Employee Turnover Prediction

Background and Goals:

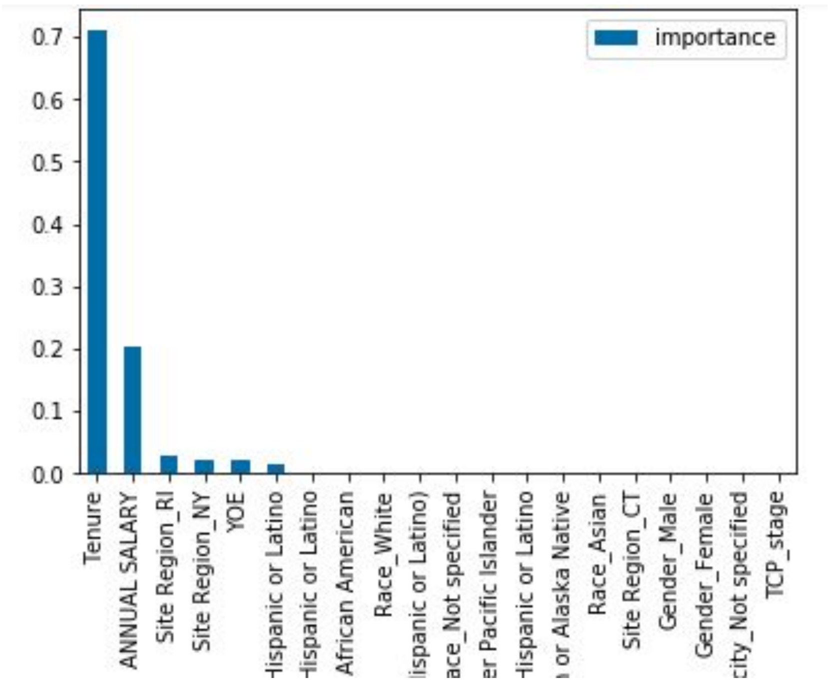
1. Employee churn could have varied impact on the whole organization: losing valued employees could cost more than expectation. Losing track of the employee's potential career path could also be concerning to Talent Management Team.
2. Predicting employees' attrition could be beneficial in multiple ways: 1) estimating headcounts 2) Reducing attrition rate by taking action in advance.
3. In this prediction, we not only identify important factors that help predict employee attrition with favourable prediction accuracy and further dive in the impact of these factors on employee's tenure.

Methods

1. We constructed decision tree models using ski-learn package on Python 3.0.
2. There are **1800+** complete cases imported to the models. The data was split into training and testing subsets for validation purpose
3. We only look at one position, the pillar role within the organization.
4. Data used from multiple resources, includes **Annual Salary, Ethnicity, Gender, Site Region, Race, Seniority Level, Tenure and Attrition Status**
5. In order to avoid overfitting(making the models more generic), we used post-pruning method: setting appropriate effective alpha
6. A follow-up survival analysis helps identify the survival(Retention) and hazard(Dropout) probability

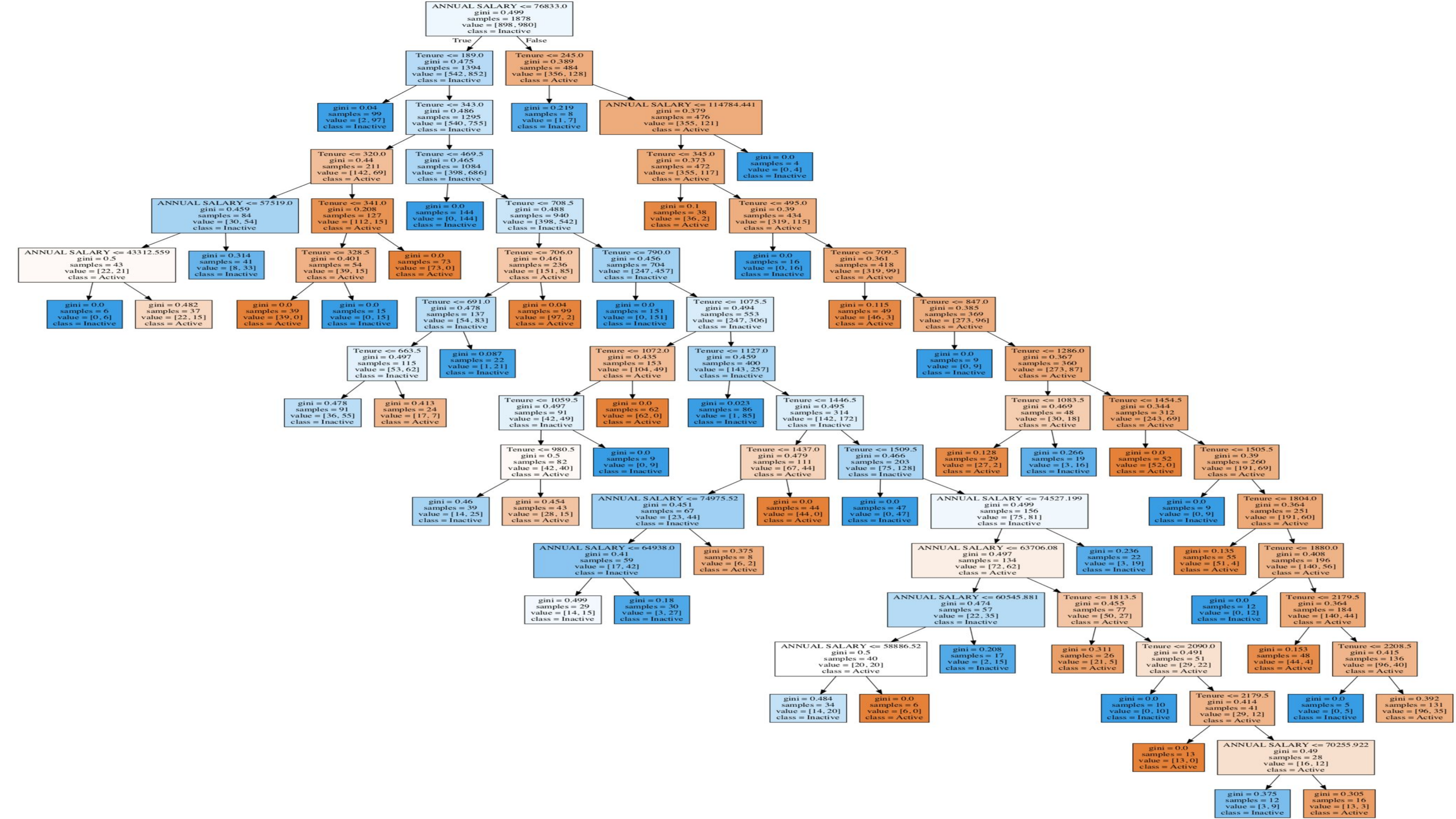
Key insights from DT model:

1. Important feature that could affect/determine this role’s turnover : **Tenure** and **Annual Salary**. Tenure is more determining than annual salary.
2. Important cutoff:
   a. Salary <= 76833 $/y is more likely to leave ; if these ‘lower-income’ employees could survive the first 6 months, their(ranging from 43312 to 57519) chance of staying could be higher, approaching 1 year.
   b. This role with salary over 76833 $/y should be paid attention at certain periods of time: 8 months since onboarding (salary higher than 114748 $/y is very likely to leave); 1.3 years; 2.3 years; 3.5 years; 4.1 years; 5.1 years; 6 years. After 6 years, their potential turnover tend to be small.



Top 2

feature	importance
Tenure	0.710
ANNUAL SALARY	0.205
Site Region_RI	0.028
Site Region_NY	0.021
YOE	0.020
Ethnicity_Not Hispanic or Latino	0.015
Ethnicity_Hispanic or Latino	0.000
Race_Black or African American	0.000
Race_White	0.000



ANNUAL SALARY <= 76833.0
gini = 0.499
samples = 1878
value = [898, 980]
class = Inactive
True
False
Tenure <= 189.0
gini = 0.475
samples = 1394
value = [542, 852]
class = Inactive
Tenure <= 245.0
gini = 0.389
samples = 484
value = [356, 128]
class = Active
gini = 0.04
samples = 99
value = [2, 97]
class = Inactive
Tenure <= 343.0
gini = 0.486
samples = 1295
value = [540, 755]
class = Inactive
gini = 0.219
samples = 8
value = [1, 7]
class = Inactive
ANNUAL SALARY <= 114784.441
gini = 0.379
samples = 476
value = [355, 121]
class = Active
Tenure <= 320.0
gini = 0.44
samples = 211
value = [142, 69]
class = Active
Tenure <= 469.5
gini = 0.465
samples = 1084
value = [398, 686]
class = Inactive
Tenure <= 345.0
gini = 0.373
samples = 472
value = [355, 117]
class = Active
gini = 0.0
samples = 4
value = [0, 4]
class = Inactive
ANNUAL SALARY <= 57519.0
gini = 0.459
samples = 84
value = [30, 54]
class = Inactive
Tenure <= 341.0
gini = 0.208
samples = 127
value = [112, 15]
class = Active
gini = 0.0
samples = 144
value = [0, 144]
class = Inactive
Tenure <= 708.5
gini = 0.488
samples = 940
value = [398, 542]
class = Inactive
gini = 0.1
samples = 38
value = [36, 2]
class = Active
Tenure <= 495.0
gini = 0.39
samples = 434
value = [319, 115]
class = Active
ANNUAL SALARY <= 43312.559
gini = 0.5
samples = 43
value = [22, 21]
class = Active
gini = 0.314
samples = 41
value = [8, 33]
class = Inactive
Tenure <= 328.5
gini = 0.401
samples = 54
value = [39, 15]
class = Active
gini = 0.0
samples = 73
value = [73, 0]
class = Active
Tenure <= 706.0
gini = 0.461
samples = 236
value = [151, 85]
class = Active
Tenure <= 790.0
gini = 0.456
samples = 704
value = [247, 457]
class = Inactive
gini = 0.0
samples = 16
value = [0, 16]
class = Inactive
Tenure <= 709.5
gini = 0.361
samples = 418
value = [319, 99]
class = Active
gini = 0.0
samples = 6
value = [0, 6]
class = Inactive
gini = 0.482
samples = 37
value = [22, 15]
class = Active
gini = 0.0
samples = 39
value = [39, 0]
class = Active
gini = 0.0
samples = 15
value = [0, 15]
class = Inactive
Tenure <= 691.0
gini = 0.478
samples = 137
value = [54, 83]
class = Inactive
gini = 0.04
samples = 99
value = [97, 2]
class = Active
gini = 0.0
samples = 151
value = [0, 151]
class = Inactive
Tenure <= 1075.5
gini = 0.494
samples = 553
value = [247, 306]
class = Inactive
gini = 0.115
samples = 49
value = [46, 3]
class = Active
Tenure <= 847.0
gini = 0.385
samples = 369
value = [273, 96]
class = Active
Tenure <= 663.5
gini = 0.497
samples = 115
value = [53, 62]
class = Inactive
gini = 0.087
samples = 22
value = [1, 21]
class = Inactive
Tenure <= 1072.0
gini = 0.435
samples = 153
value = [104, 49]
class = Active
Tenure <= 1127.0
gini = 0.459
samples = 400
value = [143, 257]
class = Inactive
gini = 0.0
samples = 9
value = [0, 9]
class = Inactive
Tenure <= 1286.0
gini = 0.367
samples = 360
value = [273, 87]
class = Active
gini = 0.478
samples = 91
value = [36, 55]
class = Inactive
gini = 0.413
samples = 24
value = [17, 7]
class = Active
Tenure <= 1059.5
gini = 0.497
samples = 91
value = [42, 49]
class = Inactive
gini = 0.0
samples = 62
value = [62, 0]
class = Active
gini = 0.023
samples = 86
value = [1, 85]
class = Inactive
Tenure <= 1446.5
gini = 0.495
samples = 314
value = [142, 172]
class = Inactive
Tenure <= 1083.5
gini = 0.469
samples = 48
value = [30, 18]
class = Active
Tenure <= 1454.5
gini = 0.344
samples = 312
value = [243, 69]
class = Active
Tenure <= 980.5
gini = 0.5
samples = 82
value = [42, 40]
class = Active
gini = 0.0
samples = 9
value = [0, 9]
class = Inactive
Tenure <= 1437.0
gini = 0.479
samples = 111
value = [67, 44]
class = Active
Tenure <= 1509.5
gini = 0.466
samples = 203
value = [75, 128]
class = Inactive
gini = 0.128
samples = 29
value = [27, 2]
class = Active
gini = 0.266
samples = 19
value = [3, 16]
class = Inactive
gini = 0.0
samples = 52
value = [52, 0]
class = Active
Tenure <= 1505.5
gini = 0.39
samples = 260
value = [191, 69]
class = Active
gini = 0.46
samples = 39
value = [14, 25]
class = Inactive
gini = 0.454
samples = 43
value = [28, 15]
class = Active
ANNUAL SALARY <= 74975.52
gini = 0.451
samples = 67
value = [23, 44]
class = Inactive
gini = 0.0
samples = 44
value = [44, 0]
class = Active
gini = 0.0
samples = 47
value = [0, 47]
class = Inactive
ANNUAL SALARY <= 74527.199
gini = 0.499
samples = 156
value = [75, 81]
class = Inactive
gini = 0.0
samples = 9
value = [0, 9]
class = Inactive
Tenure <= 1804.0
gini = 0.364
samples = 251
value = [191, 60]
class = Active
ANNUAL SALARY <= 64938.0
gini = 0.41
samples = 59
value = [17, 42]
class = Inactive
gini = 0.375
samples = 8
value = [6, 2]
class = Active
ANNUAL SALARY <= 63706.08
gini = 0.497
samples = 134
value = [72, 62]
class = Active
gini = 0.236
samples = 22
value = [3, 19]
class = Inactive
gini = 0.135
samples = 55
value = [51, 4]
class = Active
Tenure <= 1880.0
gini = 0.408
samples = 196
value = [140, 56]
class = Active
gini = 0.499
samples = 29
value = [14, 15]
class = Inactive
gini = 0.18
samples = 30
value = [3, 27]
class = Inactive
ANNUAL SALARY <= 60545.881
gini = 0.474
samples = 57
value = [22, 35]
class = Inactive
Tenure <= 1813.5
gini = 0.455
samples = 77
value = [50, 27]
class = Active
gini = 0.0
samples = 12
value = [0, 12]
class = Inactive
Tenure <= 2179.5
gini = 0.364
samples = 184
value = [140, 44]
class = Active
ANNUAL SALARY <= 58886.52
gini = 0.5
samples = 40
value = [20, 20]
class = Active
gini = 0.208
samples = 17
value = [2, 15]
class = Inactive
gini = 0.311
samples = 26
value = [21, 5]
class = Active
Tenure <= 2090.0
gini = 0.491
samples = 51
value = [29, 22]
class = Active
gini = 0.153
samples = 48
value = [44, 4]
class = Active
Tenure <= 2208.5
gini = 0.415
samples = 136
value = [96, 40]
class = Active
gini = 0.484
samples = 34
value = [14, 20]
class = Inactive
gini = 0.0
samples = 6
value = [6, 0]
class = Active
gini = 0.0
samples = 10
value = [0, 10]
class = Inactive
Tenure <= 2179.5
gini = 0.414
samples = 41
value = [29, 12]
class = Active
gini = 0.0
samples = 5
value = [0, 5]
class = Inactive
gini = 0.392
samples = 131
value = [96, 35]
class = Active
gini = 0.0
samples = 13
value = [13, 0]
class = Active
ANNUAL SALARY <= 70255.922
gini = 0.49
samples = 28
value = [16, 12]
class = Active
gini = 0.375
samples = 12
value = [3, 9]
class = Inactive
gini = 0.305
samples = 16
value = [13, 3]
class = Active

Question from audience:

We know that the salaries vary from region to region: how could we just simply draw a generic salary cutoff line to determine who is more likely to leave?

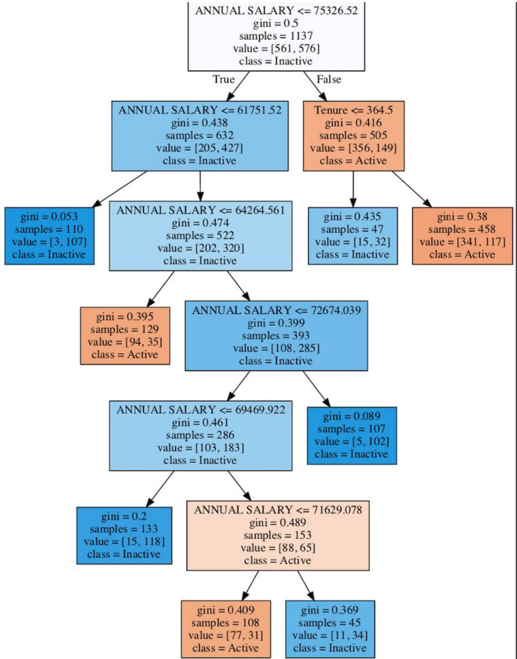
We should also notice that in the feature importance section, the model captures the predictability of the geographical features though relatively small compared with salary and tenure, indicating regional differences in leave and stay is **small**. The key insights from the DT model simply serves as elementary reference and predictive tool for this type of employee’s turnover in general. It provides a couple of numbers to reflect on the salary as well as tenure status at large and potential equity issues. More deep dive into region is needed to finally offer a customized inference.

NY

Key insights from NY’s DT model:

1. Highly driven by Annual Salary, a basic cutoff will be around 75326.52 $/year. For those who earn higher than this will be prone to be more stable after 10 months of stay.

	importance
feature	
ANNUAL SALARY	0.93
Tenure	0.07

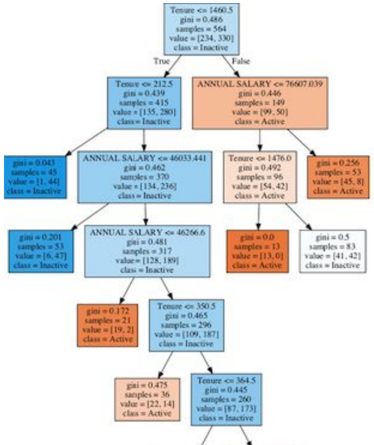


CT

Key insights from CT's DT model:

1. Compared with NY, CT Teacher's turnover are more sensitive to tenure.
2. In the first four years of this role's career life cycle, they could experience higher turnover hazard with few exceptions (some teachers with an annual salary of lower 46000$/year: part-time).
3. They are less likely to leave after staying for 4 years.

feature	importance
Tenure	0.782
ANNUAL SALARY	0.218

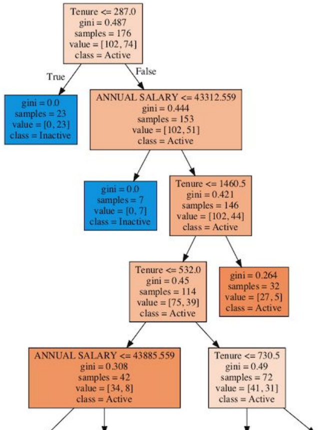


RI

Key insights from RI’s DT model:

1. Compared with CT and NY, RI’s turnover is more sensitive to tenure. They are more stable in the first few months regardless of the salary band. But after 9.5 months, those with an annual salary of lower than 43315.6 $/ year are more likely to leave
2. A few time dots require attention to: 2 years, 2.4 years, 4 years,（this applies to the roles with various salary bands）after 4 years, they are prone to be stable.

feature	importance
Tenure	0.859
ANNUAL SALARY	0.141



Results in short:

1. Among all the selected features, tenure and annual salary play the most important role in predicting the attrition results and thus leads to further analysis of what these two factors can tell us in the following slide
2. The model works well with an overall accuracy of 86% on test data while 89% on training data.
3. The DT models could work efficiently with **annual salary** and **tenure** as inputs.
4. The model is more likely for estimation of headcount instead of flagging employees on the brink of the dropout (For real-life practice's sake)

Conclusion

1. Using decision tree model quickly gives us an insight of features that impact AF’s teacher attrition and also generates a basic cutoff of salary and tenure that Talent Operations practitioners could be aware of.
2. This model also works well with high predictability with ~86% accuracy rate).
3. However, the generic timeline and salary cutoff won’t be able to tell us how likely (the probability) teachers of different levels or who just meet the threshold will actually leave or stay. This leads to using another model, survival model for estimation.
4. Additionally, why there is this drastic regional difference with regard to the feature importance also encourages us to have a further research on the exit surveys and local people strategy? Is there any way we could shorten the ‘safe’ period for attrition? How we could optimize the situation?