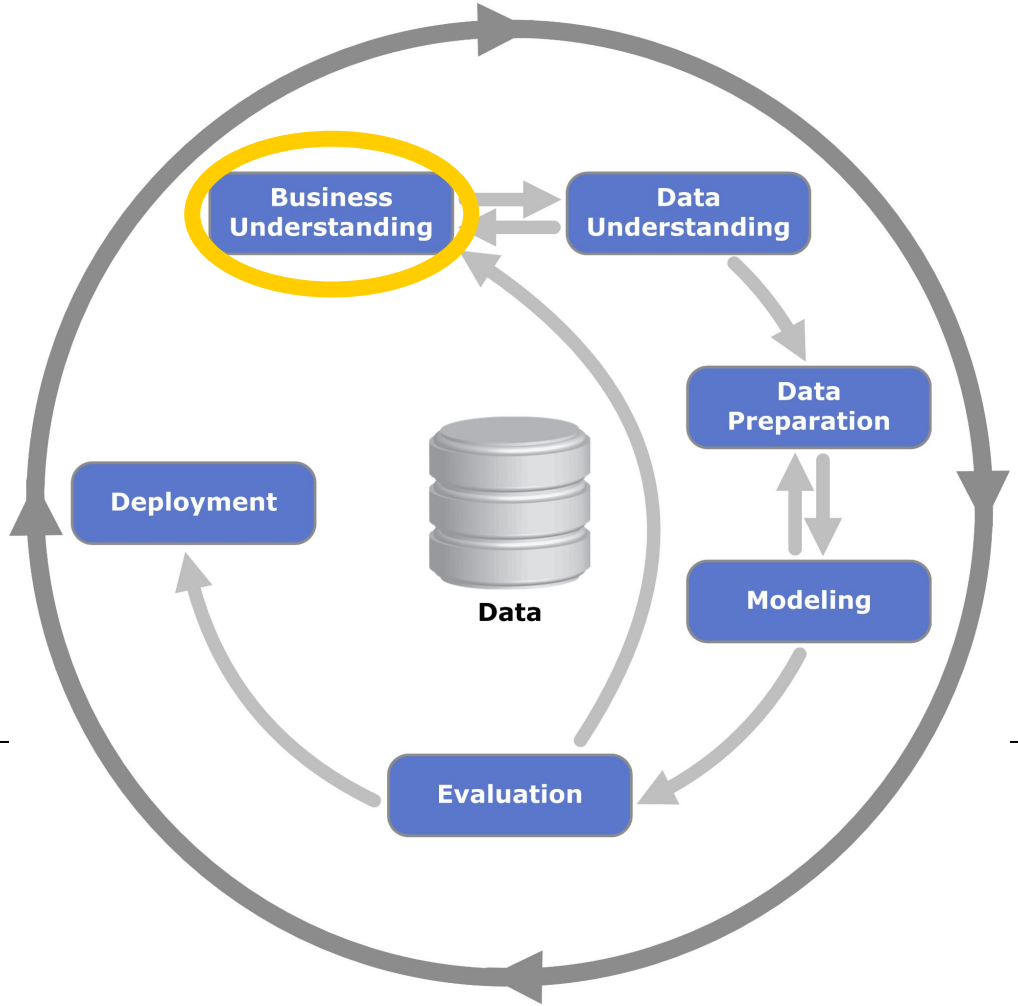
Predicting and Identifying Causes of Employee Turnover

Apurva Bhargava


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Why is this useful?

Employee turnover is the loss of talent in the workforce over time.

Here, we focus on **voluntary terminations**.

Why is predicting employee turnover important?

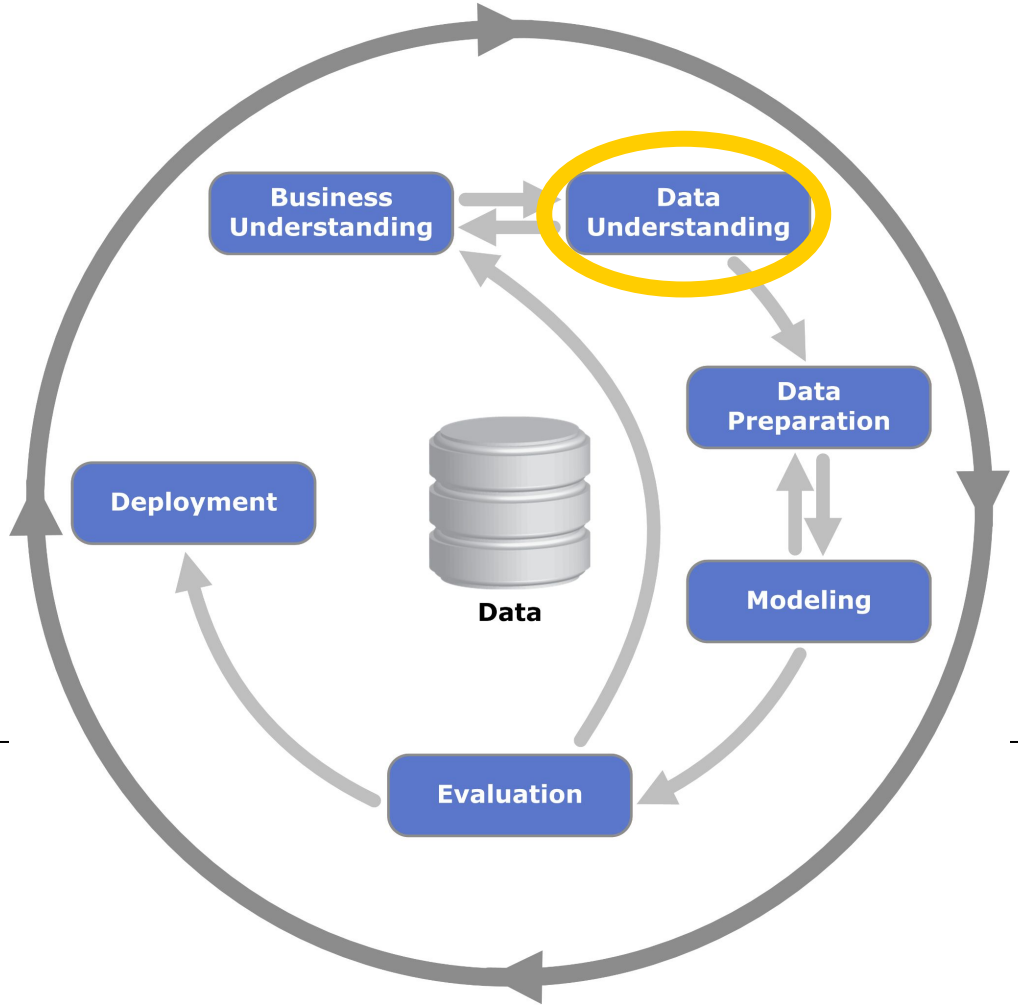
*According to a Gallup's 2015 Workforce Panel study, **51% of employees are actively looking for a new job** at any given time.

*The cost of recovery is around **20% of that position's salary in hiring, recruiting, and onboarding costs**. The hiring process takes **36 days on average**.

Prediction helps in improving **employee retention**.

Identifying causes helps in improving the **recruitment process**.

*Source: https://www.zenefits.com/workest/what-is-employee-turnover-and-why-it-matters/


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

What does our data look like?

Four tables:

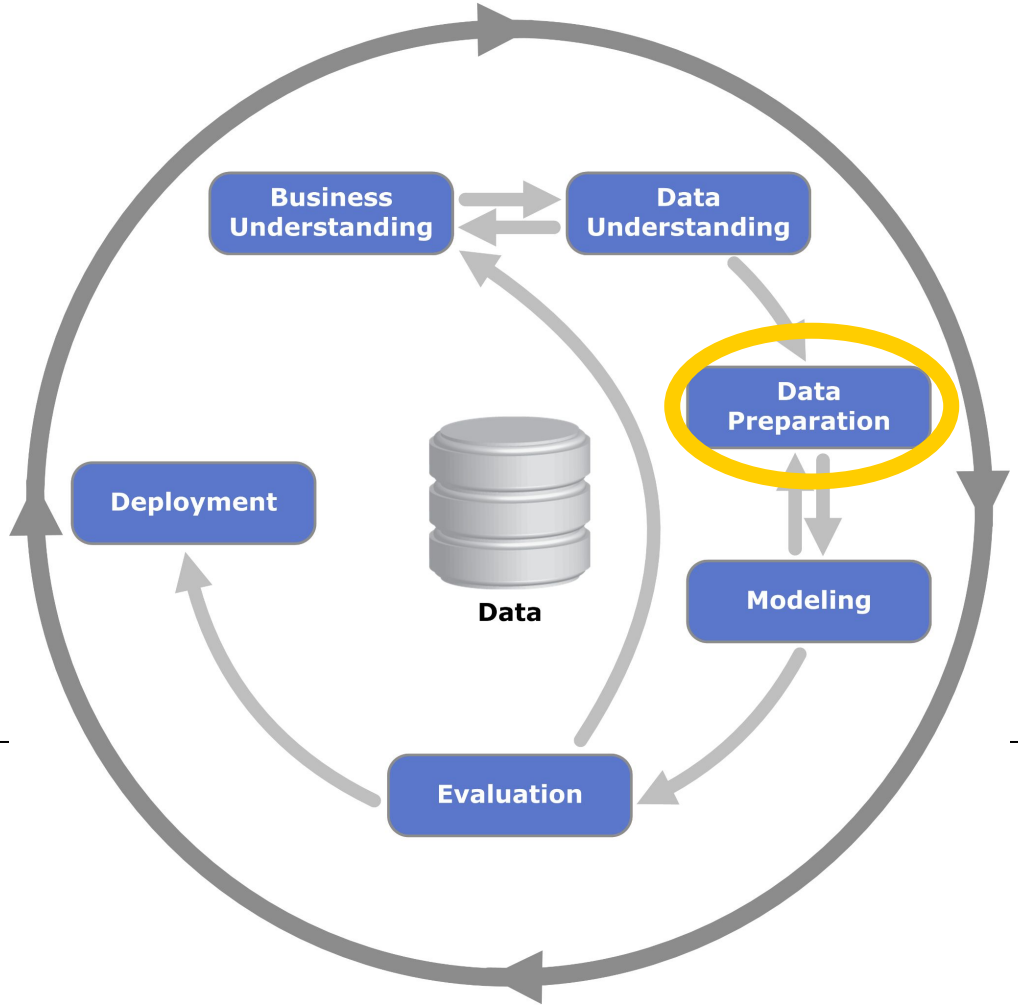
HRDataset_v9: employee age, gender, race, marital status, department, position, pay rate, employment status (active or left), reason for termination, source

salary_grid: hourly minimum, median, maximum pay rates for every position

production_staff: performance score, 90-day complaints, daily error rate

recruiting_cost: monthly and total costs of recruiting from a given source

*Source: Dataset from IFF (created by Dr. Rich Huebner and Dr. Carla Patalano for their graduate HRM course)


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Data Preparation

Joining Tables:

Employee database **HRDataset_v9 ⋈ production_staff**

on **‘Employee Name’**

(Employee Number was not common.)

Employee database **HRDataset_v9 ⋈ salary_grid**

on **‘Position’**

(Some positions were missing in salary_grid, so their summary statistics were calculated from the HR_Dataset_v9 itself.)

Data Preparation (continued)

Row and column selection:

- Employee records with non-voluntary terminations were dropped, and ‘leave of absence’ -> ‘active’ status. # of records dropped from 310 to 285.
- Redundant columns were removed by identifying one-on-one relationships using a self-written function.
- Manager Name, Location, DOB, Employee Name/ID, Dates were removed due to inability to use or make a generalization (small dataset).

Missing Value Imputation:

103 records had missing ‘90-day complaints’ and ‘daily error rate’.

These were filled in with 0, since it is the majority non-conflict value.

Data Preparation (continued)

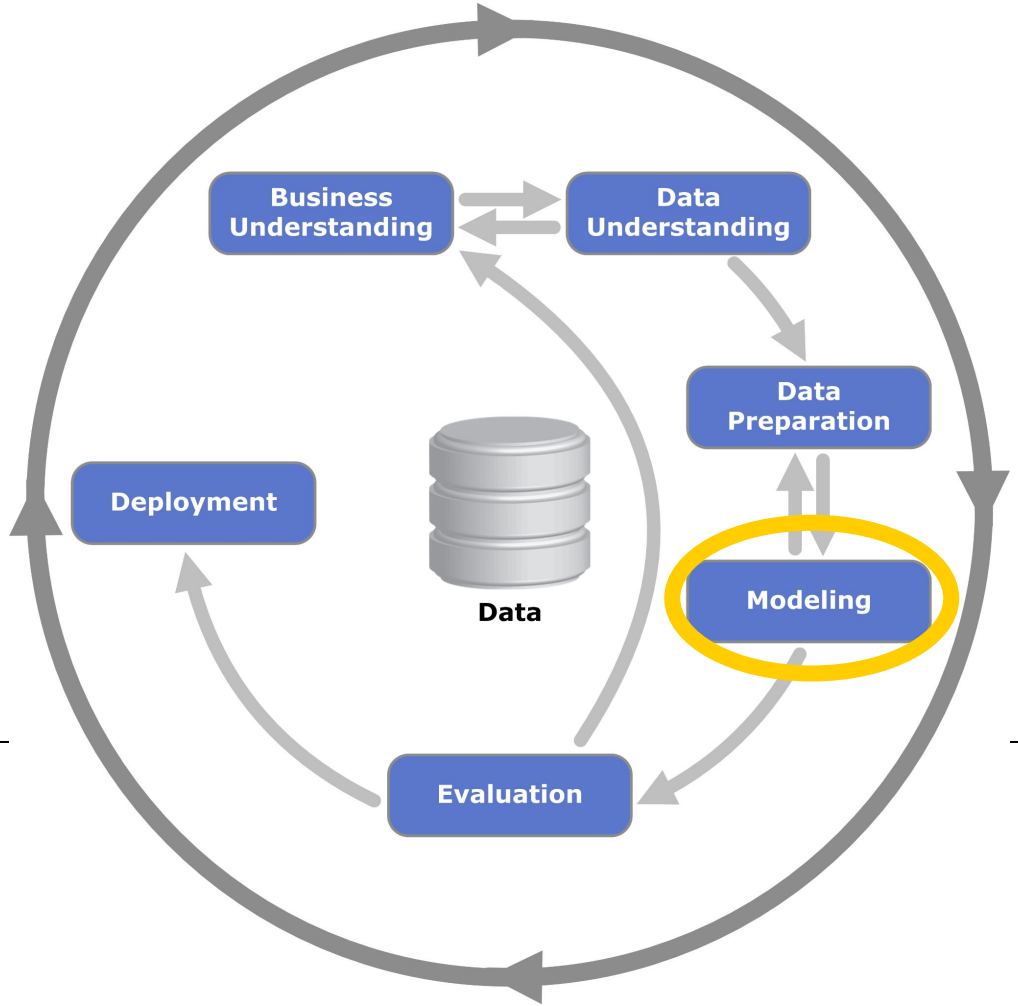
Feature Engineering:

- ‘Days employed’ -> ‘Years Employed’ to scale it down.
- Added ‘payRate>=mid’: is employee receiving more than the median pay for his/ her position type (also possible at gender/ department/ age level)
- Binary categorical variables were encoded as 0s/1s.

  All non-binary categorical variables were one-hot encoded.

  ‘Performance Score’ was coded as ordinal variable: In order: PIP, Needs Improvement, 90-day meets, N/A- too early to review/ Fully Meets, Exceeds, Exceptional.

  ‘Employment Status’ encoding: 'Active':0, 'Voluntarily Terminated':1.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Correlation

Top 39 correlations of various variables with ‘voluntary termination’ are given on the right.

Positive correlation implies that as the value of the variable increases, the chances of the employee voluntarily leaving the job increases.

Modeling

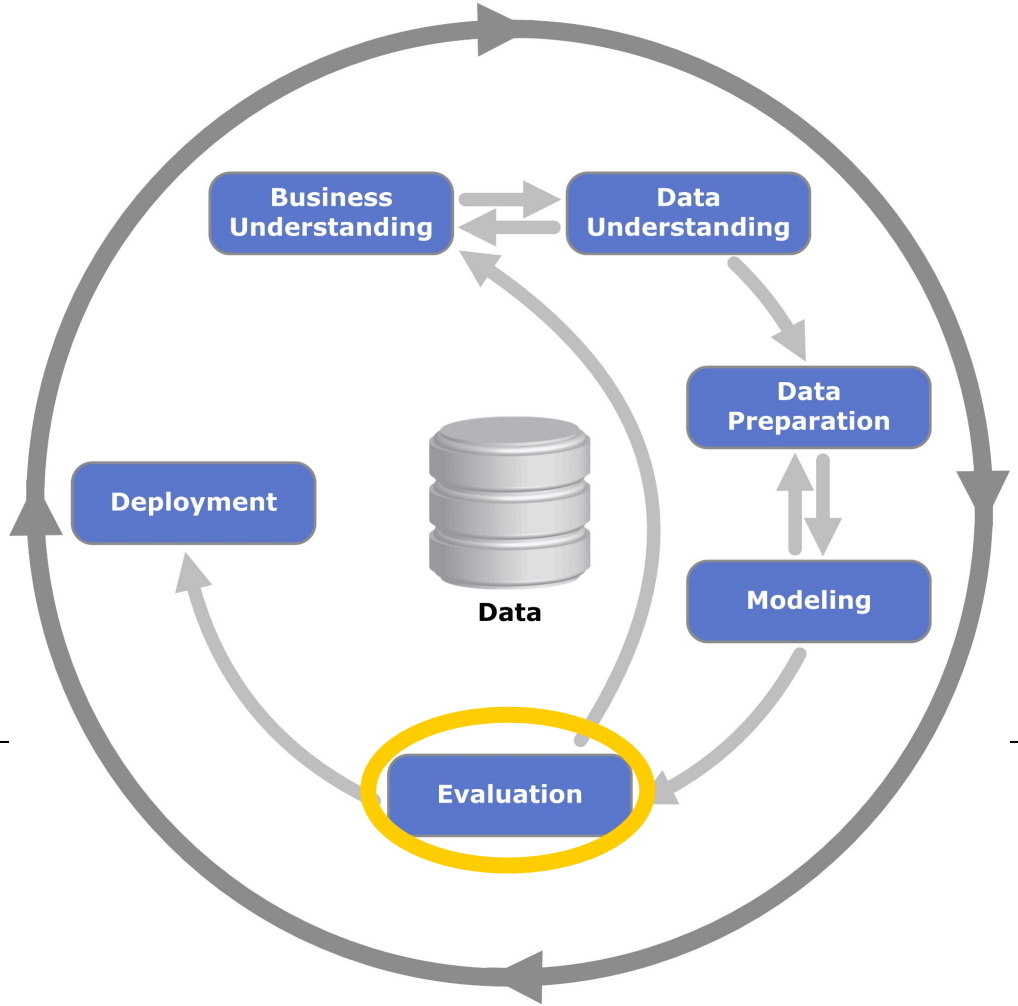
Objective:

Binary prediction (employment status: active or terminated) using 82 variables.

Models used:

Decision Tree: To identify combinations of variables that lead to a certain prediction in the form of nested if-then rules. **Feature informativeness is ranked on purity of split made on a variable.**

Logistic regression: A simpler linear model since decision tree is prone to overfitting (memorizing) small datasets; also ranks informativeness of every variable unlike decision tree. **Feature informativeness is ranked using absolute coefficients of variables.**


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Evaluation Methodology

72% train – 18% test split

161 active, 72 terminated in training data, 36 active, 16 terminated in test data.

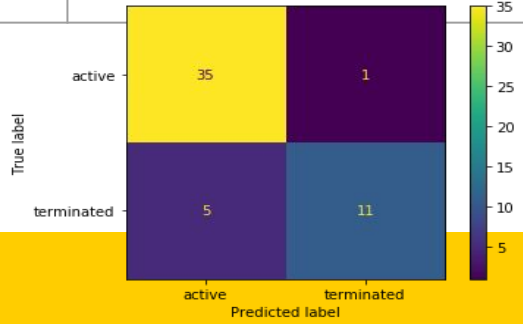
Best parameters were selected using brute grid search over all parameter combinations.

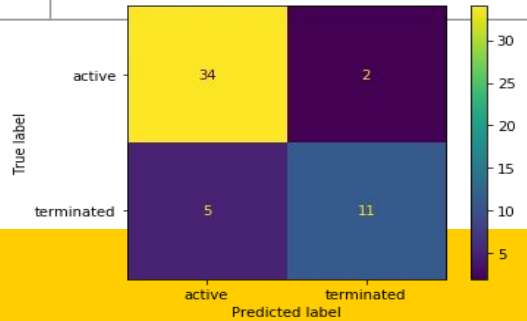
k-fold cross validation with k=5 was used for selecting best parameters (since dataset is small). AUC ROC metric used for grid search.

Thereafter, accuracies were computed on the training and test sets.

Results

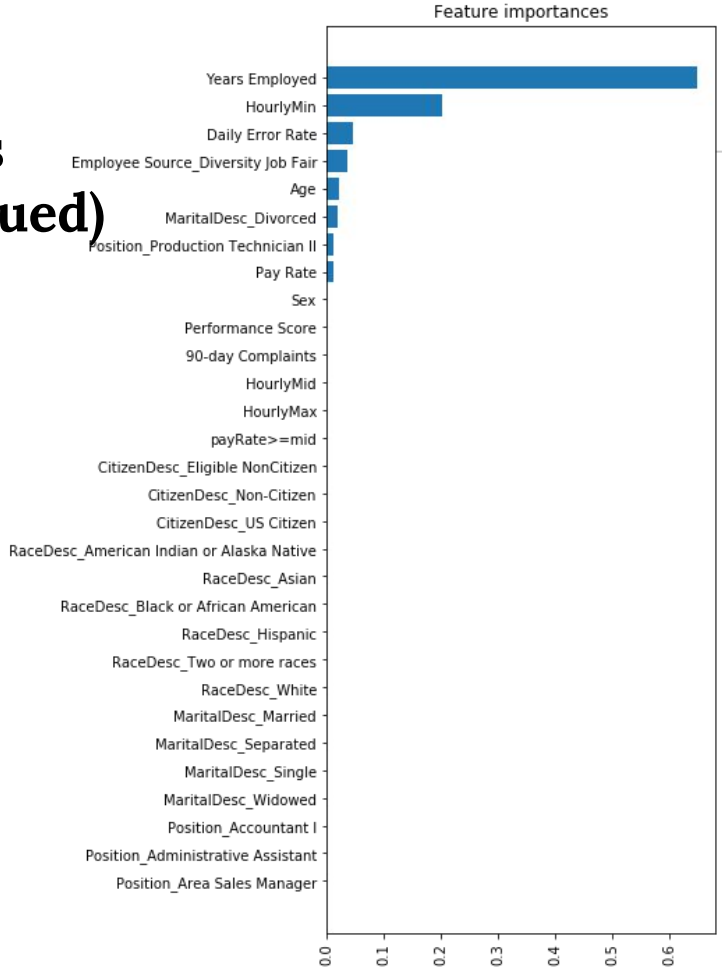
	Decision Tree	Logistic Regression
Parameters tuned	max_depth: range(1,15) min_samples_split: range(2,15) min_samples_leaf: range(1,15)	C: [1e-3, 1e-2, 1e-1, 1, 10] (C is inverse of regularization strength)
Best parameters	max_depth: 6 min_samples_split: 12 min_samples_leaf: 8	C: 0.1
Accuracy	Train: 87.12% | Test: 88.46%	Train: 90.55% | Test: 86.54%



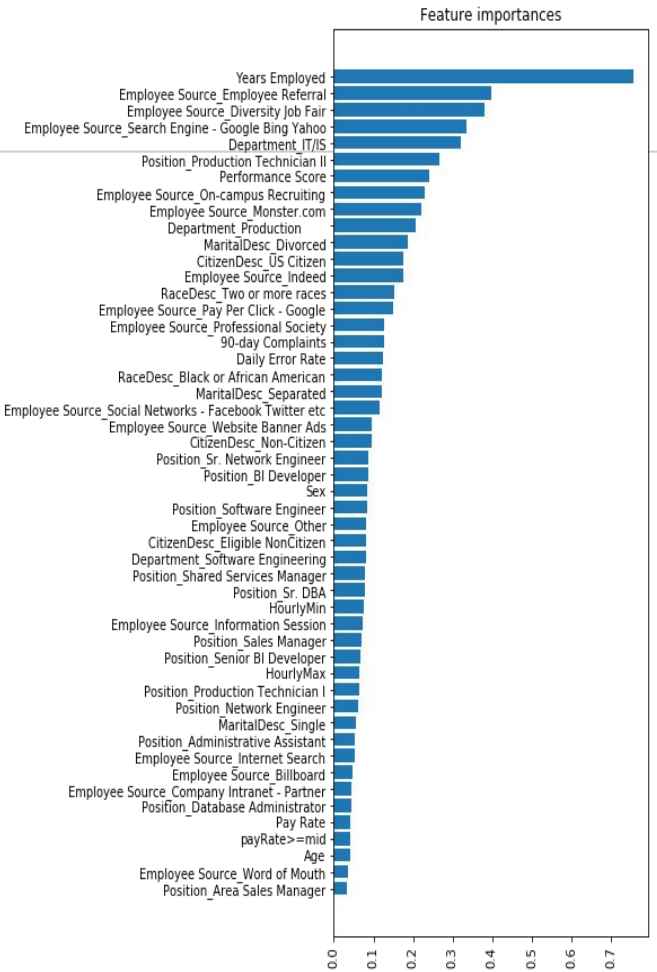


Results (continued)

Decision Tree

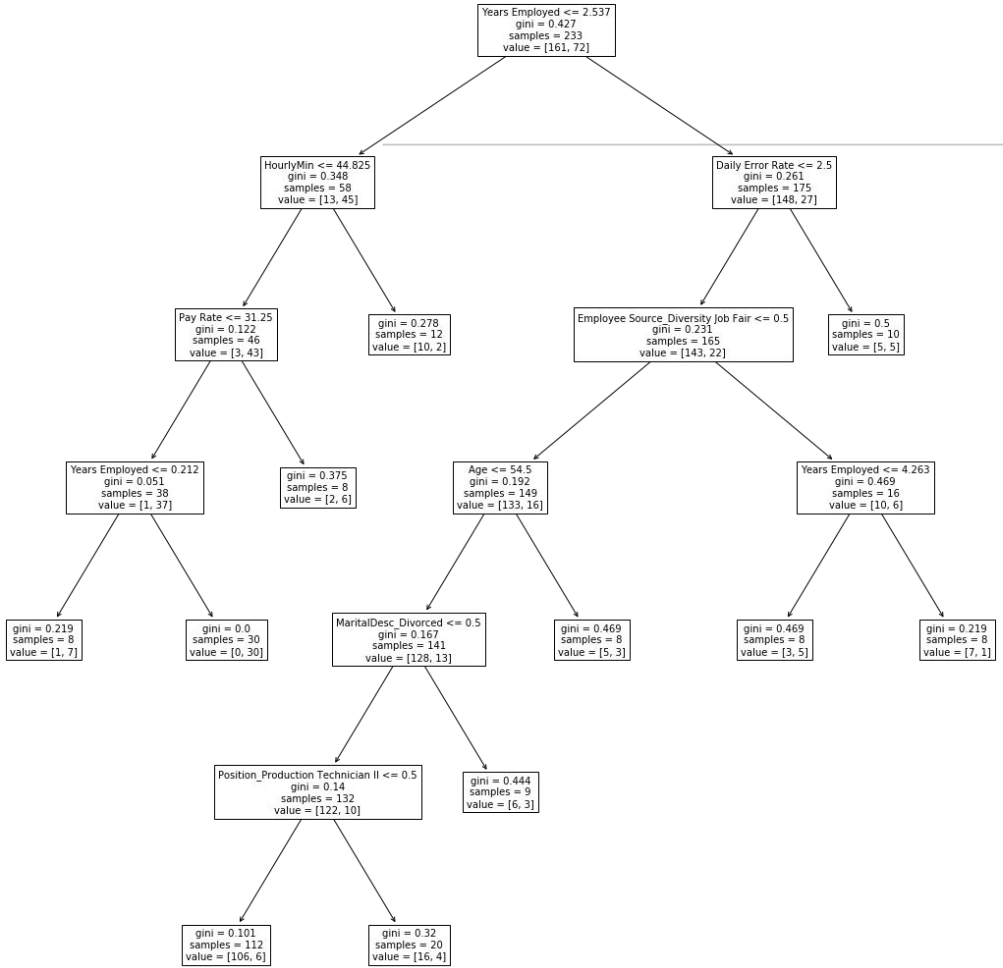


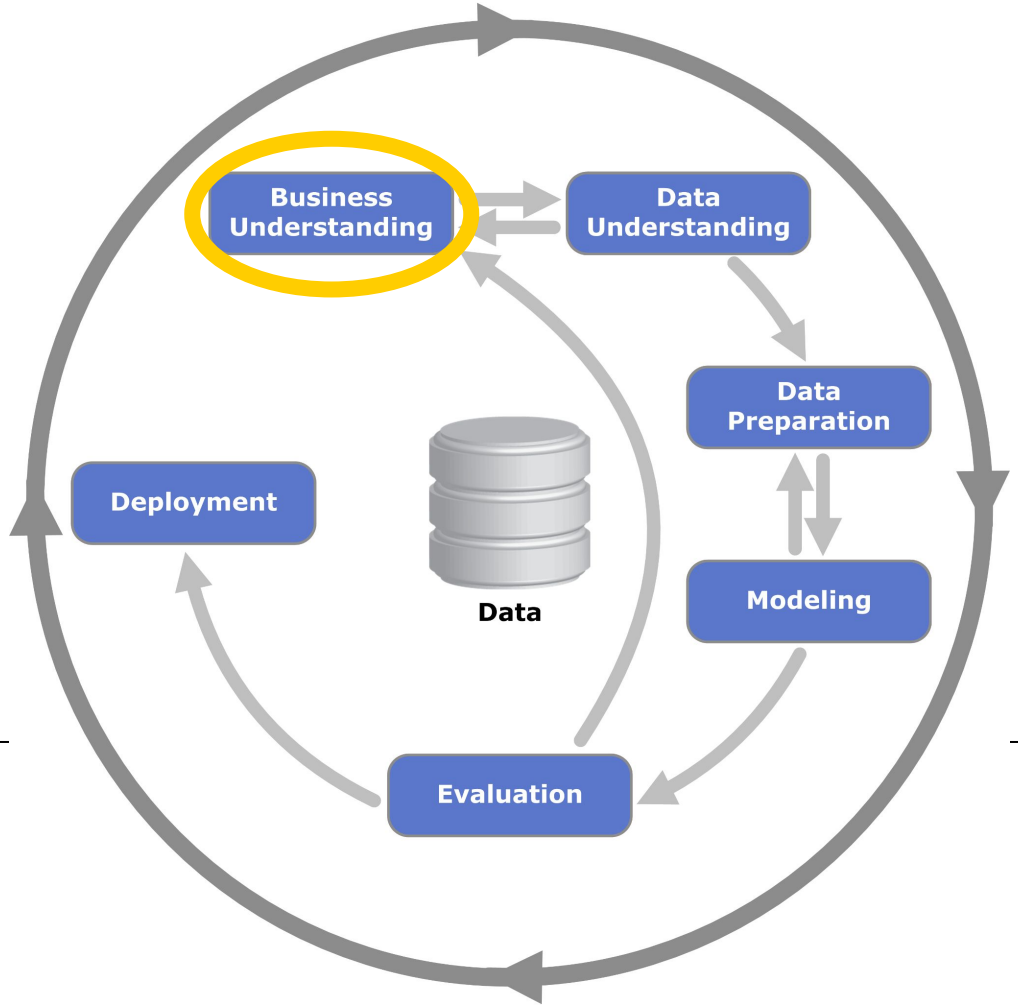
Logistic Regression



Results (continued)

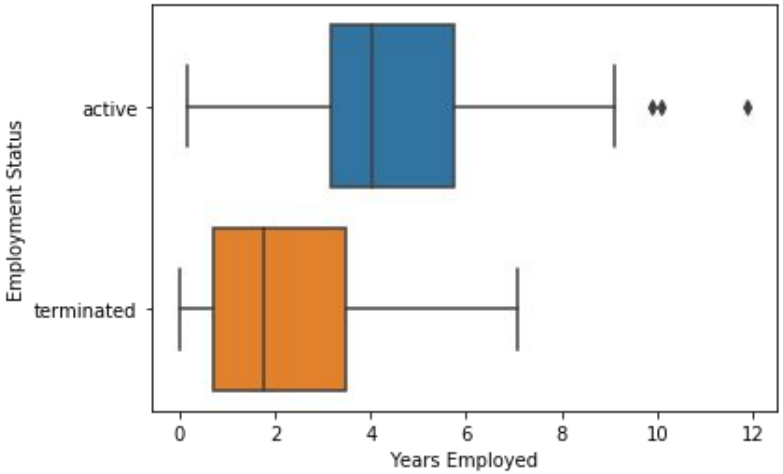
Decision Tree Visualized




Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

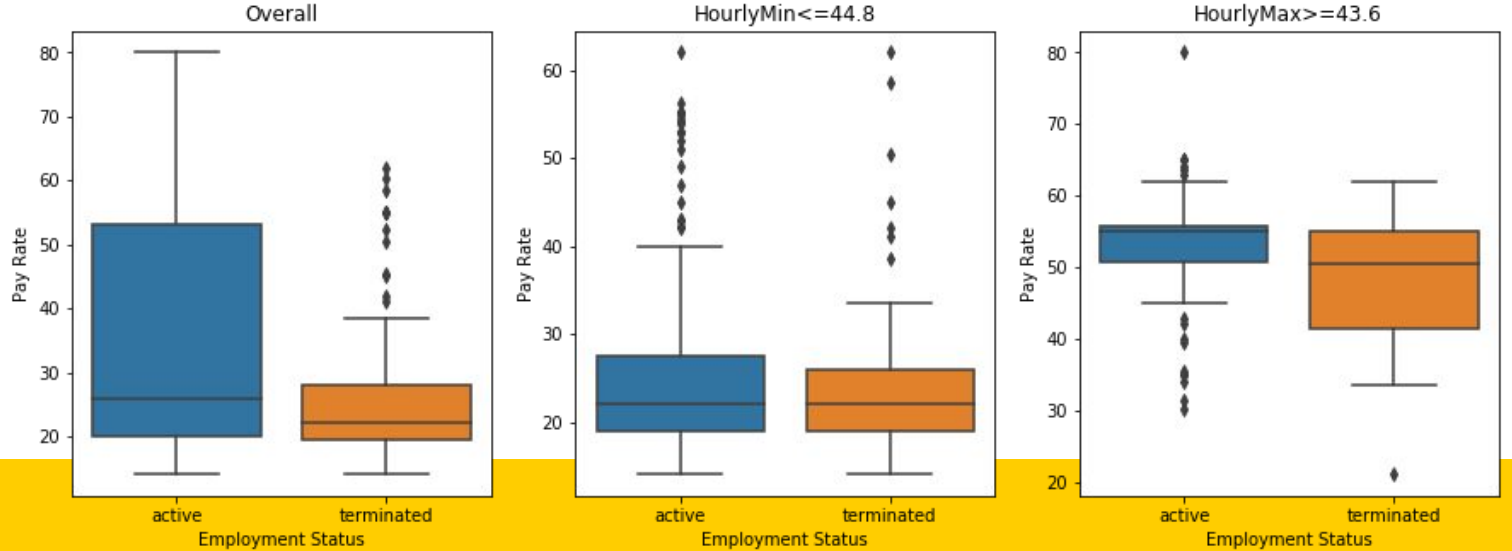
Insights (1)

Number of years employed is the biggest factor. Employees who have been working since many years are less likely to leave.



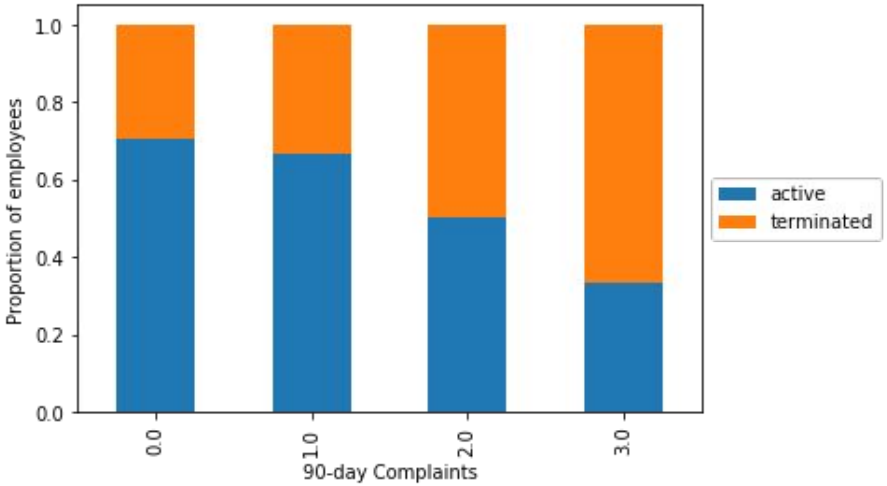
Insights (2)

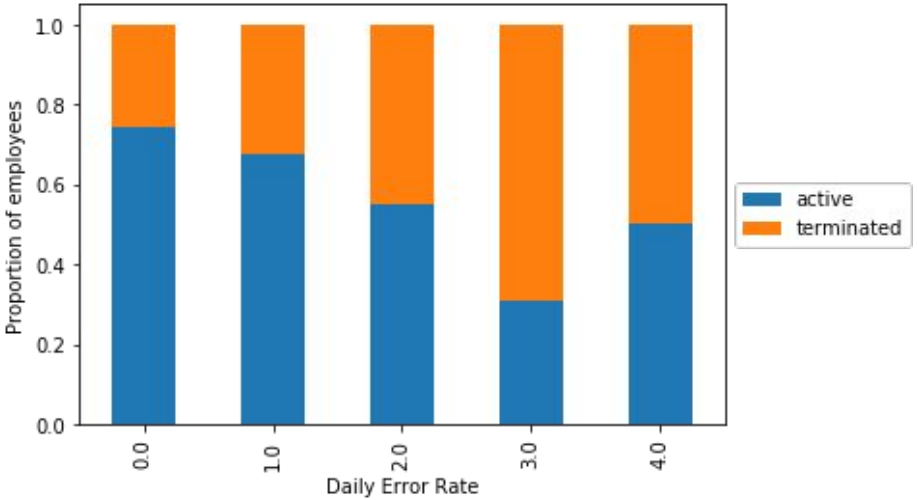
Employees with lower pay rate are likely to leave. From the decision tree, it can be observed that despite the individual pay rate, people working in positions with low hourly minimum are more likely to leave even if they have high individual pay rate, and people working in positions with high hourly maximum are more likely to stay even though they have a low individual pay rate.



Insights (3)

Poor performance is also a good indicator of employee departure. More terminations found in higher 90-day complaints and higher daily error rates.

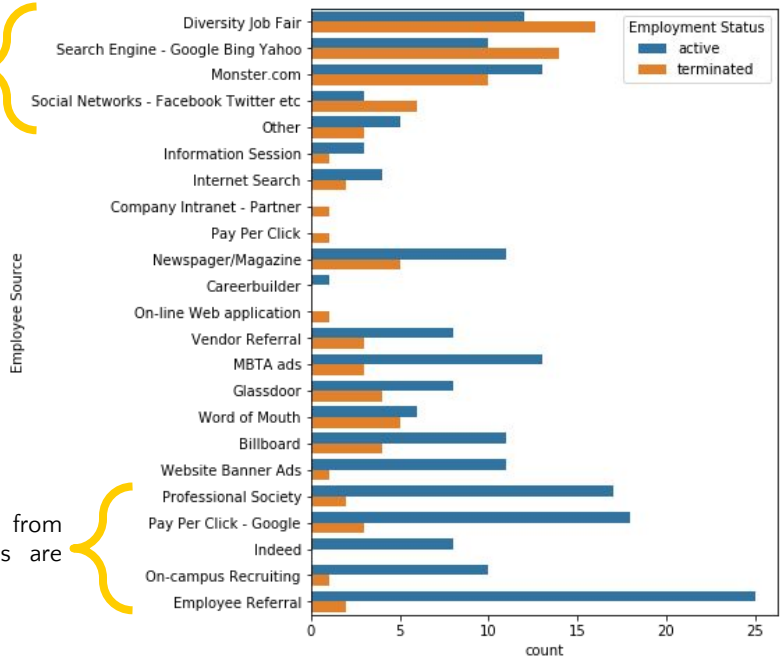




Insights (4)

Employees from certain sources are more likely to depart. The sources in the visualization are sorted by feature importances from logistic regression model.

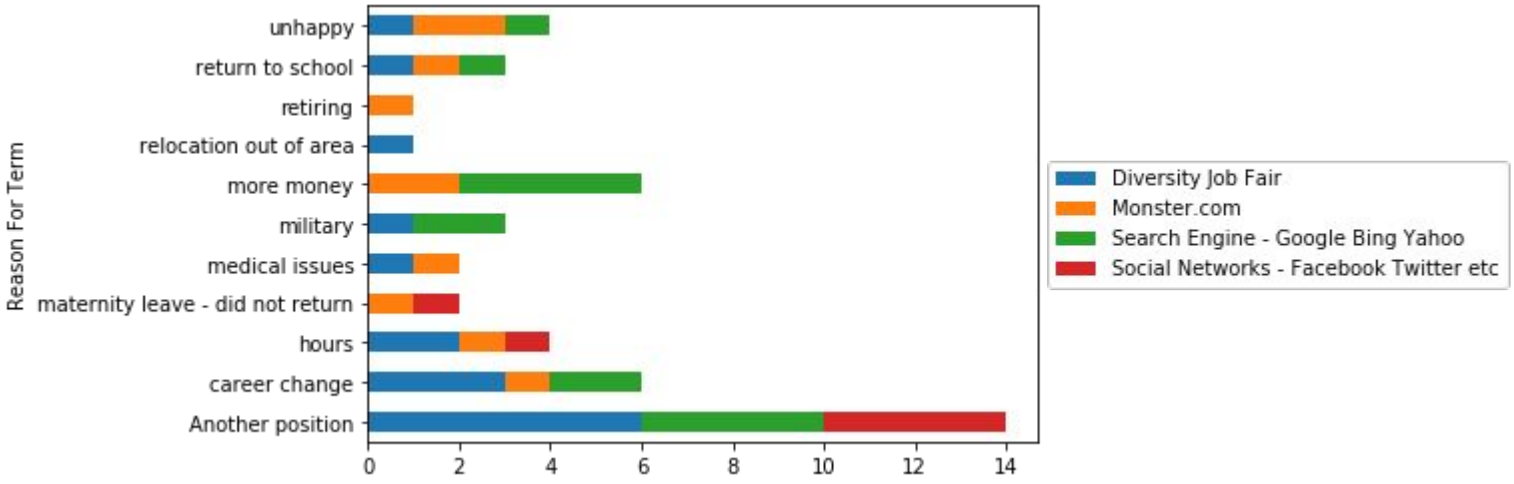
More employees from these sources have left than stayed. The counts are high, indicating some level of significance.



Employees from these sources are likely to stay.

Insights (4) (continued)

The reasons for terminations among 4 of the riskiest employee sources are given below. Employees for these sources mostly leave for other positions, career change or more salary.



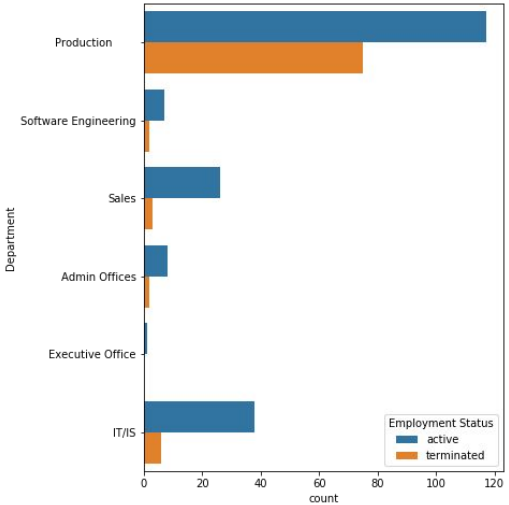
Insights (4) (continued)

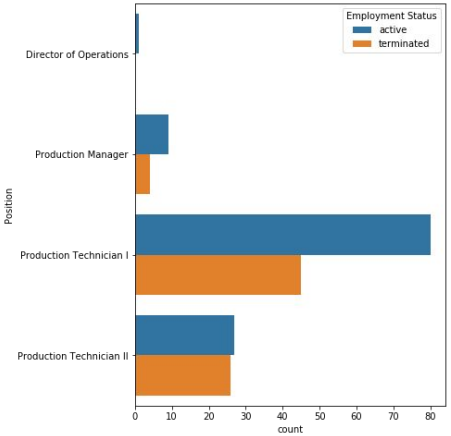
Effectiveness of source = percentage of employees retained / total recruitment costs
The table in increasing order of effectiveness is as follows:

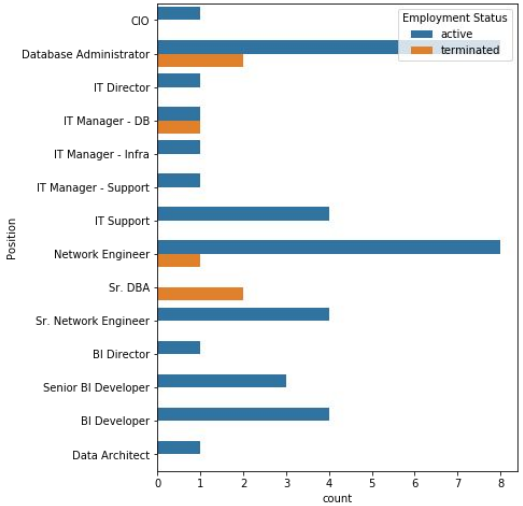
Pay Per Click	0.000000
Diversity Job Fair	0.000043
Social Networks - Facebook Twitter etc	0.000060
MBTA ads	0.000074
Search Engine - Google Bing Yahoo	0.000080
Newspager/Magazine	0.000083
Monster.com	0.000098
Billboard	0.000118
On-campus Recruiting	0.000121
Website Banner Ads	0.000128
Careerbuilder	0.000128
Other	0.000156
Pay Per Click - Google	0.000244
Professional Society	0.000746

Insights (5)

It was observed that among the departments, ‘Production’ sees the most voluntary terminations and ‘IT/IS’ sees the least (as compared to active). The departments in the visualization are sorted by feature importances from logistic regression model.





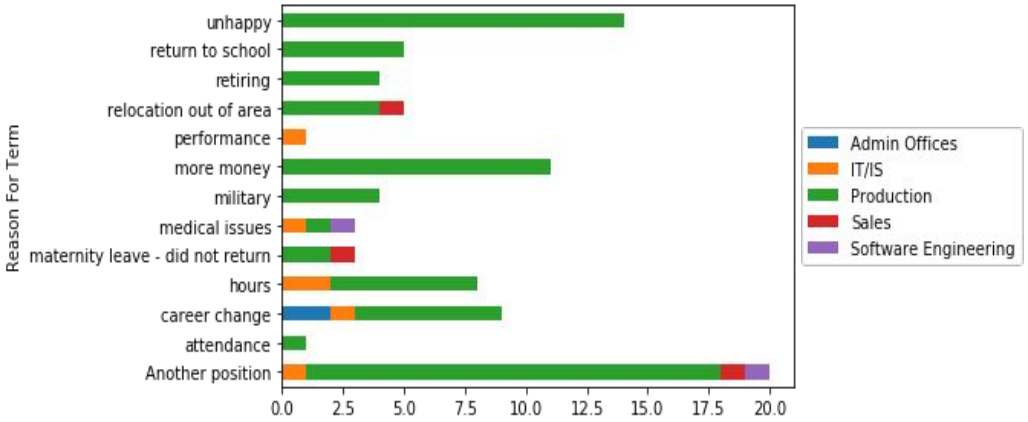


Production department

IT/IS department

Insights (5) (continued)

The reasons for terminations among all the departments are plotted below. The table shows the department-wise statistics of hourly pay rates, where it can be seen that production department is indeed paid significantly less than other departments.



Department	min	mean	median	max
Production	14.00	23.153385	22.00	60.00
Admin Offices	16.56	31.896000	28.75	55.00
IT/IS	21.00	45.104091	45.00	65.00
Software Engineering	27.00	48.683333	49.25	57.12
Sales	54.00	55.560345	55.00	60.25
Executive Office	80.00	80.000000	80.00	80.00

Conclusions

Overall, years of employment, performance and pay rates are significant but obvious factors. It is notable that working in positions with generally higher pay somehow plays into retention (through either expectation or prestige).

Employee source is also a good indicator of employee turnover. It can be utilised directly by evaluating credibility of the source or indirectly by generalizing the kind of employees that come from a particular type of source.

While ‘Department’ may be a proxy for other factors and ‘Production’ is over-represented in the data, analysing it allows us to pinpoint areas of workforce loss and estimate the cost of its recovery.

Decision tree can be used to analyse the effect of more complex relationships between variables on the probability of employee turnover, with more data.

Thank you