Why do employees voluntarily leave their job?

Apurva Bhargava

Time taken: 3.5 hours for writing code, 1 hour for writing report

1. **Preprocessing:**

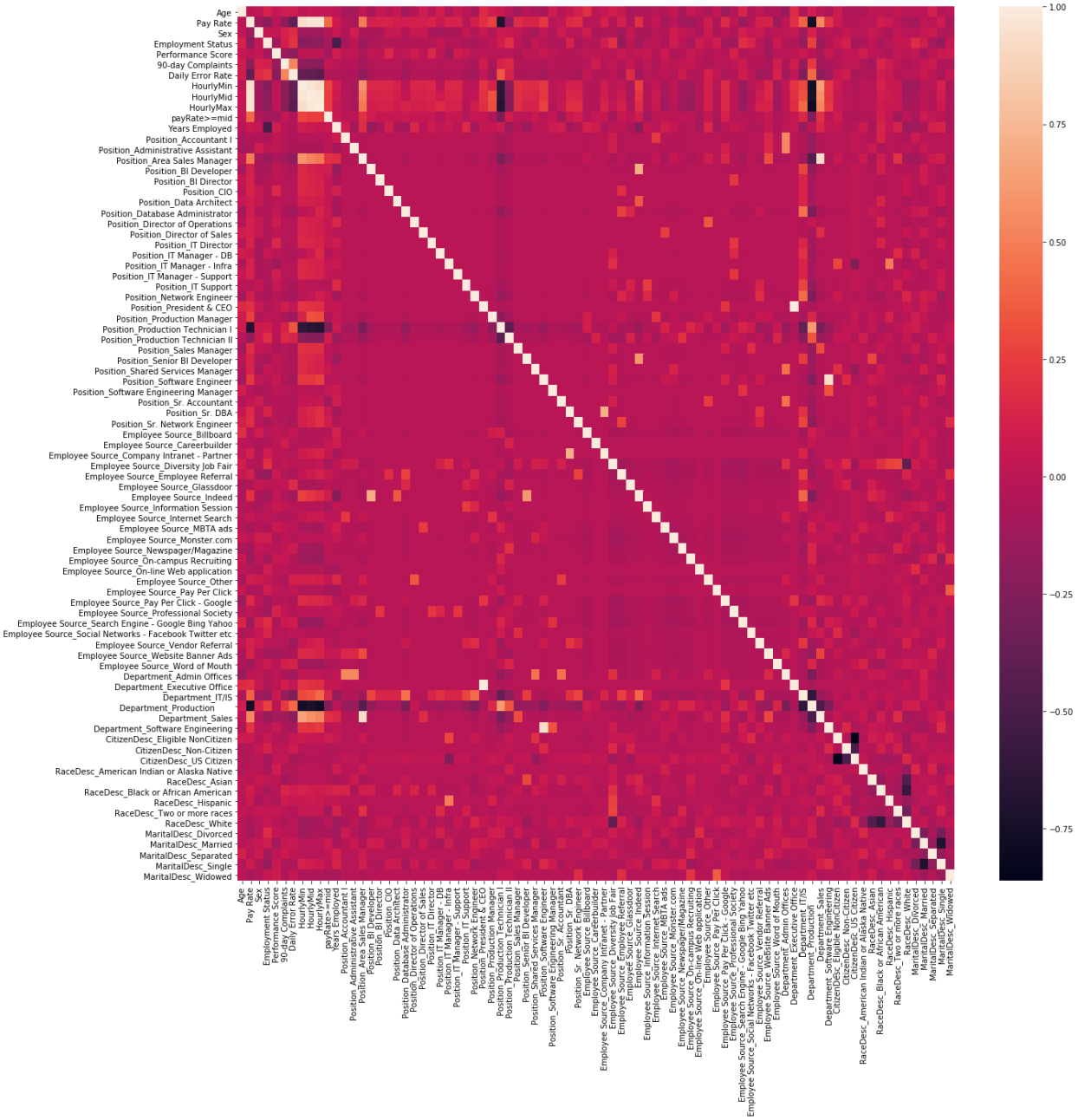
   1.1 Joining Datasets: Primary dataset HR_Dataset _v9 was joined with '90-day Complaints', 'Daily Error Rate' from production_staff dataset using ‘Employee Name’. Hourly minimum, maximum and mids of salaries from the salary dataset was joined with HR_Dataset_v9 using the ‘Position’ attribute. There were additional positions in HR_Dataset_v9 that were missing in ‘salary_dataset’. The hourly values for those positions were computed from HR_Dataset_v9 by grouping on the missing positions. Recruitment costs dataset was left for later analysis.

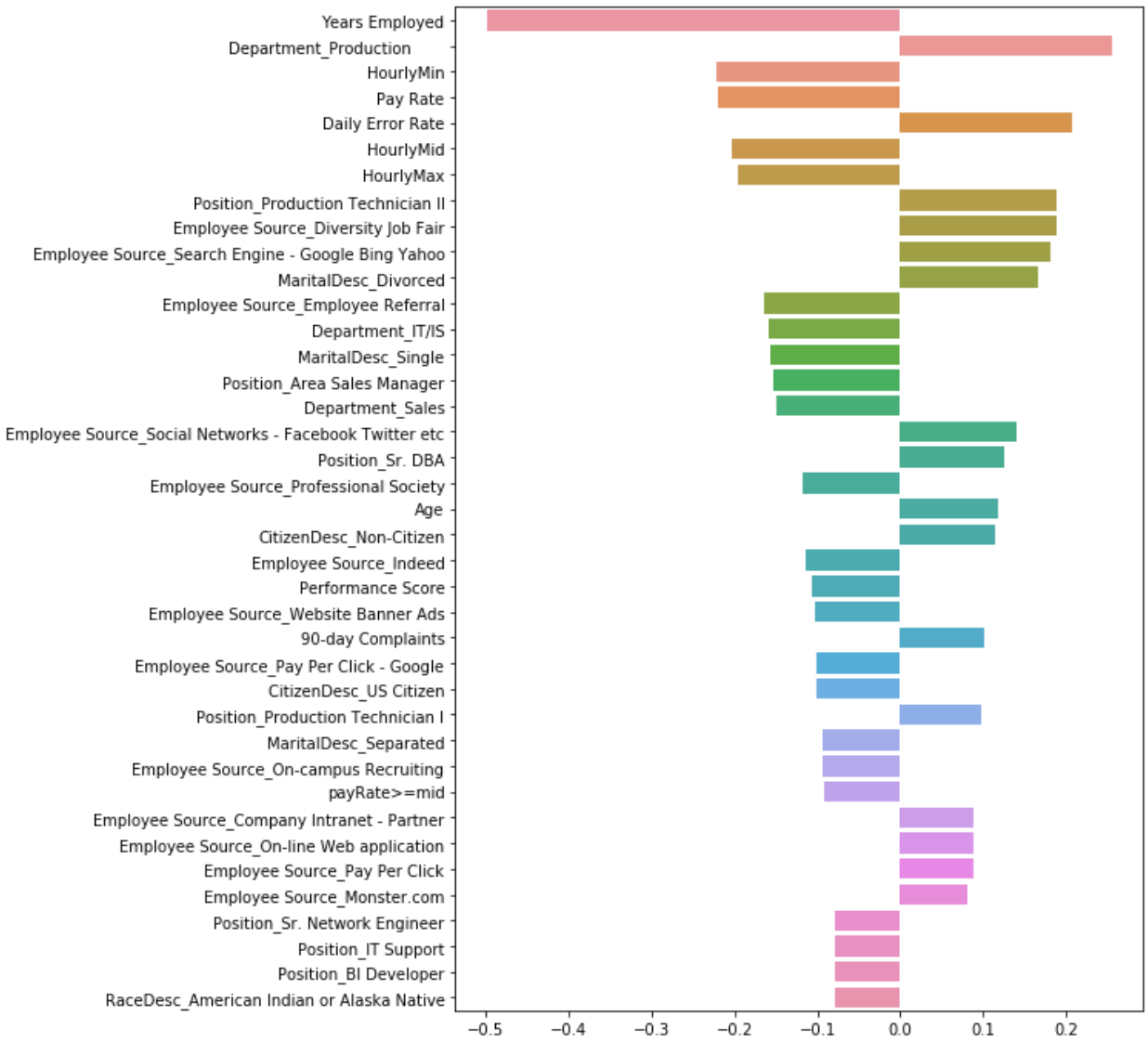
   1.2 Missing Value Imputation: 103 missing values from '90-day Complaints', 'Daily Error Rate' were imputed using 0, since it is the majority and non-conflict value in both attributes. Years were converted to datetime type.

   1.3 Feature Selection: Redundant columns were removed by identifying one-on-one relationships using a self-written function. Manager Name, Location, DOB, Employee Name/ID, Dates were also removed.

   1.4 Feature Engineering: ‘Days employed’ was converted to ‘Years Employed’ to scale it down. A new binary feature ‘payRate>=mid’ was added to mark people receiving more than the median pay for their position type. This could also have been done at gender/ department/ age level, or at a quantile different from median, but was skipped to narrow down the analysis. Binary categorical variables were encoded as 0s/1s. All non-binary categorical variables were one-hot encoded. ‘Performance Score’ was coded as ordinal variable: {'PIP':1, 'Needs Improvement':2, '90-day meets':3, 'N/A- too early to review':4, 'Fully Meets':4, 'Exceeds':5, 'Exceptional':6}. **Target variable selected was ‘Employment Status’. In order to select only voluntary terminations, ‘terminated for cause’ and ‘future start’ were dropped.** Encoding: 'Active':0, 'Leave of Absence':0, 'Voluntarily Terminated':1.

   1.5 Visualizing correlations in final numerical dataset:


Age
Pay Rate
Sex
Employment Status
Performance Score
90-day Complaints
Daily Error Rate
HourlyMin
HourlyMid
HourlyMax
payRate>=mid
Years Employed
Position_Accountant I
Position_Administrative Assistant
Position_Area Sales Manager
Position_BI Developer
Position_BI Director
Position_CIO
Position_Data Architect
Position_Database Administrator
Position_Director of Operations
Position_Director of Sales
Position_IT Director
Position_IT Manager - DB
Position_IT Manager - Infra
Position_IT Manager - Support
Position_IT Support
Position_Network Engineer
Position_President & CEO
Position_Production Manager
Position_Production Technician I
Position_Production Technician II
Position_Sales Manager
Position_Senior BI Developer
Position_Shared Services Manager
Position_Software Engineer
Position_Software Engineering Manager
Position_Sr. Accountant
Position_Sr. DBA
Position_Sr. Network Engineer
Employee Source_Billboard
Employee Source_Careerbuilder
Employee Source_Company Intranet - Partner
Employee Source_Diversity Job Fair
Employee Source_Employee Referral
Employee Source_Glassdoor
Employee Source_Indeed
Employee Source_Information Session
Employee Source_Internet Search
Employee Source_MBTA ads
Employee Source_Monster.com
Employee Source_Newspager/Magazine
Employee Source_On-campus Recruiting
Employee Source_On-line Web application
Employee Source_Other
Employee Source_Pay Per Click
Employee Source_Pay Per Click - Google
Employee Source_Professional Society
Employee Source_Search Engine - Google Bing Yahoo
Employee Source_Social Networks - Facebook Twitter etc
Employee Source_Vendor Referral
Employee Source_Website Banner Ads
Employee Source_Word of Mouth
Department_Admin Offices
Department_Executive Office
Department_IT/IS
Department_Production
Department_Sales
Department_Software Engineering
CitizenDesc_Eligible NonCitizen
CitizenDesc_Non-Citizen
CitizenDesc_US Citizen
RaceDesc_American Indian or Alaska Native
RaceDesc_Asian
RaceDesc_Black or African American
RaceDesc_Hispanic
RaceDesc_Two or more races
RaceDesc_White
MaritalDesc_Divorced
MaritalDesc_Married
MaritalDesc_Separated
MaritalDesc_Single
MaritalDesc_Widowed
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

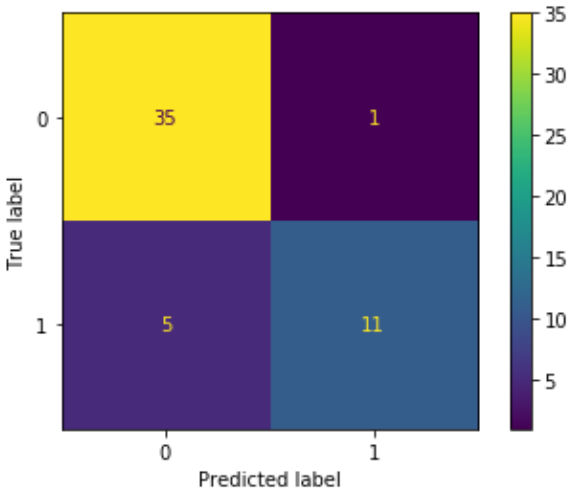


Above figure shows correlation of different variables against the target variable. Positive correlation implies that as the value of the variable increases, the chances of the employee voluntarily leaving the job increases.

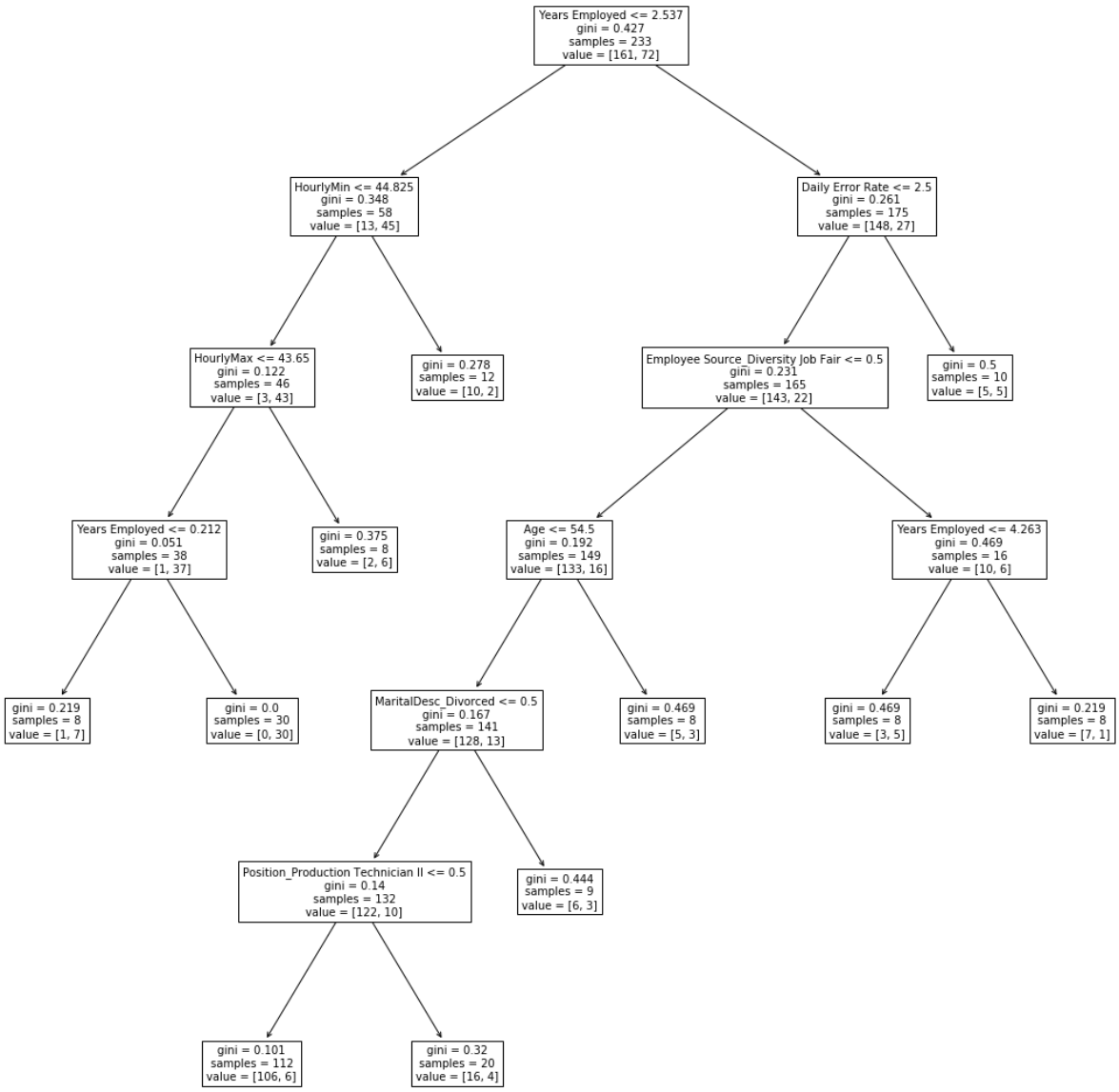
2. Modeling and Evaluation

2.1 Evaluation Methodology: The entire training set was split into train (72%, 161 active and 72 terminated) and test (18%, 36 active and 16 terminated) sets. Two simplistic models were used: Decision Tree Classifier and Logistic Regression. The best models (or hyperparameters) were selected using brute grid search over all parameter combinations and k-fold cross validation with k=5, instead of using a validation set because of the very small size of the dataset. AUC ROC score was the metric used by the grid search. Thereafter, accuracies were computed on the training and test sets.

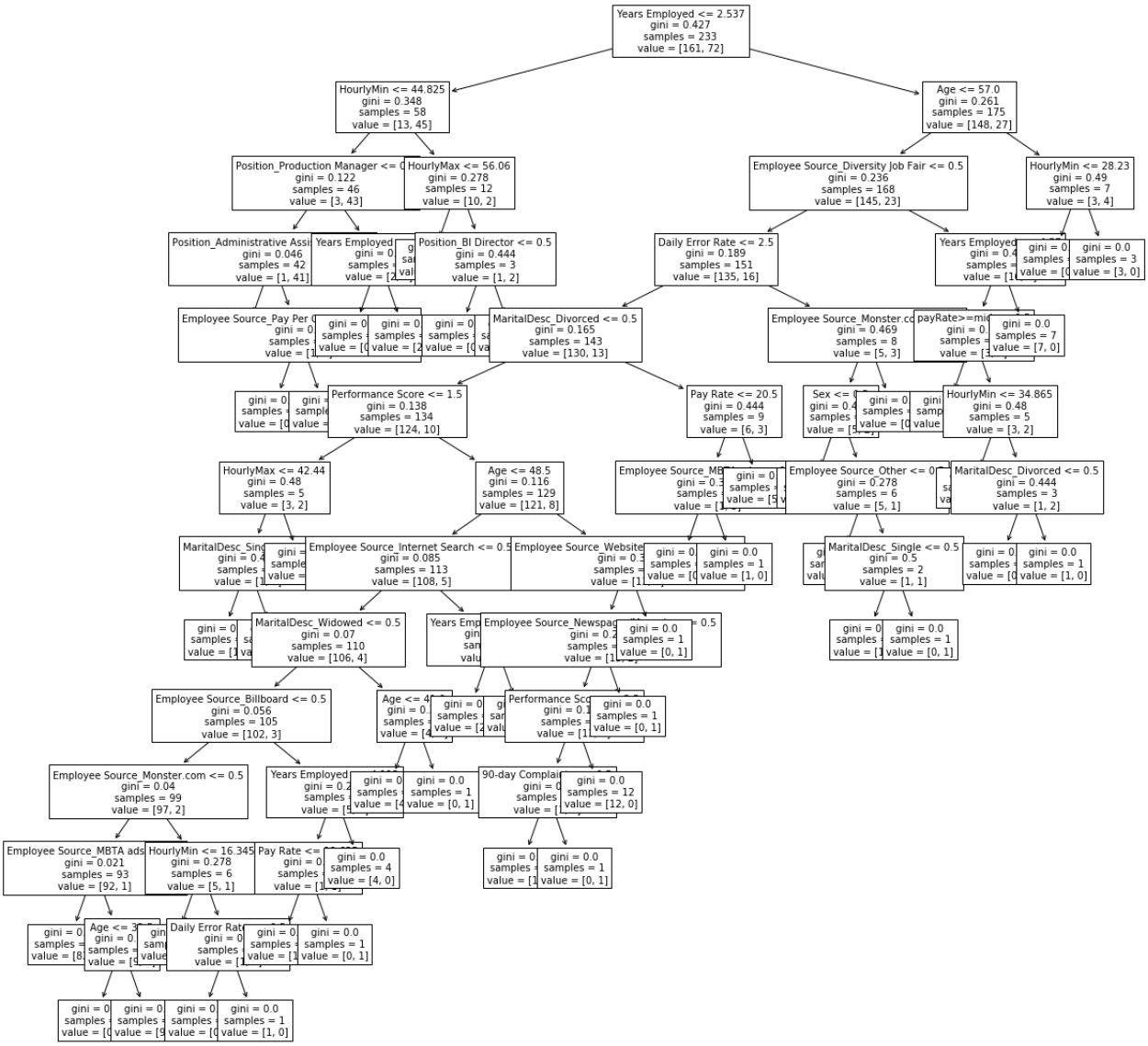
2.2 Decision Tree: Parameters tuned: {'max_depth':np.arange(1,15), 'min_samples_split':np.arange(2,15), 'min_samples_leaf':np.arange(1,15)}. Best parameters: {'max_depth': 6, 'min_samples_leaf': 8, 'min_samples_split': 12}. Train set accuracy = 87.12% and test set accuracy= 88.46%. Confusion matrix on test set is given below:



The visualized decision tree is as follows:

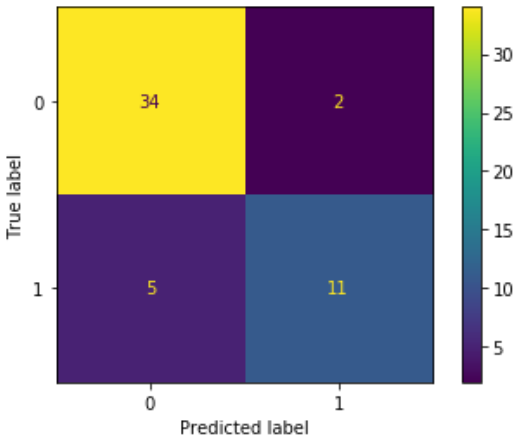


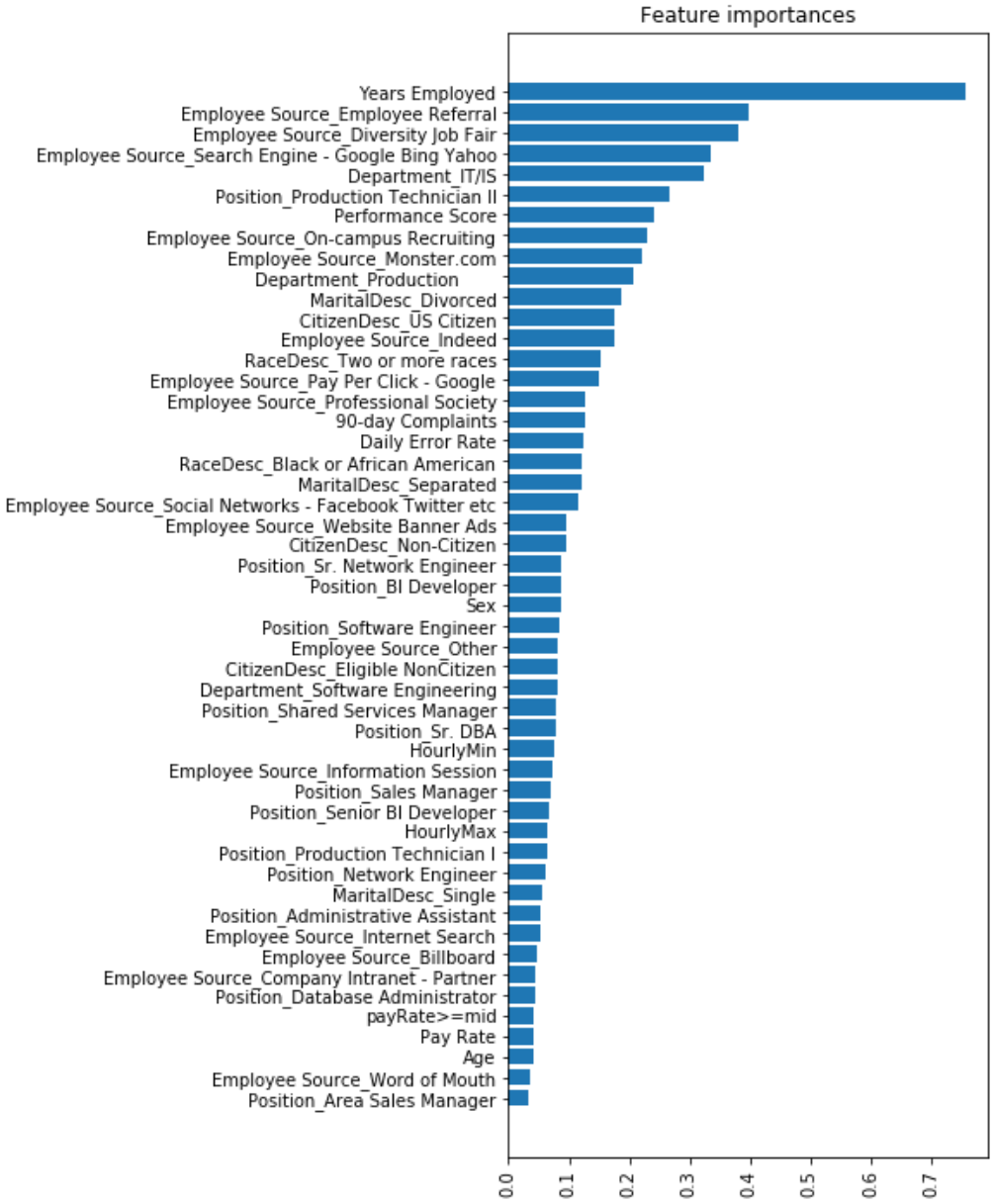
The unpruned, overfitted decision tree would have looked like:



Clearly, the decision tree model suffered from stumping on a few variables, and could not give a good idea of the most important features. Hence, we use logistic regression.

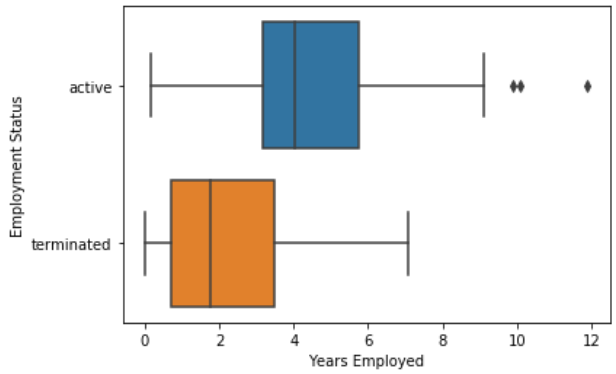
2.3 Logistic Regression: Parameters tuned: {'C':[1e-3, 1e-2, 1e-1, 1, 10]}. Best parameter: {'C': 0.1}. Training set accuracy = 90.55% and test set accuracy = 86.54%. The confusion matrix is as follows:

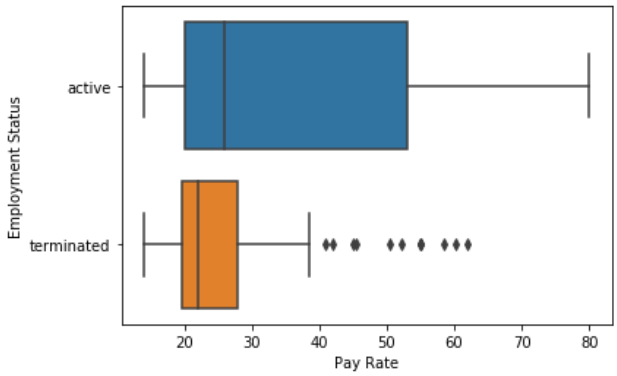

34
2
5
11
0
1
True label
Predicted label
30
25
20
15
10
5

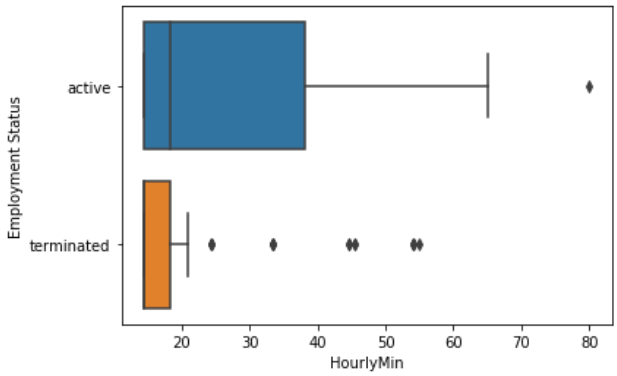

Feature importances
Years Employed
Employee Source_Employee Referral
Employee Source_Diversity Job Fair
Employee Source_Search Engine - Google Bing Yahoo
Department_IT/IS
Position_Production Technician II
Performance Score
Employee Source_On-campus Recruiting
Employee Source_Monster.com
Department_Production
MaritalDesc_Divorced
CitizenDesc_US Citizen
Employee Source_Indeed
RaceDesc_Two or more races
Employee Source_Pay Per Click - Google
Employee Source_Professional Society
90-day Complaints
Daily Error Rate
RaceDesc_Black or African American
MaritalDesc_Separated
Employee Source_Social Networks - Facebook Twitter etc
Employee Source_Website Banner Ads
CitizenDesc_Non-Citizen
Position_Sr. Network Engineer
Position_BI Developer
Sex
Position_Software Engineer
Employee Source_Other
CitizenDesc_Eligible NonCitizen
Department_Software Engineering
Position_Shared Services Manager
Position_Sr. DBA
HourlyMin
Employee Source_Information Session
Position_Sales Manager
Position_Senior BI Developer
HourlyMax
Position_Production Technician I
Position_Network Engineer
MaritalDesc_Single
Position_Administrative Assistant
Employee Source_Internet Search
Employee Source_Billboard
Employee Source_Company Intranet - Partner
Position_Database Administrator
payRate>=mid
Pay Rate
Age
Employee Source_Word of Mouth
Position_Area Sales Manager
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

3. Insights

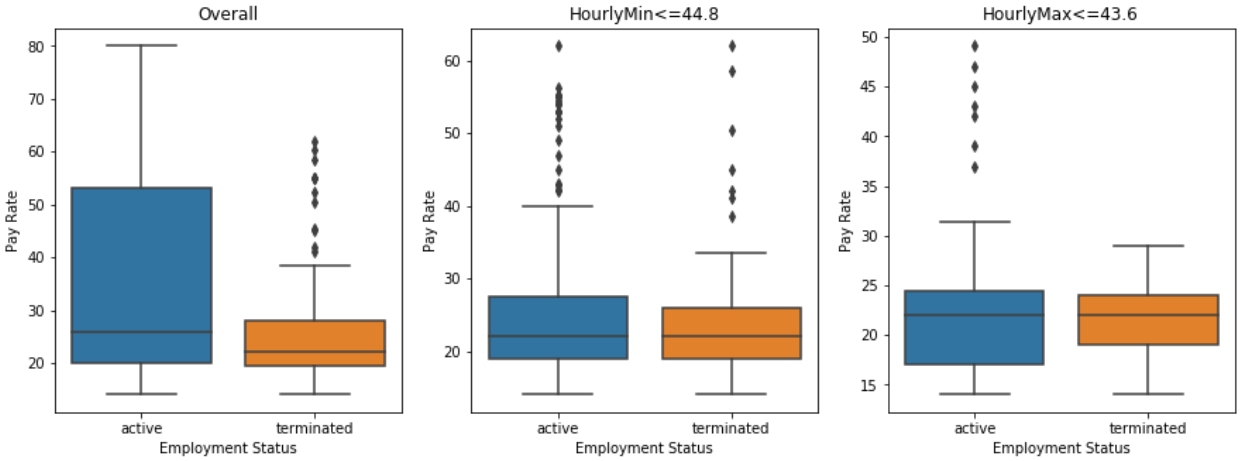
3.1 Lesser experience (low value of years employed), low individual pay rate, and generally low wage of the position of an employee has a higher probability of leading to the employee leaving the company.



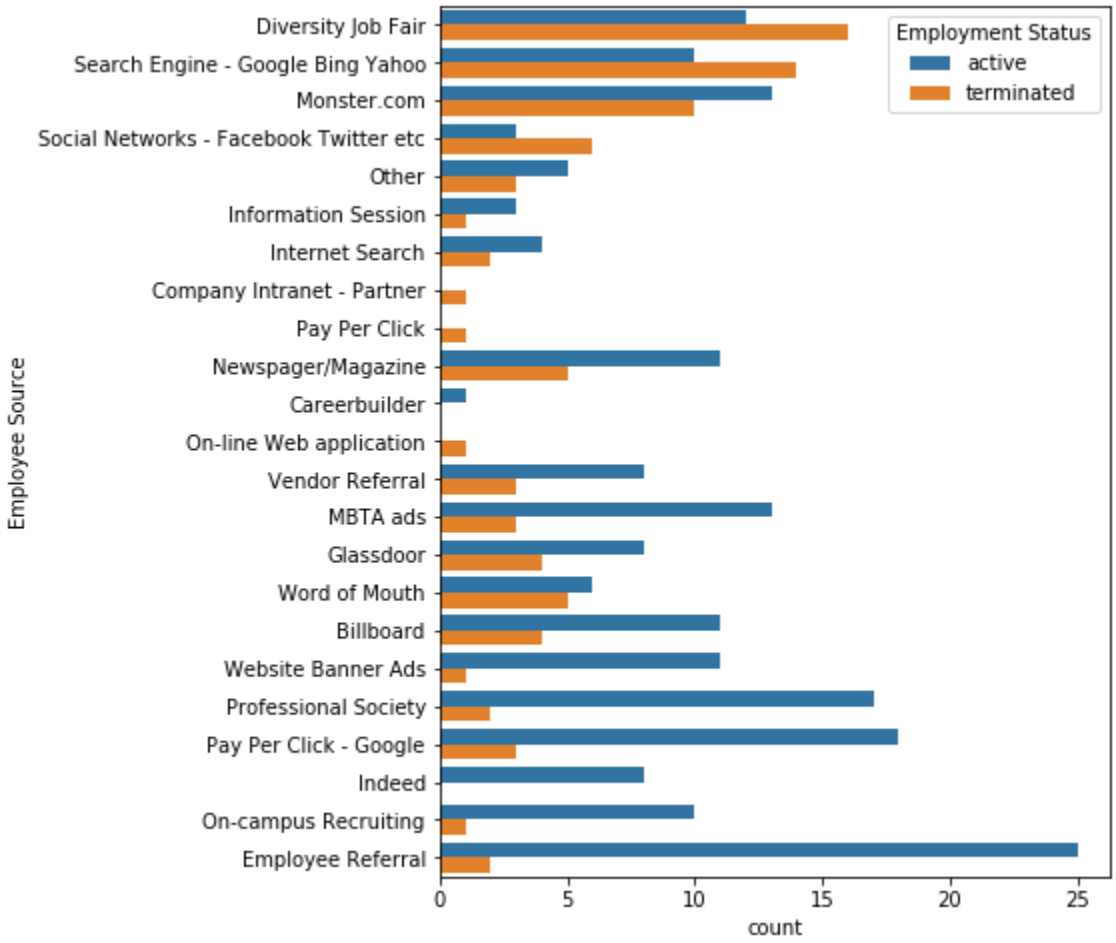




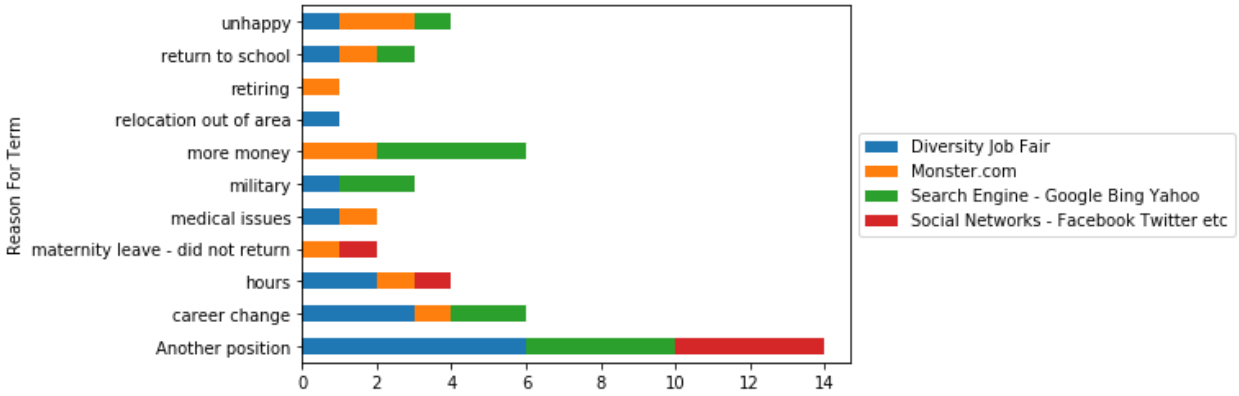
Especially when the hourly minimum and maximum for a position is quite low, the employee with the lower pay has an even higher chance of leaving (limits set from the decision tree rules).



3.2 Using the logistic regression feature importances and ordering the employee sources in order of feature importances, it can be seen that employee sources have a strong impact on the probability of the employee leaving. Clearly, employees from Diversity Job Fair, Search Engines (Google, Bing, Yahoo), Social Networks (Facebook, Twitter, etc.) leave more often than stay. On the other hand, employees from Indeed, on-campus recruiting and referrals stay more often than leave.



Next, the reasons for terminations among 4 of the riskiest employee sources were plotted.



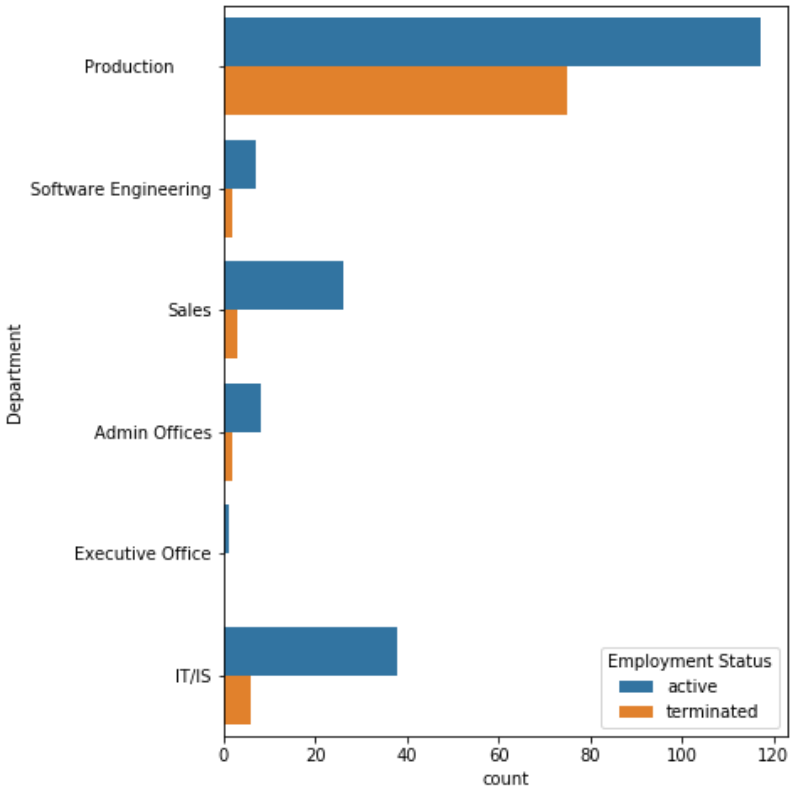
It seems that people from the Diversity Job Fair are unhappy, or dissatisfied with hours and want a change or career or position.

The recruitment costs dataset was brought back and used to calculate the effectiveness of source by dividing the percentage of employees retained by total recruitments costs. The table in increasing order of effectiveness is as follows:

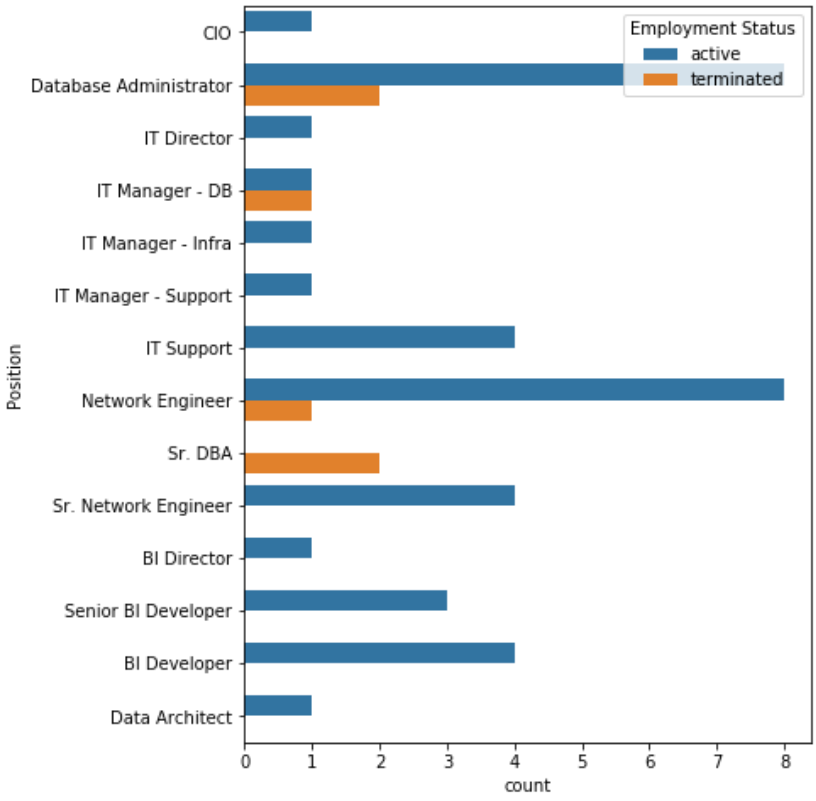
Pay Per Click	0.000000
Diversity Job Fair	0.000043
Social Networks - Facebook Twitter etc	0.000060
MBTA ads	0.000074
Search Engine - Google Bing Yahoo	0.000080
Newspager/Magazine	0.000083
Monster.com	0.000098
Billboard	0.000118
On-campus Recruiting	0.000121
Website Banner Ads	0.000128
Careerbuilder	0.000128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

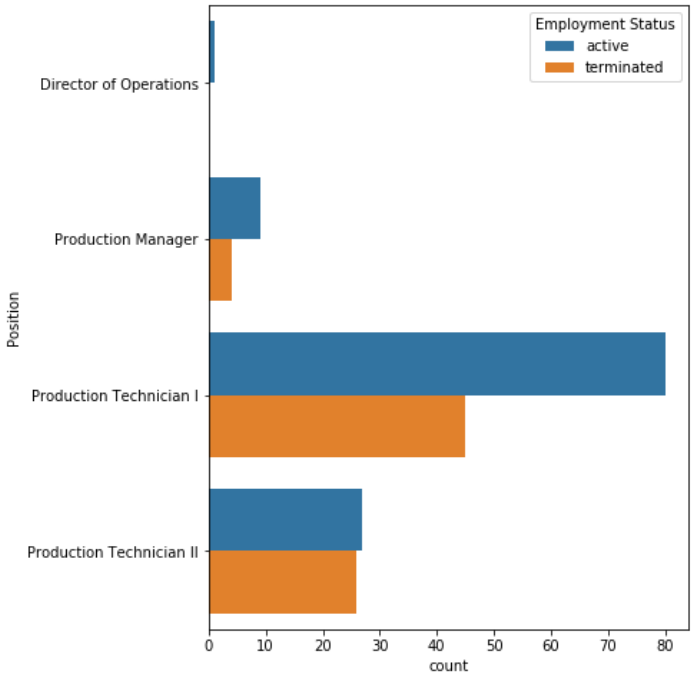
Again, we can see that Pay Per Click, Diversity Job Fair, Social Network and Social Engines are the least cost-effective sources.

3.3 It was also observed that among the departments, Production sees the most voluntary terminations and IT/IS sees the least.

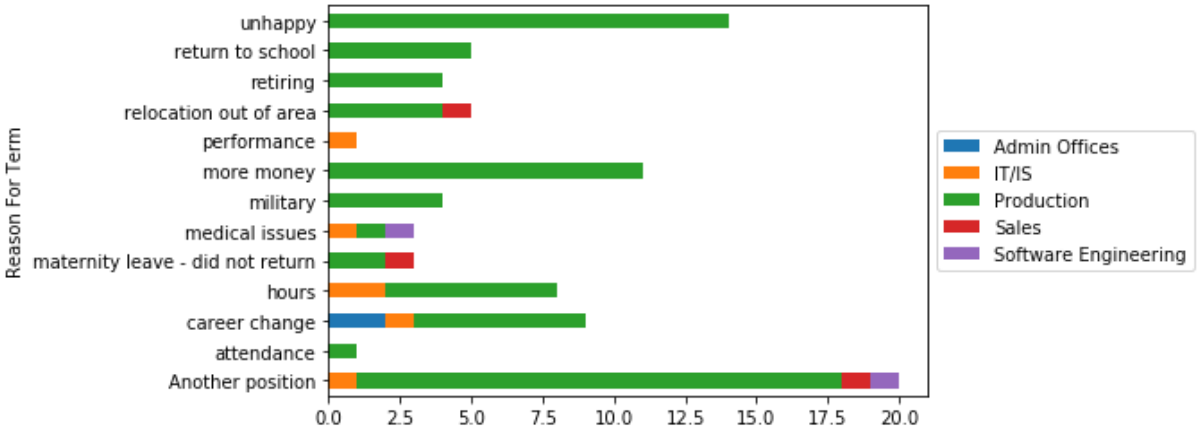


Zooming in on the positions in these two departments, we find that in IT, DBAs tend to leave whereas in Production, Production Technicians tend to leave.





Next, the reasons for termination were examined for all departments.



Only the production department reports ‘unhappy’ and ‘more money’ in the reasons for leaving. Examining the salaries of the departments, it is found that the Production department does earn the least.

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