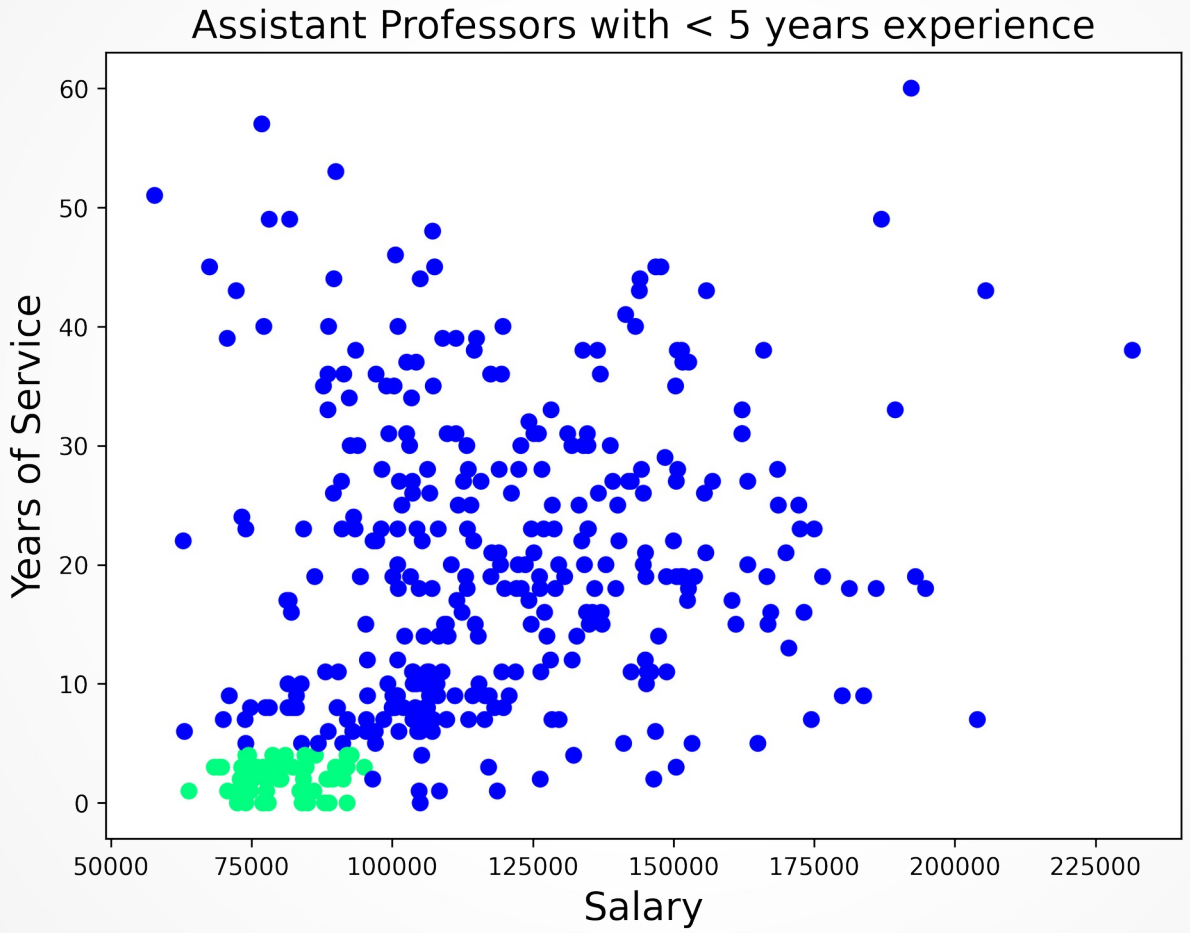
Analysis

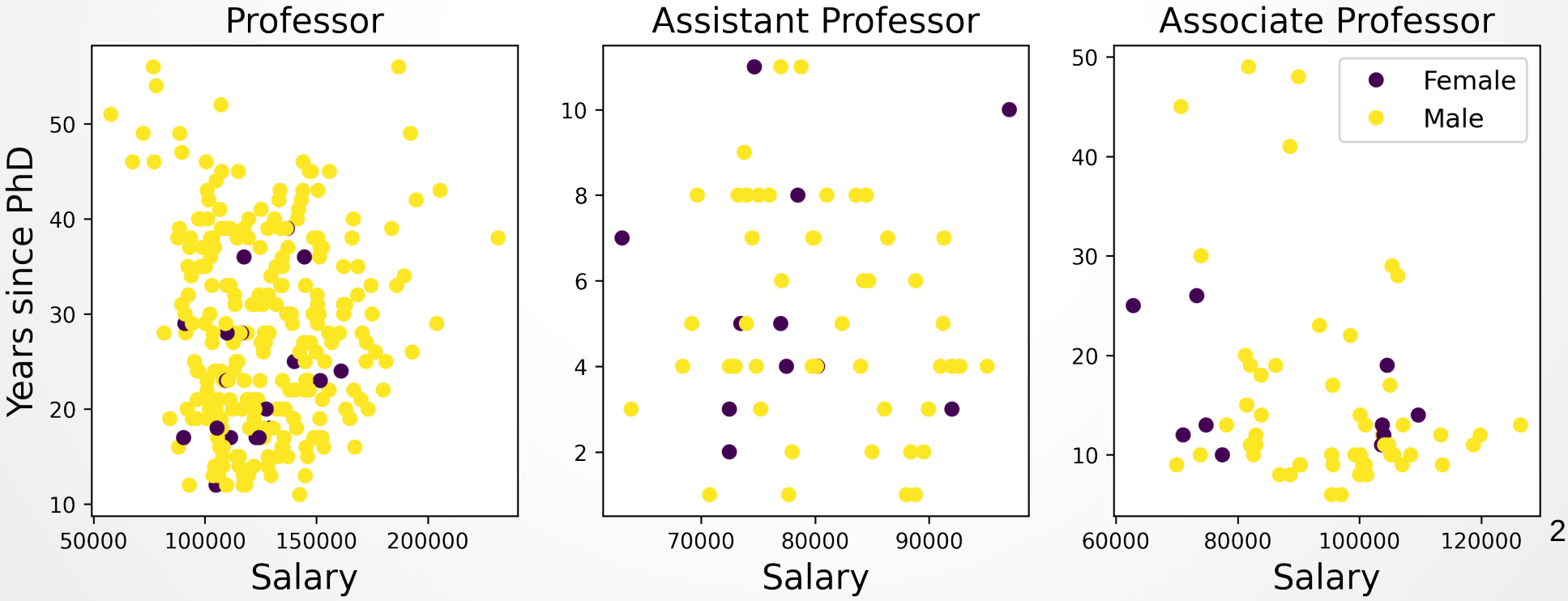
1) 16% of the records are Assistant Professors with less than 5 years of experience.



Analysis

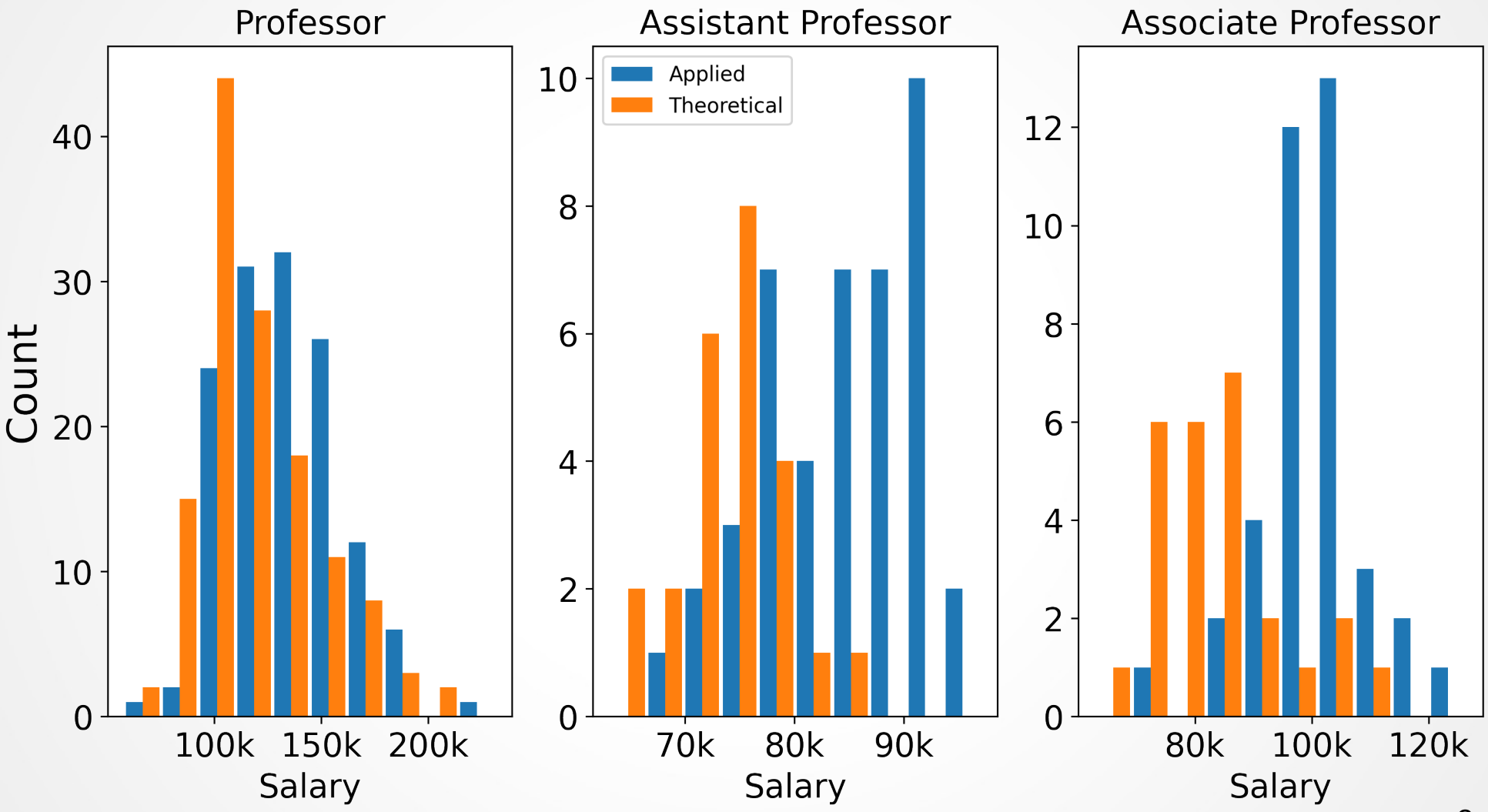
2) The average salary for female professors is 101k ± 26k, while the average salary for male professors is 115k ± 30k. This is a statistically significant difference (p = 0.006) using a two-tailed t-test.

If we look at salary by rank, the salaries of male professors is higher in each rank, but not by a statistically significant amount (p = 0.23, 0.18, 0.45 for Assistant, Asssociate, and Full professors respectively), but a significantly higher percentage of male professors are at the full professor rank (p = 0.006), which has a much higher average salary.



Analysis

3) The salary distributions by rank and discipline are as follows:



Analysis

4) For a single feature I would prefer an explicit mapping for readability (pd.Series.map), but for model building where I am encoding multiple variables the get_dummies function would be prefered.

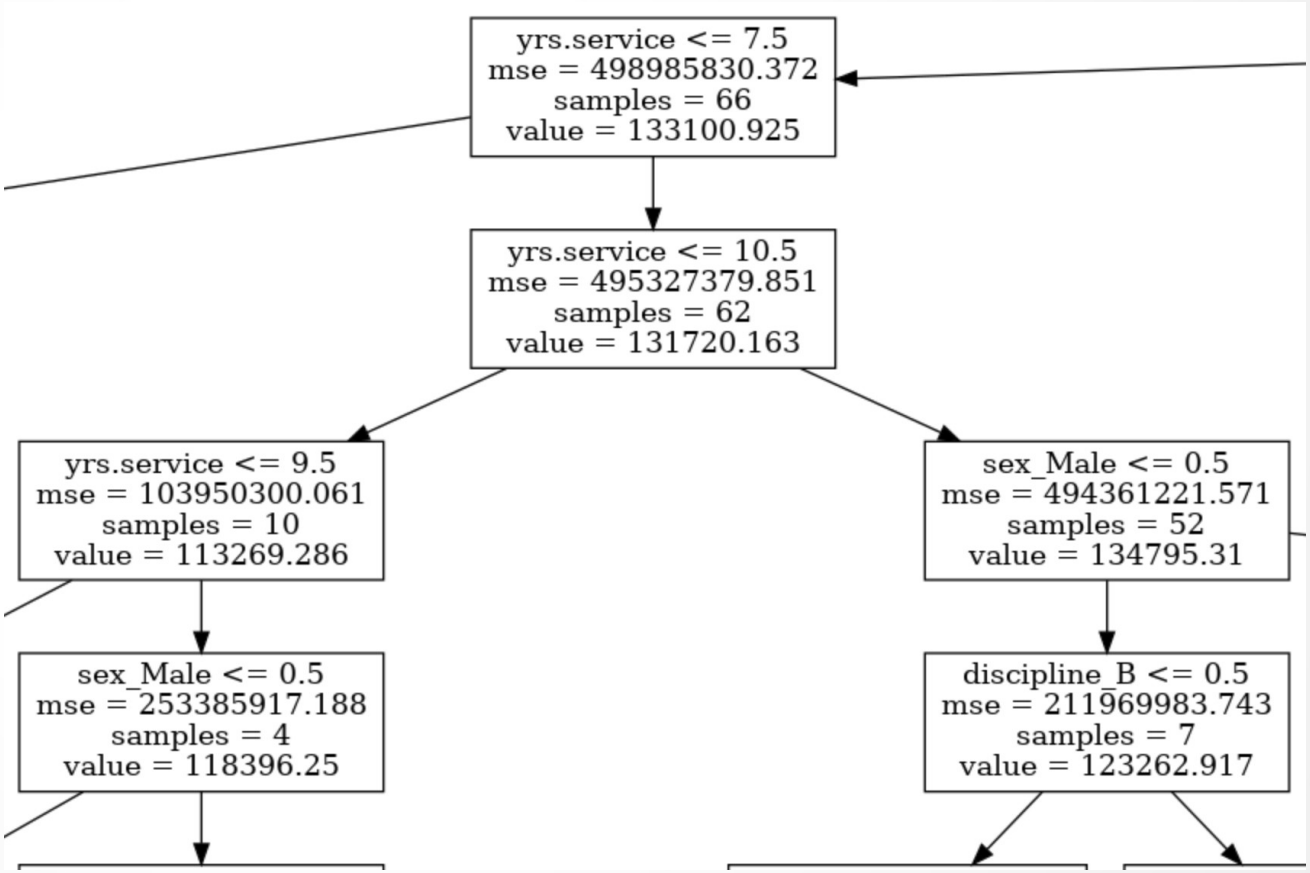
	yrs.since.phd	yrs.service	salary	rank_AssocProf	rank_AsstProf	rank_Prof	discipline_B	sex_Male	above_median_True
0	19	18	139750	0	0	1	1	1	1
1	20	16	173200	0	0	1	1	1	1
2	4	3	79750	0	1	0	1	1	0
3	45	39	115000	0	0	1	1	1	1
4	40	41	141500	0	0	1	1	1	1

Model Building

I use a random forest regression

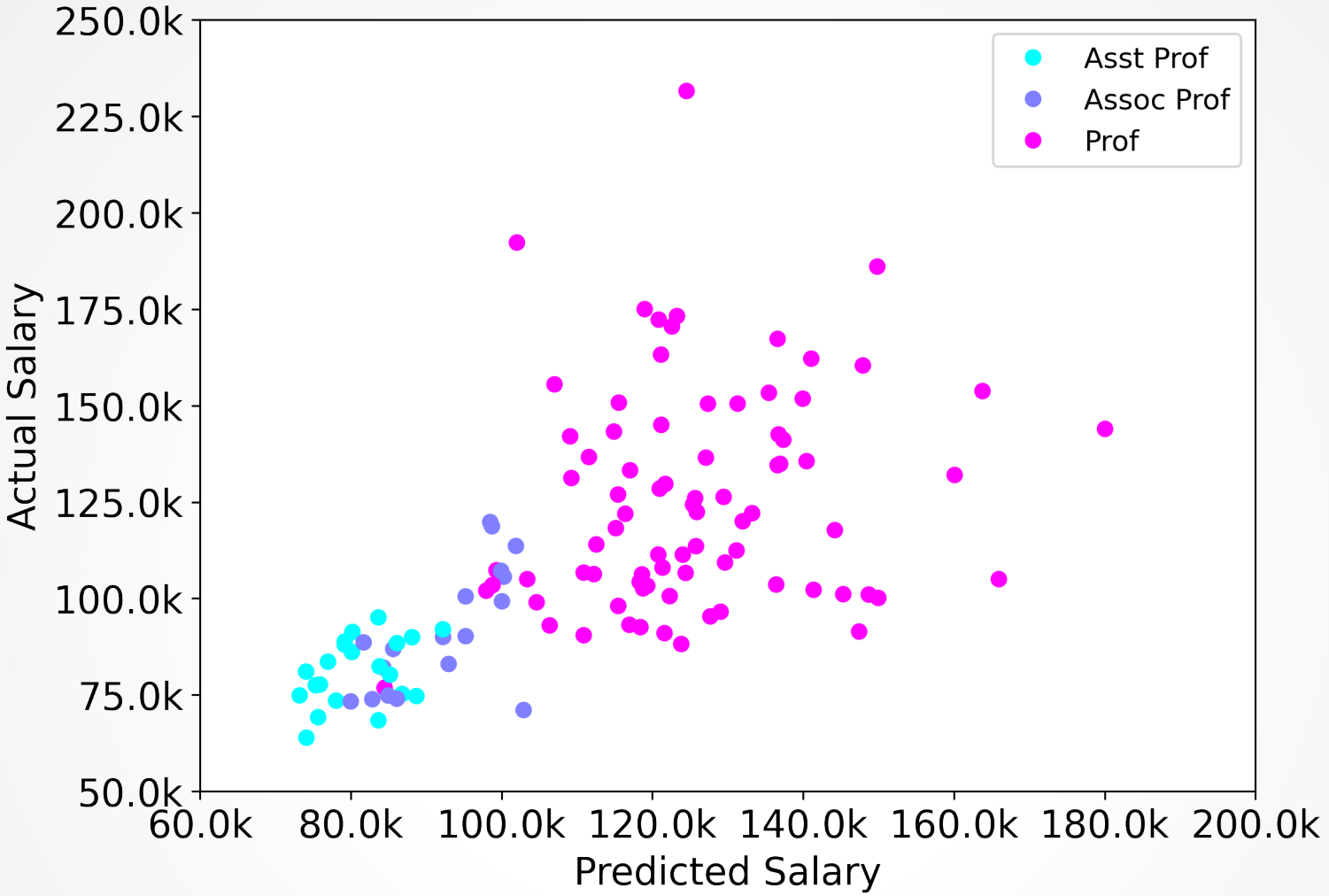
- Split on a random selection of features
- Let data determine correlations between features
- Resistant to overfitting

A section of one of the decision trees.



Trained on 70% of the data using cross-validation
Validated on the remaining 30%

Model Building

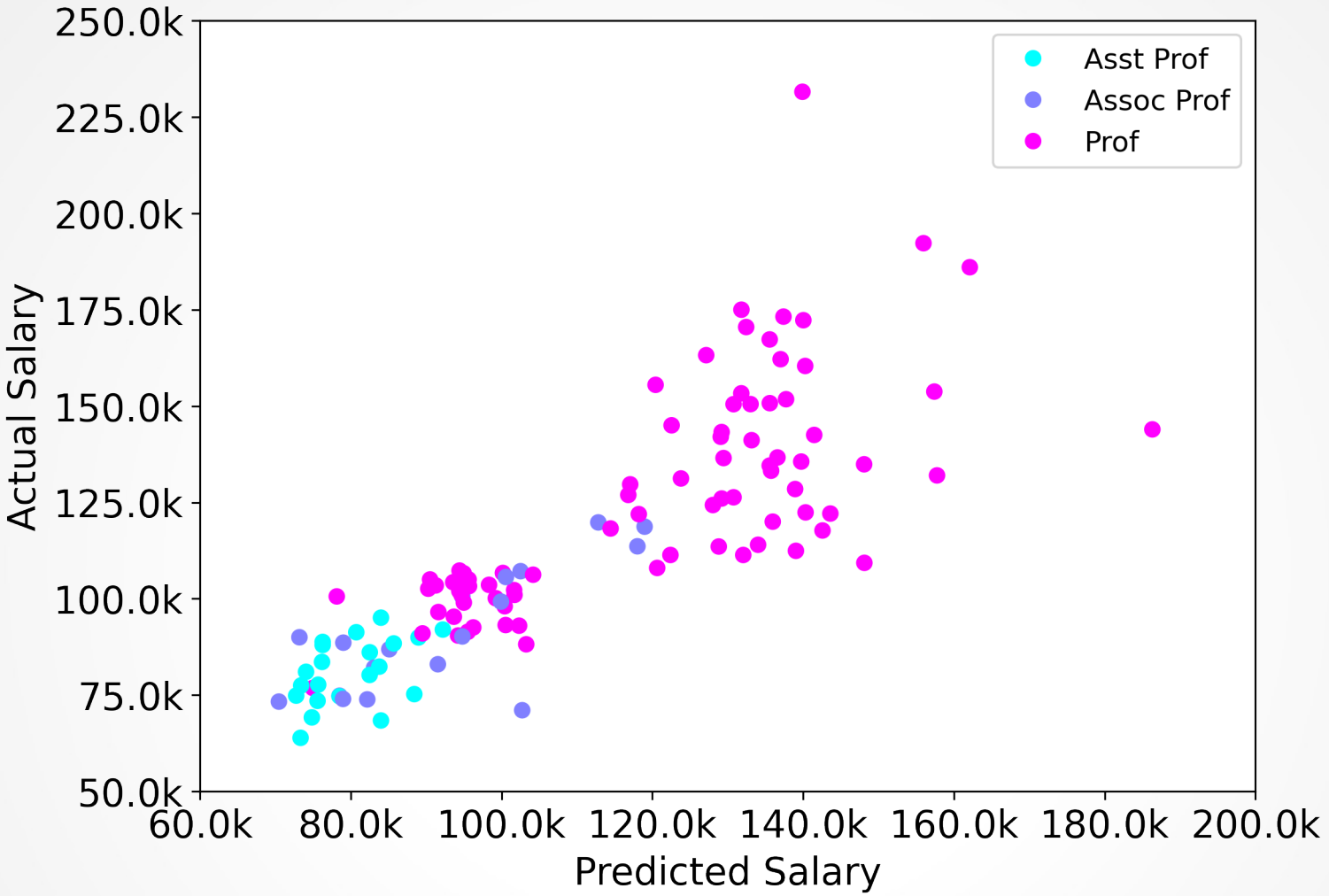


$R^2 = 0.33$

Assistant Professors salaries have low mean and variance, followed by Associate Professors. Full professor salaries have high variance and mean.

Model is effective at predicting Assistant Professor salaries, less effective for full Professors.

0/1 Salary Indicator



$R^2 = 0.66$

Vast improvement predicting Professor salaries < 113k
Doubles overall R^2 of fit.
Gap at 113k as all data points in leaves ‘know’ which side of the median they lie.

Dataset Enhancement

1) Research questions:

- Does the salary of a professor correlate to their performance?
  - Looking at the wide disparity in salaries, is there a return associated to the higher salaries of certain professors?
  - Preformance could be tracked by any of: number of publications, number of graduate students, teaching evaluations, and external grant income.

- How does the difference in salary between male and female professors vary across different regions?
  - I think it would be interesting to compare how differences in pay vary regionally.
  - Regions could include different universities within a province, between provinces, or comparative between two countries.

- How do the salaries for different ranks of professors change over time?
  - With university budget cuts in the news, I think it'd be useful to know which of the ranks of professors are most affected by cuts, and which ranks get raises over time.
  - 3 consecutive years worth of data would be interesting to examine here.

Dataset Enhancement

2) Additional attributes

- Number of publications
  - This is a typical metric to measure research output
- External grant income
  - Either a boolean (they get external funding or not), or a numeric value.
  - Some professors salary may be funded substantially by outside sources
- University Name
  - Grouping professors by type of university and location enables comparisions between larger research universities and smaller education-focused ones and examining regional variance in salaries.
- Demographic information about the Dean or Chair of the department
  - Is there a correlation between the Dean's research focus and salaries of professors in that area?
- Previous year's salaries.
  - With this data, it would be possible to check correlations between current income and the %-change in income year over year.
  - Are university cutbacks felt uniformly?

Dataset Enhancement

3) Sample Size

- Does the salary of a professor correlate to their performance?
  - As a similar feature to this assignment, a similar sample size would be sufficient. (~50 data points minimum, ~200 recommended)
  - This size of dataset was sufficient to show statistically significant differences between female (N = 39) and male (N = 358), but the findings were not significant when grouping by both rank and sex (N(female) ~15)
  - While one performance metric would be necessary, it would be helpful to consider multiple.

- How does the difference in salary between male and female professors vary across different regions?
  - Ideally the sample size would include at least 100 male and female professors. Though in this set, female professors only make up 10% of the dataset, which may mean the total sample size would need to be near 1000.

- How do the salaries for different ranks of professors change over time?
  - A similar sample size to this dataset would be sufficient. This analysis would be similar to testing the