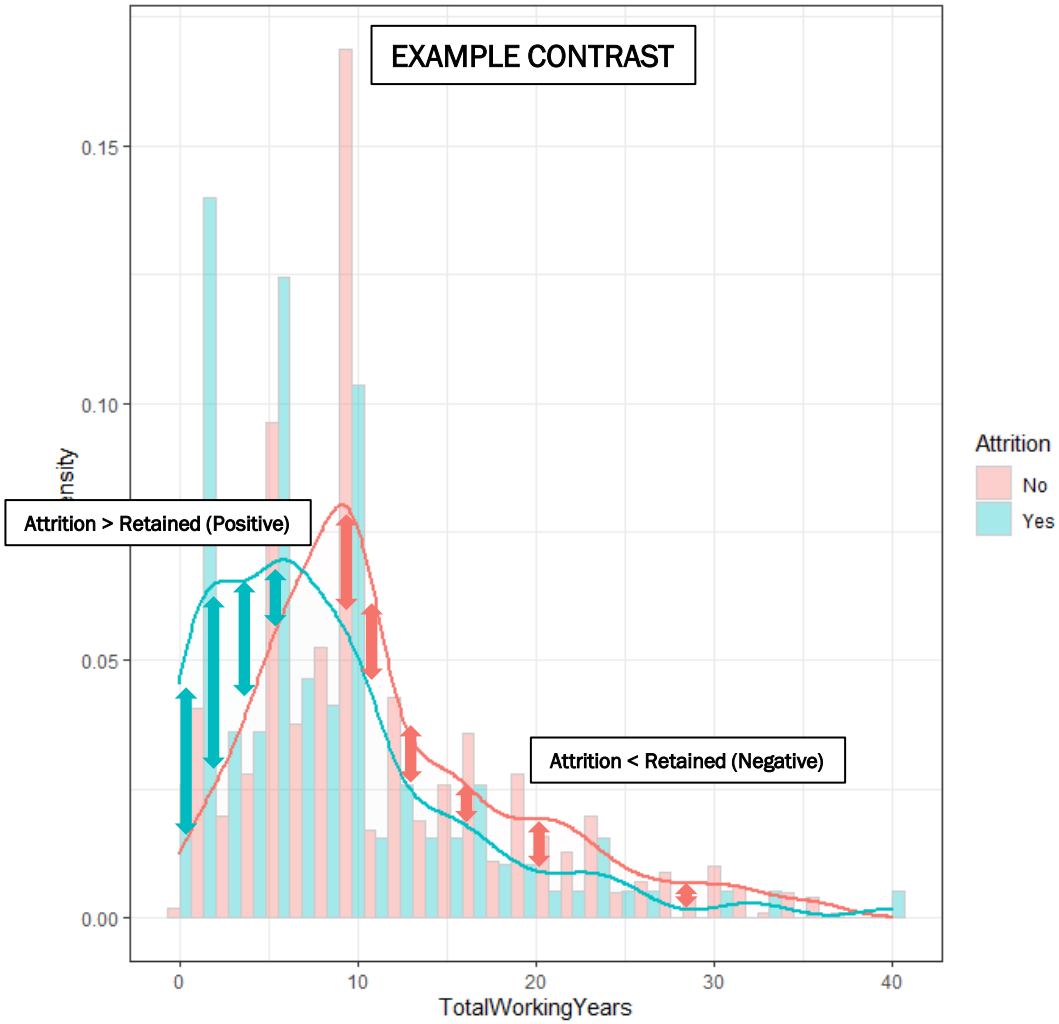
Observations

Now consider those same Company internal demographic density trends for the individual employee populations of those which have left versus those who have stayed.

- The Sales depart clearly accounts for the majority of employees who have voluntarily left the company.
- While younger employees, who by their very nature cannot posses many years of total experience nor with The Company, tend to leave at higher rates, we don't see a significant difference of attrition at any number of years of Company experience.
- Employee's attrition densities has a plateau from 0-5 years total experience before collapsing as individuals have presumably found good long-term fits with employers.

At-Risk Attrition Demographics

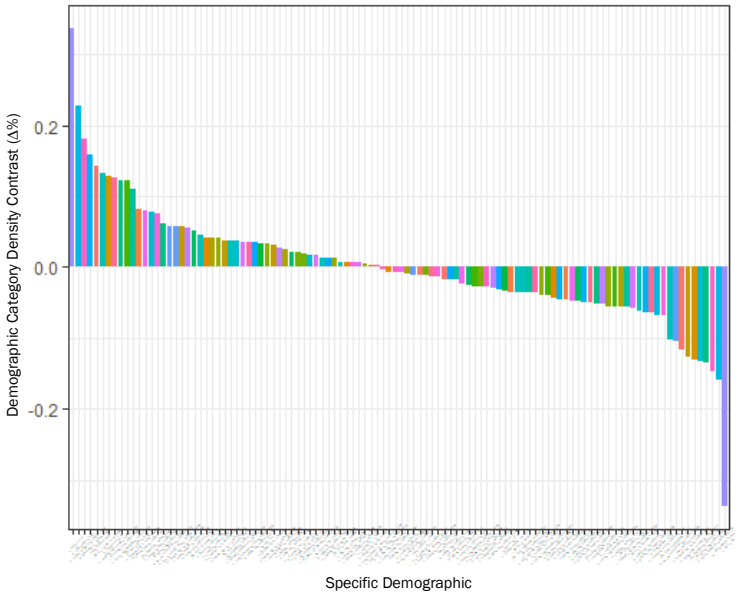
Ranking & Identification Strategy



Observations

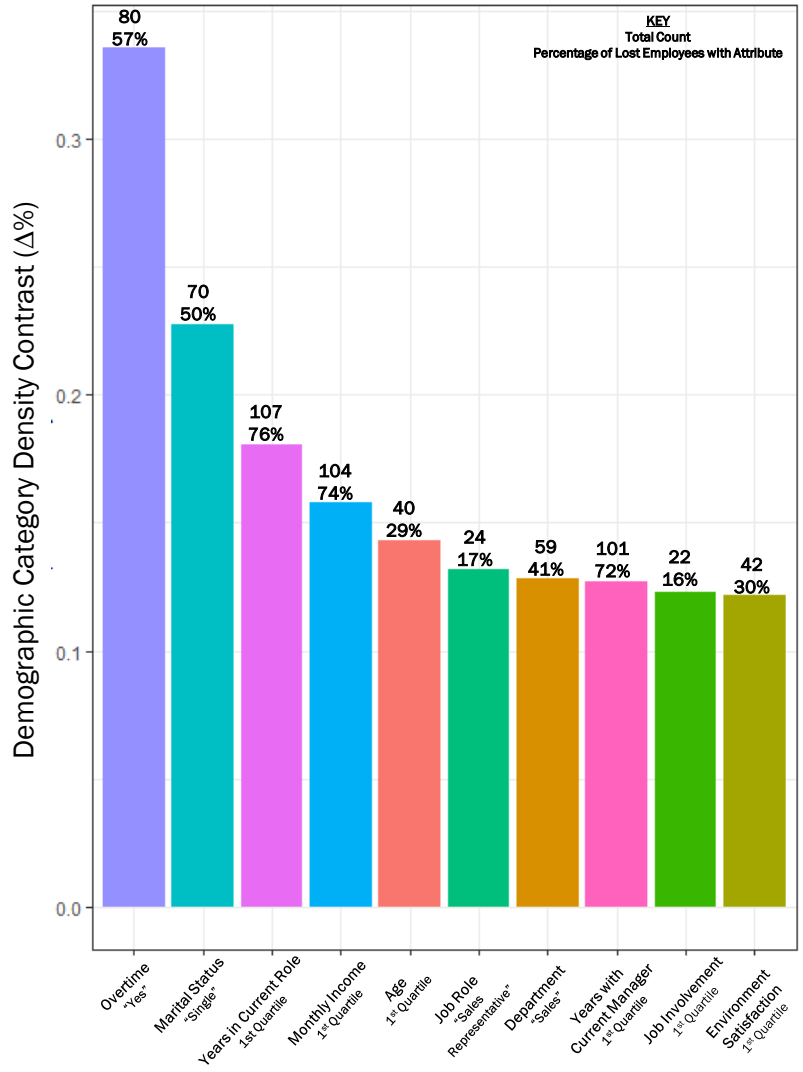
To aid identification of the most flight-risk employee demographics, we have analyzed the contrast between weights of employee between those populations of employees which have left versus those who have stayed.

- Continuous variables were bucketed into quartiles
- Positive values signify overweighting of the attrition group, negative values signify overweighting by remaining groups



At-Risk Attrition Demographics

Top 10 Most At-Risk Specific Employee Demographics



Observations

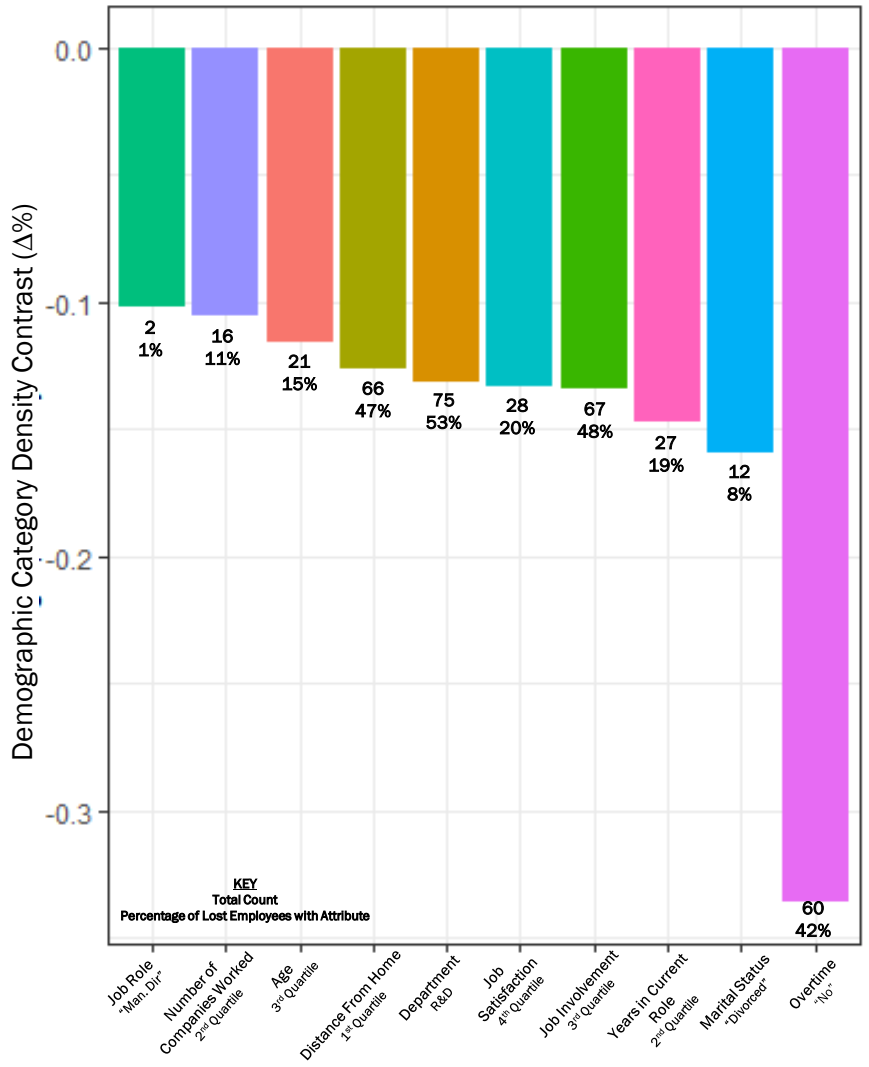
The top 10 most at-risk specific employee demographics are displayed at left. In other words, employees with these specific attributes see the highest positive contrast (delta) between density weightings of the attrition versus remaining populations.

- Most at-risks demographics in order of value of contrast:
  (1) Employees who work overtime
  (2) Single employees
  (3) Employees with (0 - 4.5) years in current role
  (4) Employees who make a monthly income ($0 – $5,811)
  (5) Employees with an age (18 – 28.5)
  (6) Employees in a "Sales Representative" role
  (7) Employees in the Sales Department
  (8) Employees with (0 – 4.25) years with their current manager
  (9) Employee who report a Job Involvement score of (1 – 1.75)
  (10) Employees who report an Environmental Satisfaction score of (0 – 1.75)

e – 1st Quartile: (0 - 4.5)
e – 1st Quartile: ($0 - 5,810.5)

Stable Attrition Demographics

Top 10 Most Stable Specific Employee Demographics



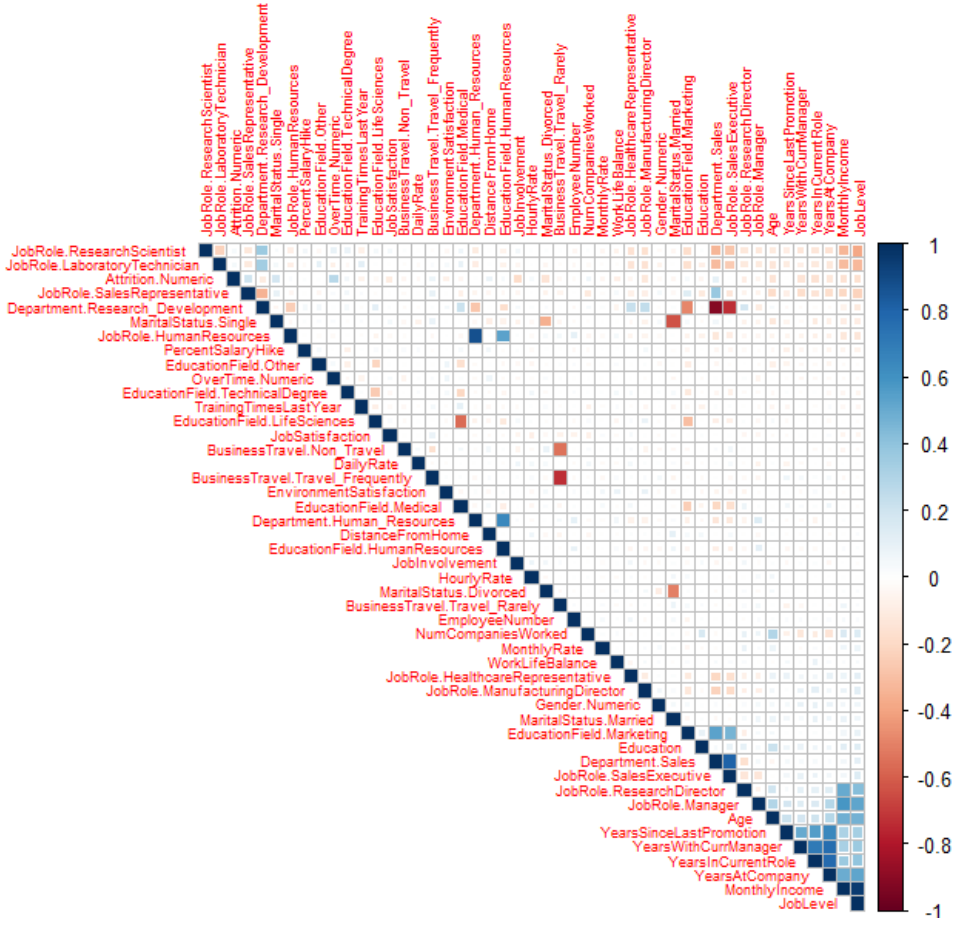
Observations

The top 10 most stable specific employee demographics are displayed at left. In other words, employees with these specific attributes see the highest negative contrast (delta) between density weightings of the attrition versus remaining populations.

- Most stable demographics in order of value of contrast:
  - (1) Employees who do not work overtime.
  - (2) Employees who are divorced.
  - (3) Employees who have been in their current roles (9 – 13.5) years
  - (4) Employees who report a Job Involvement score (2.5 – 3.25)
  - (5) Employees who report a Job Satisfaction score (3.25 – 4)
  - (6) Employees in the R&D department
  - (7) Employees who live (8 – 15) miles from the office
  - (8) Employees who are (39 – 49.5) years old
  - (9) Employees who have worked at (2 – 4) different companies
  - (10) Employees in a "Manufacturing Director" role

Variable Correlation

Correlation Matrix by Pearson Correlation Score



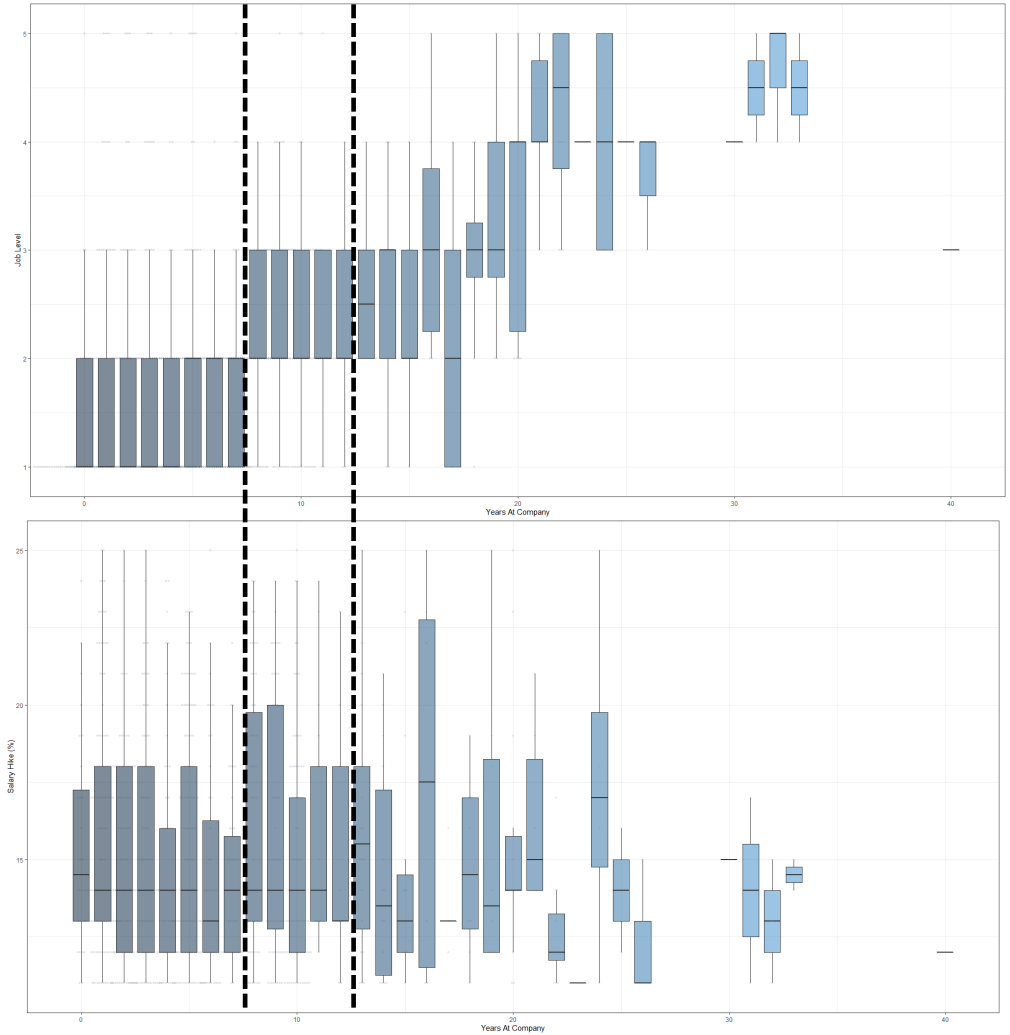
Observations

Correlations do exist between variables, but these are largely confined to those variables intuitively related to the age of an employee (see bottom right are of plot).

- Variables such as Job Level & Years at Company are inherently correlated, to a degree with age, and, in turn, these are correlated with Job Level & Monthly Income.

Surprisingly, very few other variables show strong, material correlations. The project team was much encouraged that our predictive and regression modeling efforts could utilize much of this data without fear of multilinearity effects.

Salary & Promotions



Observations

During the construction of the multiple linear regression model for Monthly Income (discussed at length later), input variables were studied in depth.

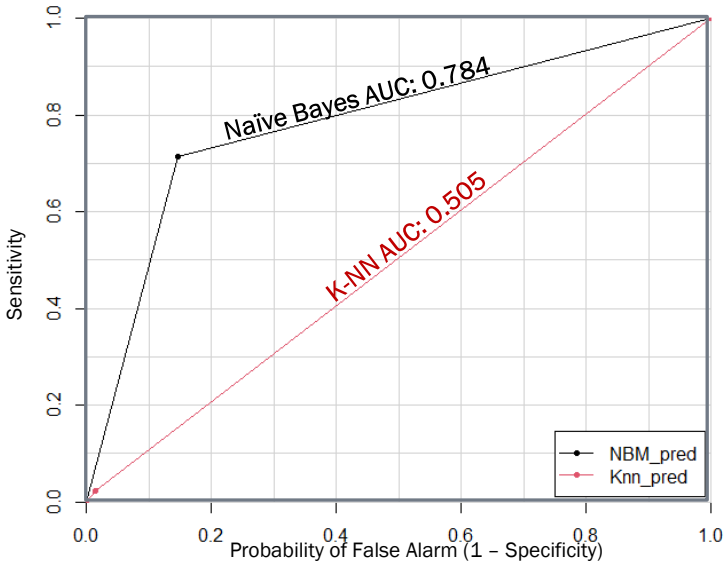
The boxplots shown at left, consider Job Level (top) and Salary Hike (bottom) versus Years at Company with dotted lines annotated to aid the visual tie between plots.

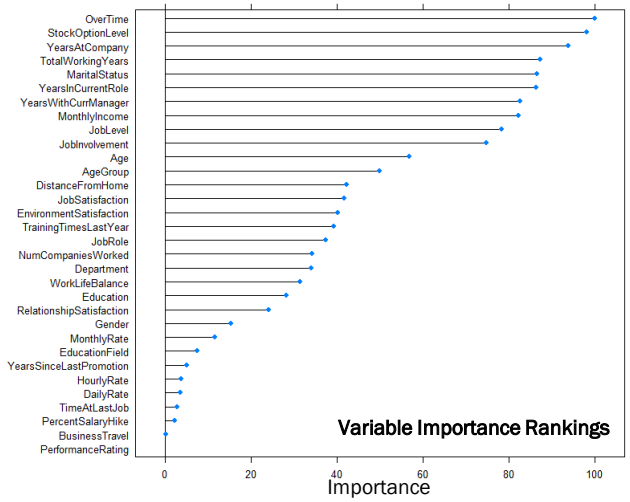
The research team found it quite interesting to observe the stability/perseverance of distribution positions and shapes for both variables until around year 12 of an employee's tenure at the Company.

These likely indicate internal structural mechanisms which control how and when an employee advances in rank and salary through the Company. If the Company is interested is preserving talent, they may reevaluate these internal controls and develop strategies to elevate and reward the younger at-risk demographics quicker.

Attrition Classification Model

Comparison & Performance





Observations

At the Company's request, a model which predicts if an employee will voluntarily leave the company, given values for the suite of variables we have been discussing, was constructed.

- All relevant variables provided by the client were utilized. Categorical variables were factored and quartiles calculated.

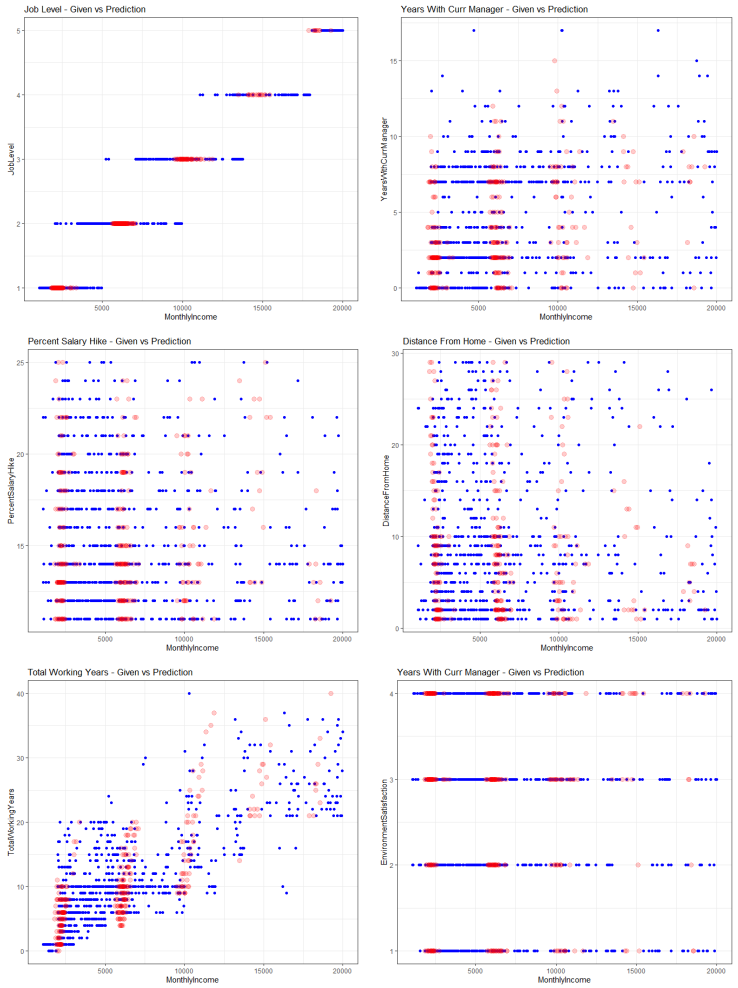
Naïve Bayes & k-NN classification strategies were compared using an AUC – ROC scheme.

- Naïve Bayes proved to be the much more powerful tool – see resulting variable importance rankings (bottom left).
    - At peak Specificity & Sensitivity, our model achieved an average accuracy of ~83%

- Naïve Bayes identified the same variables with the largest impact as our density contrast scheme (see earlier description) but now with Stock Option Level.

Monthly Income

Multiple Linear Regression Model



Observations

Also at the Company's request, a multiple linear regression model (MLR) was constructed to predict the monthly income of an employee given values for the suite of variables we have been discussing.

- Variables were first screened using a Chi-Squared test, then Forward/Backward/Stepwise selection routines were performed to further refine. The final variable set utilized by the model were manually selected using insights observed during the earlier EDA and simple intuition by the researchers – model equation shown below.
  - See further description and model performance statistics on the following slide.

- Cross-plots of the explanatory variables versus Monthly Salary are show at left where the blue points are from the training dataset provided and the red points are predicted values for the individuals for which we do not have Monthly Salary information.
  - The categorized Job Level variables proves to be the most impactful, important explanatory variables but also does introduce striping is not overcome by addition of the other variables.

$$MonthlyIncome = \beta_0 + \beta_1 Distance + \beta_2 JobLevel + \beta_3 PercSalaryHike + \beta_4 TotalWorkingYears + \beta_5 YearsWithCurrentManager$$

Monthly Income

Multiple Linear Regression Model

$$MonthlyIncome = \beta_0 + \beta_1 Distance + \beta_2 JobLevel + \beta_3 PercSalaryHike + \beta_4 TotalWorkingYears + \beta_5 YearsWithCurrentManager$$

MODEL PARAMETERIZATION

Residuals:

Min	1Q	Median	3Q	Max
-5759	-872	16	740	4035

Coefficients:

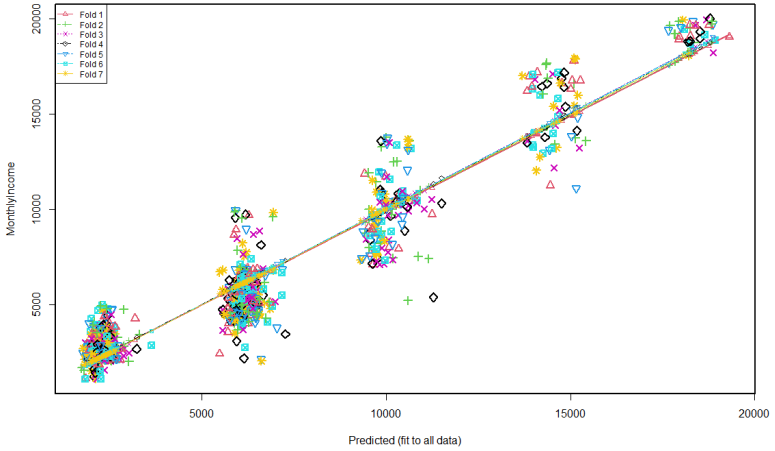
	Estimate	Std. Error	t value	Pr(>|t|)
β_0 (Intercept)	-1707.30	227.30	-7.51	1.5e-13
β_1 DistanceFromHome	-15.57	5.74	-2.71	0.0068
β_2 JobLevel	3723.77	68.43	54.41	< 2e-16
β_3 PercentSalaryHike	9.57	12.72	0.75	0.4519
β_4 TotalWorkingYears	68.12	10.41	6.54	1.0e-10
β_5 YearsWithCurrManager	-60.04	14.70	-4.09	4.8e-05

Residual standard error: 1370 on 864 degrees of freedom
Multiple R-squared: 0.911, Adjusted R-squared: 0.911
F-statistic: 1.78e+03 on 5 and 864 DF, p-value: <2e-16

INTERNAL CROSS-VALIDATION

Fold	Sum of Squares	Mean Square	RMSE	n
1	2.19E+08	1,766,391	1,329	124
2	3.45E+08	2,756,228	1,660	125
3	1.95E+08	1,558,045	1,248	125
4	2.37E+08	1,913,945	1,383	124
5	2.50E+08	2,019,304	1,421	124
6	1.89E+08	1,528,155	1,236	124
7	2.19E+08	1,768,845	1,330	124
Overall		**1,902,147**	**1,379**	

K-FOLD CROSS-VALIDATION: PREDICTED VS OBSERVED



K-FOLD CROSS-VALIDATION: PREDICTED VS RESIDUALS