Intersectional Demographic Regression Analysis of University Employee Salaries

Tyler Branscombe and **Edmund Chen**
Georgia Institute of Technology

Abstract

We preform a regression on individual and intersectional demographic factors, specifically predicted gender and predicted race in addition to generalized job role and employee rating of various public state university system employee salaries. Our findings suggest that some of these factors have significant predictive power regarding salary. We discuss ethical considerations in using predicted race and gender, and what its implications are on our findings. The code is hosted at the following link: https://github.gatech.edu/6471-pay/prof-pay

Introduction

University employee salaries vary widely and there are many potential reasons for these differences. Some factors are legitimate causes for differences in pay, such as research impact, tenure, or location. This project investigates demographic factors, specifically race and gender, in addition to intersectional groups across these two factors. Beyond demographic factors we also perform analysis on Georgia Tech Professor Course/Instructor Opinion Survey data to see if ratings here have significant predictive ability over salary.

To begin we collected public university system data from several states from different regions. We collected data from California, Georgia, Illinois, North Carolina, Tennessee and Texas. We cleaned these data set, conforming the field names, individual's names, position titles so that we could perform analysis on all of the data together with state as a factor. Ultimately with the data from all of these sources, our final list of factors to perform regression with was: State, Year, Institution, Salary, Predicted Race, Predicted Gender, and Generalized Role. For the Texas data specifically, we separated this data set to evaluate our prediction methods as it did include race and gender data. Because race and gender data was not available for the remaining public data sets, we use predictive models based on first and last name to predict gender and race respectively. In addition to predicting these values we also had to consolidate the position titles because each state had their own naming system sometimes with hundreds of individual titles. We chose to reduce down to Faculty, Instructors, Lecturers, and Professors. We perform more specific analysis amongst just professors scrutinizing to Professors, Assistant Professors, and Associate Professors.

With this data we perform regression analysis at three different scopes. We perform regression with data from all universities and generalized roles, data from all universities for each of our specific roles, and finally data from just Georgia Tech to perform analysis of Course Instructor Opinion Survey ratings. For each of these scopes we perform regression to predict salary, using multiple regression models and parameter tuning techniques. From these models we interpret the results by performing significance tests to evaluate each factor's predictive ability. For the variables determined to be significant we interpret the coefficients themselves to detect if there are biases against certain demographic groups or intersectional demographic groups.

Related Works

To gain adequate background knowledge on the prior work done in this space, we looked for studies that concerned similar statistical analyses on salaries for university faculties. From a high-level overview, many of the studies we found addressed specific institutions but did not consider a broader scope to compare different states or types of institutions. Many studies implied they ran into the issue of gaining enough standardized data to execute their analyses on all the factors desired, limiting their scope to venues where they already had access to a high number of data features. All in all, much of the prior work presented interesting insights on data gathering and regression techniques that helped inform our project. For instance, (Claypool et al. 2017) provided an insightful methodology in using linear regressions to implicate differences in salary for demographic-related factors.

(Konsor 2010) and (Claypool et al. 2017) used regression analysis to analyze salary based on different traits. The former specially used regional income, enrollment, tuition, and discipline concentration to explain average salary determination. With the studies that showed more granularity such as (Schrouder et al. 2019) however, they were not able to get as many standardized quantities and restricted their analyses to only one department of one university. In a similar vein, (Webster 1995) and (Cheng et al. 2019) considered a larger pool of data point but did not have many factors for each,

specifically (Cheng et al. 2019) took a survey of the American Medical Informatics Association (AMIA).

For factors past the demographical, we found studies that touched on teaching effectiveness in a limited scope. For instance, (Hoyt and Reed 1977) used an approach that considered three quantitative metrics of teaching effectiveness for 266 faculty at Kansas State University. They then used this to determine whether it affected salary percent and dollar increases, taking into account discipline partitions. They concluded that there was a modest but significant correlation between better teaching and better salaries and that such an effect was more apparent in the humanities than sciences. In another vein, (Tsikliras, Robinson, and Stergiou 2014) sought to determine whether rankings led to higher salaries. Their correlative analysis implicated an asymptotic trend for American universities but none for Canadian universities when considered in their local contexts. However, in the global contexts both showed a strong correlation, supporting their hypothesis that funding leads to higher rankings.

All in all, these studies helped us gain an idea of what approaches past work in the field had taken on. We sought to strike a careful balance between having more data so we could study and analyse a wider scope of states and institutions while also having enough standardized data where we could provide meaningful analyses. Critically, we considered the intersectional factors that none of the related studies have. For instance, we considered the different combinations of demographic factors as opposed to considering them each independently.

Ethical Considerations

One major assumption the conclusions of our paper rely upon is the accuracy of our race and gender predictions. When collecting our data, gender and race data was generally not available. Once this was apparent our options were either to not include race and gender data which would make our main goal impossible, or to use some method to predict gender and race from the data we had. We decided that we wanted to predict the race and gender and we would make sure to address the implications and ethics of this choice. To predict race and gender we came up with three options: Use a statistical approach, perform web scraping to collect employee photos and perform a computer vision analysis, or to contact the university systems and request the data.

Each of these options came with their own advantages and disadvantages. For the computer vision approach the main hurdles obviously would have been web scraping for all of these photos across dozens of individual university web pages and potentially several department pages for each university. In addition to this major obstacle, staff photos are not always available for numerous reasons, for example, if they are not offered, the employee used to work there but no longer does, the employee is relatively new, just to name a few. The consequences of this is that it would have drastically reduced the size of our data, and potentially introduced some underlying biases associated with having an employee photo online. (Serengil 2019) addresses some of the ethical implications of using or even creating Artificial intelligence demographic prediction technologies. The main surface concerns of technologies like this would of course be those of malicious use in the hands of some racist or sexist party, in addition to privacy concerns. They conclude, perhaps with some bias as a developer of these technologies, that there is nothing wrong with them being available as there are beneficial use cases, and that it should be governing bodies in charge of regulations to prevent malicious use.

The next alternative was to request data from the public university systems directly. The drawbacks of this approach are more logistical in nature, mainly being that the university systems may not have the data in the first place to give, they may not be allowed to give us the data, they may not respond at all. While this would be the most accurate in terms of labeling, it is unlikely that we would have gotten the data we needed in our time-frame. Additionally this method would have to come with additional security concerns as it is official personal data not publicly available.

That left us with our final option, a statistical approach. The drawbacks of the statistical approach are obviously accuracy and potential bias inherent in the data that is the basis for the model we use. The benefits of this approach were that it could be done relatively quickly, it would allow us to use all of the data we had collected, and would eliminate the need to clean these factors of the data for the most part.

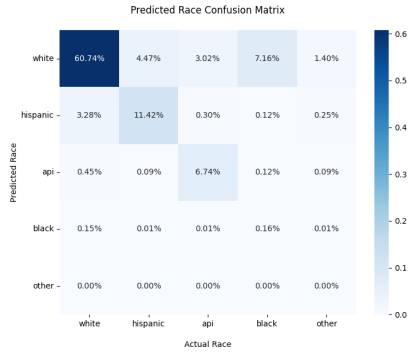

Figure 1: Race Prediction Confusion Matrix

In order to mitigate the negatives of the statistical method, we did our best to find models that had minimal apparent biases. What we found were two Python packages each one with a model for predicting gender or race. For race we used the ethnicolr package, specifically the pred_census_ln function to predict race from last name. This model is based on the 2010 census data, and this model has been updated and maintained within the past year, so we believe this is as unbiased of a data source as we can hope to get and the model is well kept, both promising sings. One thing to note is that the model outputs posterior probabilities for each name being each of the available races, but we elected to just use the highest probability race, because we could not come up with an approach to model intersectional data if our race variable was not discrete.

Race Predictor Evaluation

To evaluate the effectiveness of this model on professor data sets, we collected salary data from Texas which did include race data. To evaluate our model, we ran it on the Texas data so we had an actual and predicted race columns. We then created a confusion matrix shown in Figure 1 to visualize the effectiveness of our model. We first had to consolidate the Texas race data down into the four category outputs of our model for convenience, and also left an 'other' category for races that our model could not output to help determine how our model affects minority groups such as American Indians and Native Americans. Summing the diagonal of Figure 1 shows that the model predicts race correctly 79.06% of the time. The most significant issue is that it seems to mis-classify Black individuals as White with 7.16% of all of the data falling in this category, while generally incorrectly classifying Black individuals 94.7% of the time. Only 1.75% of the data fell into the 'other' category and the model seemed to split that data cross the four categories in a similar proportion to the rest of the data. Also worth considering is that this test set of data is from Texas and likely has a significantly different proportion across the races than the other data sets from other states. This model did not perform as well as we hoped and its biggest flaw of mis-classifying Black individuals really detracts from any findings with respect to them. This is however the best model we could find and at least classifies Whites, Hispanics, and Asians/Pacific Islanders correctly at rates of 94.0%, 71.4%, and 66.9% respectively.

Figure 2: Gender Prediction Confusion Matrix

Gender Predictor Evaluation

We performed a similar evaluation for our gender prediction methods. For gender prediction we used Python's gender-guesser package to predict gender based on first name. We chose the gender-guesser package due to the care in its data collection as they had several native speakers of multiple languages vet the data in helping determine their classification, taking into account different cultures. Additionally it's output is not on a binary scale, expressing confidence in its answer with additional categories beyond male and female being mostly male, mostly female, androgynous if the data for the model is evenly split between male and female, and unknown, if there is not enough data to make a guess. in Figure 2 we show the confusion matrix between the Texas data and the raw prediction which show pretty impressive results of 74.6% exactly correct, and 12.1% unknown. For the mostly_male and mostly_female categories, if you coerce these categories into their corresponding category as you can see in figure 3, the exact match percentage rises to 83.3% only mis-classifying 2.5% of the data. As might be expected, individuals labeled as having androgynous names are evenly split between being male and female, and a similar split is found in the unknown category, though slightly more female names are labeled as unknown. We are very happy with the performance of this gender model and we believe that the findings regarding any differences specifically with gender would be realistic.

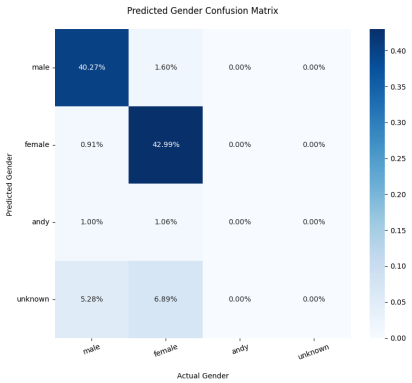

Figure 3: Consolidated Gender Prediction Confusion Matrix

We believe that our use of predicted demographic data is ethical because it is done with the intent of identifying bias, not perpetuating it, and we hope that any findings be interpreted with the knowledge that they are predicted values. It is important to note that any of our findings specifically regarding black university employees should be take with skepticism. Other things to note that could potentially be improved upon in future work is including more genders beyond male and female, and more races beyond the four output by our model, or splitting Asian/Pacific Islanders into more specific groups. Additionally our modeling does not take into account people of multiple races. We compromised these aspects in order to use more data, we hope that future works can potentially use smaller datasets with true race and gender values and use this paper to supplement any findings.

Data Sources and Collection

In order to perform our analyses, we needed to start off by collecting data from a wide range of salary data sources for a large scope of research universities. We wanted to ensure there were enough datapoints to produce a meaningful analysis as the strength of our resulting statistics depends on the associated sample sizes. As public institutions are bound by American law to release salary data under the freedom

of information act and public record laws, we targeted our analyses around large public university systems in the US. Through such means, we are able to better ensure there will be similar dimensions in all our data sources. Public universities are a critical part of American higher-education, especially at the undergraduate level. According to (Hanson and Checked 2022), 77.7% of undergraduates and 48.8% of graduate students attend public institutions in the US. By considering public institutions, we focused on the preeminent university systems in certain states in question and produced our analyses based on those datapoints.

Data Gathering

In essence, we first set off to gather data from institutions in different states and determine what other fine-grained data points we may be able to obtain about professors such as department, role, seniority, teaching effectiveness, and research impact. Drawing back to the balance alluded to in the related work section, we were careful to draw from enough data sources that we could implicate a meaningful analysis while still preserving enough standardized categories of data that we could draw from multiple states and institutions. All in all, we went through to survey which university systems reported the most similar data values and ended up gathering data for the California State University system, University of North Carolina system, the University of Illinois system, the University of Tennessee system, the University of Texas system, and the University System of Georgia.

The University of Texas system had the most granular records, with provided data values on duration of employment, department, and other salary breakdowns. The University System of Georgia unfortunately did not provide specific departments to which each employee was associated with, but all other university systems had such a column. Upon initial gathering, we either used data downloads or scrapers to systematically aggregate the data.

State	Years	Instances	Features
TN	2021	6,057	7
TX	2019-2021	65,206	11
CA	2011-2020	2,945,939	12
NC	2021	46,950	10
IL	2011-2021	195,070	6
GA	2011-2021	1,228,011	6

Table 1: Data Statistics

Across all schools possible, we sought to maintain the same data categories and dimensions where possible. Any categories which were not already standardized we took care of manually in the later data processing section.

For most of the datapoints in question, namely California, North Carolina, Georgia, Texas, and Illinois, we were able to obtain salary records through public data portals where the respective institutions published their salary data on a yearly basis. For Tennessee, since there was no readily downloadable data source, we used a website scraper to collect the data from a table that was available online. We were able to use this extension to convert 121 pages of 50 entries into 14 csv files, merging them and doing appropriate data postprocessing where needed.

Teaching Effectiveness

Another data source that would render helpful in our analysis were the teaching effectiveness ratings. As we did not have access to other universities, we were only able to gain meaningful teaching effectiveness data points for Georgia Tech. To do so, we used the Course Instructor Opinion Survey (CIOS), a semesterly survey administered to Georgia Tech students in evaluating their instructors. In order to collect CIOS data, we used a combination of a Selenium headless browser to operationally access the smartevals site, where the ratings are hosted, and use BeautifulSoup4 to scrape the rating entries off the frontend interface to the rating values. All in all, this yielded 67,816 records of distinct sections taught by professors.

Algorithm 1 CIOS Teaching Evaluation Data Gathering

Require: Python, BS4, Selenium
  Use Selenium to initialize headless browser session
  Navigate to CIOS login page and input login details
  wait until DUO Login 2FA success
  Navigate to ratings site, check success

  for each ratings category **do**
    scrape and process header column
    for each every page provided **do**
      scrape all column
      drop columns that do not abide by data format
      press next page button with Selenium
    end for
  end for

As the primary dimensionality of our analysis will be each professor, we primarily concerned ourselves with the professor table of each year in question. These tables have one record for each unique course a professor has taught, so they are not strictly one record to one professor mappings. We were able to turn this into one record to one professor mappings by using a weighted average aggreagation, seen later in the data processing section.

Overall, taking into account all the different data sources and inputs we ended up with, we had a process akin to Figure 4. In the end, everything was aggregated into CSVs for ease of version control and sharing, using SQL would have proved much more cumbersome on the data sharing and version control front.

Data Processing

After we had all data values sufficiently aggregated, our next steps concerned standardizing them and populating the columns which needed to be implicated from the provided datapoints. In having multiple data sources, we also had to match the same entity from multiple data sources together in order to build a cohesive dataset from which we could extrapolate meaningful analyses from.

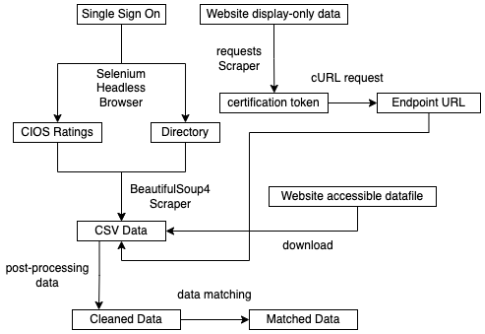

Figure 4: Data Aggregation Process

Primary Normalization and Generalization

For the primary consolidation and standardization, we sought to combine them into one cohesive dataset which could then be used for an unified analysis. In order to accomplish this we had to choose which features to select that could be taken from as many of our data sources as possible. We ended up selecting the following: *First Name, Last Name, Institution, Department, Title, Salary* across all university systems where possible. This was processed by renaming corresponding columns of the various data sources to match, checking the merged datasets for duplicate data and null/zero values and removing them. Other adjustments were also performed, for example, the Illinois and California data had a "Name" feature, which we had to split into a "First Name" and "Last Name" column, accounting for some entries having a middle initial, and people with more than one last name. We were able to account for most cases with the one notable exception occurring if someone has no middle initial and a last name that is only one letter. We performed all of these operations in python and with the Python pandas package.

Past the initial normalization, we sought to generalize certain columns so that we could perform analyses based on the categories certain professors took for each category without overwhelming with too many possibilities, which would have made our analyses hard to decipher. For instance, for the role column, we sought to only consider teaching faculty, split into categories of professor, instructor, and lecturer. Furthermore, professor would be split up into assistant, associate, adjunct, and full professors. By clearly delineating these professor and teaching roles we would be able to section our analyses to consider only similar job levels at a time. This is important since if one university was perhaps more research-based than teaching-based, then naturally there may be more full professors than, say adjunct professors or lecturers, and the analyses would be have the confounding factor of a different distribution of employee roles. In that scenario, an university who is more focused on teaching would hypothetically show up as having a worse salary distribution but perhaps only have that as a consequence of having more lower-paid staff due to their university function.

In order to accomplish this, we used a technique known as fuzzy matching. For fuzzy matching, we sought to identify keywords and other common abbreviations aided by generous margins to identify which generalized role each professor would fall into. As each of the many states we had data for included many different teaching roles, it was at times hard to determine specifically which professor category or employee category a specific record would fall into. As we mostly were concerned with half-time or above faculty, we excluded any emeritus and visiting faculty whose salary fell significantly outside the average of their colleagues' salary ranges. This was accomplished in Python.

Demographic Determination

For demographics determination, we used the first and last names of employees to map to best guess their demographic backgrounds as all except the Texas dataset did not already provide this dimension. However, we recognize there are complications with this, namely any people have names that may not necessarily correspond to the majority vote from a purely statistical view. For the process in determining demographic data and the ethical considerations wherein, we have outlined that in the ethical considerations section.

Ratings Normalization

As mentioned in our data collection section, the ratings garnered from CIOS were section by unique course each professor taught in a given year. To effectively map this to the salary data we had, we needed to reduce this to only have one rating per professor. Drawing on our own past recreational work in normalizing ratings, we determined an aggregation function that would account for the variability between the different rating subcategories and effectively implicate a final rating for each record. This final rating was furthermore taken on an average over all the professor's nontrivial courses, excluding seminars and reading courses. The final rating was then a one on one mapping from teacher to rating. The histogram of the constituent components of each rating can be seen in Figure 5, with the final aggregate rating denoted by the thick cyan line.

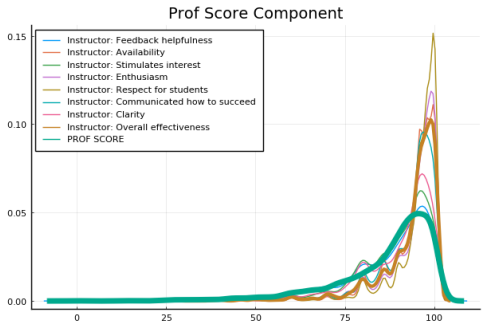

Figure 5: Professor Scoring Breakdown

Research Impact

We attempted to implicate a h-index and citation count for the professors in our datasets but ultimately could not reli-

ably populate a large enough sample without massively slicing our overall dataset size. Trying both python interfaces and scraping approaches, we utilized the Google Scholar and ORCID interfaces to garner such data. Such interfaces were heavily rate-limited unfortunately, and it took almost 20 hours to populate one of the smaller datasets, University of Illinois, pre-filtered down into University of Illinois Urbana Champaign. Upon analyzing the resulting metrics, there were not enough matched records to continue down this path. Ultimately, we believe the issue is not enough junior faculty have rendered a Google Scholar page at that stage in their career. Additionally, it is hard to exactly match researchers; for more senior professors who may have switched institutions, simply matching up the institution on an online source is insufficient as they may have moved to another institution.

Data Filtering

Lastly, in order to further hone a fair analysis between the different institutions and states, we performed data filtering before going into our final analysis. Firstly, we filtered out any non-generalized employees so that all records would be succinctly within the bounds of the roles we were concerned with. Next, we went through and did basic statistic counts of each split by department, institution, and overall data values to determine which universities we had sufficient final data to perform analyses on. In the end, we identified the California State University System, the University of Illinois System, the University of North Carolina System, The University of Tennessee System, the University of Texas System, and the University System of Georgia (limited to Georgia Tech) to perform our final analyses on. As noted above, we only used the Georgia Tech data from the University System of Georgia dataset since we only had the ratings from Georgia Tech. All in all, this left us with a concise set of generalized datasets which we could then perform regression analyses on.

Regression Methods

Once we had consolidated our data our next step was to perform regressive analysis on our data. We had 16 different sets of data that we wanted to perform regression on and wanted to see what the best model would be. Our intuition was that linear regression would be best, but it was very difficult for us to know the shape of the data for certain especially because a majority of our factors were categorical. To settle this uncertainty, we performed a fed different regression techniques on each of our 16 data sets, and would evaluate these models against one another to determine the best model for each specific data set.

Linear Regression

The first regression technique we performed was simple linear regression, which we believe would be the best model based on the dimensions and type of the data. Before we performed our regression we first transformed each of our categorical data fields into n different binary fields, where n is the number of categories for that field. What this accomplishes is that it allows the linear regression to include each category as its own variable in the regression equation (1). How the simple linear regression works is that it take each of our field, that is all of our categorical field which have been converted to binary variables, and find a linear combination of coefficients for each of these variables that minimizes the average distance of the models prediction from the actual result. One caveat with using categorical variables in a linear regression like this is that for each categorical variable, one category is selected to be normalized against. That is one category is assigned a coefficient of 0, and all of the other categories for that variable are assigned coefficient which represent changes in salary relative to the normalized category. We will come back to this when we analyze the significance of our variables.

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \tag{1}$$

Multivariate Adaptive Regression Splines

The next regression we performed was Multivariate Adaptive Regression Splines or MARS. Before we discussed how we weren't certain what the shape of our data was, so we wanted to try a model which would fit nonlinear data better in case that was the true shape of our data. The same principle for turning categorical variables into binary variables for each category applies. The difference between MARS and Linear regression is that it will parse through the data searching for "knots" at which to change the linear regression function. Once it has completed its search for these knots across all variables, it will go back and "prune" the knots which do not significantly contribute towards the predictive ability as a way to decrease over fitting. MARS is an effective model for non linear data, and can be used with categorical data and as such was a perfect candidate to apply to our data to see if there was a significant improvement over the linear model to suggest our data was in fact not linear.

Support Vector Regression

The third regression method was Support Vector Regression (SVR). SVR works very similar to linear regression, the main difference being the way that they are optimized. Where Linear regression minimized the squared error, effectively the average distance of the prediction from the actual values, SVR minimizes the l2-norm of the coefficients. The effect of this is that it puts less weight on the coefficients, which allows less bias to be caused by the training data set. In addition to less bias, in determining the best SVR equation to use, there is an additional parameter to be optimized which lets you choose tolerance for error, which allows this regression to balance good fit with reduced bias. We elected to attempt using this regression method to see if our other models were over-fit to the training data, because if they were, the performance of a SVR would notably outperform them. Unfortunately one drawback of SVR was that it is not very time efficient optimizing the additional parameter. For this reason we were unable to perform SVR on our large data sets, specifically the comprehensive data set with data

from all states, and our lecturer data set, and their respective intersectional data sets.

Logistic Regression

Our final regression was Logistic Regression. Due to the fact that were were uncertain oft he shape of our data, we wanted to include one more additional method to test for another non-linear shape to see if it fits better. With Logistic regression, we would expect a good fit if there was more for an S-shaped curve (or sigmoid function) rather than a line. What this can be interpreted is if there is some specific set of values that change, after which salary significantly increases. It accomplished having this shaped fit by instead of starting with a simple linear equation and optimizing it, it optimizes (2). And important aspect of this function to note is that when performing regression with it rather than classification, you must regularize your dependent variable between 0 and 1. We did this only for this function as it was necessary and doing it for the other functions would not change performance as everything is still scaled the same, but it would add an additional step in the end when we want to derive meaning from the coefficients.

$$y = \frac{1}{1 - e^{-(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)}} \tag{2}$$

To evaluate our model before we trained all of our models, we split the data into a training and a test data set. We divided the length of each of our sets of data such that 90% would be training data and 10% would be test data. We decided on a 90/10 split because some of our data sets were relatively small due to how many partitions we made into the data. Our Georgia Tech data for example had under 3000 data points, so we wanted to maximize the amount of data used to train our models for our smaller data sets, while still leaving enough data to effectively test and compare our models, and with these goals in mind we selected the 90/10 split. Once we had split the data, we trained 3-4 models on each of our data sets, and we collected some measures on the models effectiveness using the reserved test data. The metrics we used to evaluate our models were R-Squared, Root Mean Square Error, and Mean Absolute Error.For each of our data sets we compiled these metrics for each model into a data frame and manually evaluated these metrics against one another to determine the best model for each data set.

Analysis

Once we have created all of our models, the next step was to evaluate them against one another to determine the best models for each of our data sets. For each of our 16 data sets we created a table like the one shown in Table 2 which is for our Assistant Professor non-intersectionalized data, of course with the exception of our larger data sets not including SVR.

Now with theses tables we identify the ”best” model for each by selecting the model which led in at least 2 of our metrics. There were no instances where 3 different models led in each metric so we did not have to come up with a contingency plan for this. For all of our models we found

Table 2: Model Comparison for Non-Intersectional Ast. Prof.

	R-Squared	RMSE	MAE
Linear Regression	0.3471	40279	28177
MARS	0.3442	40368	28347
SVR	0.3250	41186	27348
Logistic	0.3330	2.4540	2.4290

Table 3: R-Squared for Best Model for Each Data set

	R-Squared	Intersectional R-Squared
All Data	0.3931	0.3964
Professor	0.2125	0.1988
Asc. Professor	0.2771	0.2930
Ast. Professor	0.3471	0.3451
Faculty	0.5184	0.5773
Instructors	0.5440	0.5150
Lecturers	0.0416	0.0345

that the linear model was the best fit, with the exception of our Instructor data. In this case the edge SVR had over Linear regression was very small, and for this reason combined with the fact that every other data set's best model was linear, suggesting that the general shape of our data truly was linear. This means that it is possible that some random variance caused SVR to perform better. These two data sets were most susceptible to this as they were some of our smallest. Switching from our data set with gender and race separated to our intersectional data sets made no differences in which model was the best, and as shown in Table 3 had no consistent effect on the R-Squared performance of the linear model for each data set. All of our data had moderate success in explaining the variance of salary with the exception of our lecturer data with R-Squared values of 0.04 and 0.03. For this reason we did not perform further analysis of the Lecturer data set, as it is clear our data is not explanatory towards salary. As for the rest of the data now that we know our data is at least somewhat explanatory of salary, we can look at our coefficients of interest.

Significance Analysis

First lets evaluate the results for our data sets which had race and gender separate. In Tables 4, 5, and 6 we have our data sets on the left and the columns marked with '*' indicate which categories were identified as significant at a 95% confidence interval, or a p-value of less than 0.05.

Table 4: Significance for Race Variables

	Hispanic	API	Black
All Data	*	*	*
Professor	*	*	
Asc. Professor		*	
Ast. Professor			
Faculty			
Instructors			

Table 5: Significance for Gender Variables

	Female	Mostly Female
All Data	*	*
Professor	*	*
Asc. Professor	*	*
Ast. Professor	*	*
Faculty		
Instructors		

Table 6: Significance for Gender Variables cont.

	Male	Mostly Male	Androgynous
All Data	*	*	*
Professor	*		
Asc. Professor	*		*
Ast. Professor			
Faculty			
Instructors		*	

If it is marked as being significant at a 95% confidence interval it indicates that there is a greater than 95% likelihood that the variance explained by that category is not due to random chance. This in turns means we can say with relative confidence that that category has significant predictive ability towards salary, and in the case of these demographic variables that there may be bias towards these groups. An important note before looking at these tables is that for each categorical variable one category must be selected to be normalized against. This means the coefficient of one category must be set to 0. A consequence of this is that with a coefficient of 0 it cannot explain any variance and cannot be identified as significant. This also means that the variables marked with "*" indicate significance relative tot his normalized category. More explicitly, the categories labeled with a "*" mean that there is a greater than 95% likelihood that the variance explained by that category relative to the normalized category, is not due to randomness. With the intent of our paper being identifying biases towards minority groups, we selected 'white' as our normalizing category for race, and for gender we wanted to select one of androgynous or unknown as both had an even split on men and women, so we selected those identified as unknown. It is because of this that the White and Unknown categories are excluded from Tables 4, 5, and 6.

Next we will perform the same analysis on our intersectional data. The results of our analysis are shown in Table 7. For our intersectional significance testing we removed the unknown categorizations of gender (which also include the androgynous labels) as they were a minority of the cases and we did not feel conclusions drawn from their significance would have useful interpretation. We selected male-unknown as the normalization category for our intersectional data under similar reasoning to our selection or normalization categories in the segmented race and gender data. This being said, Tables 7 and 8 shows the intersectional groups and those marked with an "*" indicate a 95% likelihood that the variance explained by that category relative to the male unknown category is not due to randomness

Table 7: Significance for Intersectional Variables

Male				
	White	Hispanic	API	Black
All Data	*	*		*
Professor	*			
Asc. Professor			*	
Ast. Professor	*	*	*	
Faculty				
Instructors				

Female				
	White	Hispanic	API	Black
All Data	*	*	*	*
Professor	*	*		
Asc. Professor	*			
Ast. Professor	*	*	*	
Faculty				
Instructors			*	

Table 8: Significance for Intersectional Variables cont.

Coefficient Analysis of Significant Variables

Taking a look at which variables were determined to be significant,it is pretty evident that there is not much significance between gender and racial factors with respect to faculty and instructor roles. If we take a look at the professor, associate professor and assistant professor roles however, we not only see significance on some of these factors, but we see significance on the same categories in each, notably all three found significant explanatory value in the female and mostly female categories in the data with race and gender separated, and all three also found significance in the white-female category and two of the three in Hispanic-female category. To get a better grasp of the significance of these variables, we took the coefficients of the significant race and gender variables for both of our data sets and plotted them for each of the professor roles.

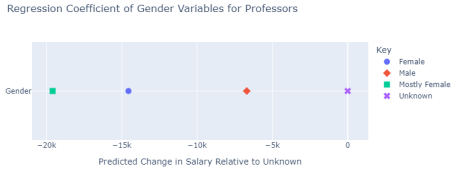

Figure 6: Professor Gender Regression Coefficients

Beginning with the gender data from each of the professor roles, in Figures 6, 7, and 8 we see that female and mostly female categories are consistently to the left, significantly below the regularization category, and in Figure 6 and 8 they are also below male and or androgynous categories. Also worth noting is that the Unknown gender category is

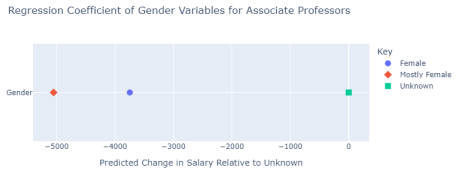

Figure 7: Asc. Professor Gender Regression Coefficients

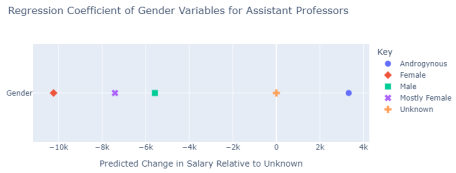

Figure 8: Ast. Professor Gender Regression Coefficients

predicted to be paid more than any other significant category with the exception of androgynous in Figure 7. From this we can conclude than People with predicted female or mostly female names earn significantly less than those with unknown predicted gender across all generalized professor roles.

Looking at the gender roles for each professor role (except Assistant Professors which had no significant race categories) in Figures 9 and 10, the most clear conclusion is that professors who are Asian/Pacific Islanders earn more than White professors, predicted to earn close to $5000 more, holding all else constant. Further, in the case of Hispanic professors, they earn more than both white and Asian/Pacific Islanders.

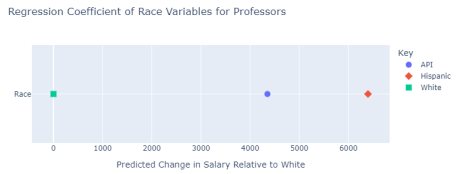

Figure 9: Professor Race Regression Coefficients

Finally we can look at our intersectional models for each professor role in Figures 11, 12, and 13. With these we see that white females are consistently to the left of the chart in all three with the exception of regular professors, where they are very close with Hispanic Females. This signifies that relative to the other present intersectional groups on their respective charts, they are consistently predicted to earn the least. Looking closer at Figure 13, which nearly all intersectional groups were significant we see another interesting pattern. From left to right the points in Figure 13 are White female, Hispanic Female, API Female, Hispanic Male, White Male, API Male, and finally the normalization category, White Unknown. This would seem to suggest that

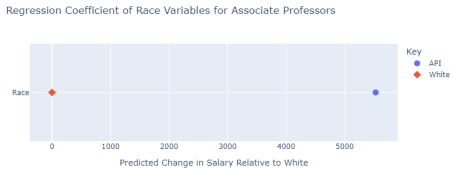

Figure 10: Asc. Professor Race Regression Coefficients

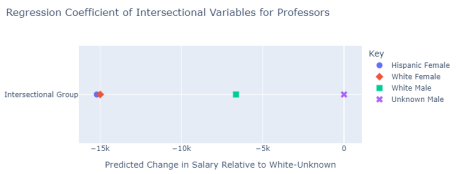

Figure 11: Professor Intersectional Coefficients

the highest paid woman race is predicted to be paid less than the lowest paid male race. Additionally amongst each gender, Asian/Pacific Islanders, are paid the most.

Georgia Tech Ratings Analysis

Now that we have gone over our demographic analysis, we can look into our analysis of the Georgia Tech Data, looking at how Course/Instructor Opinion Survey ratings perform in predicting salary. For the Georgia Tech data we included the intersectional demographic data. Following a similar analysis process as we did with our demographic data we begin my looking at the performance of our four regression models on the Georgia Tech data. Like with the Instructor data, the SVM was the best model, by a very slight margin over Linear regression. Again we decided to select the Linear regression model due to the fact a significant majority of the other data sets were linear, giving us confidence this data would be too. The statistics of our linear models with the Georgia Tech that are shown in Table 9

Right in the same area of success as our other data models, based on the R-Squared value our model is moderately successful in explaining the variance of the data. So we feel comfortable progressing to analyze the significance of the variables. Our single variable of interest in this case is rating, and with a p-value of .0055, we can say that at a 99% confidence level gender is significant. More explicitly there is at least a 99% likelihood that the predictive ability rating's coefficient has on salary is not due to random chance. Now that we are confident that rating is significant we can look at the value itself. The coefficient for rating in this model is 6821.08. Once interpreted this suggest that for each unit rating increases, each additional point a professor gets on

Table 9: Significance for Intersectional Variables cont.

	R-Squared	RMSE	MAE
GT	0.291335	52133	37818

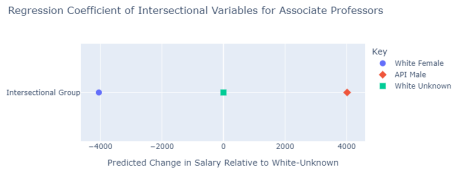

Figure 12: Asc. Professor Intersectional Coefficients

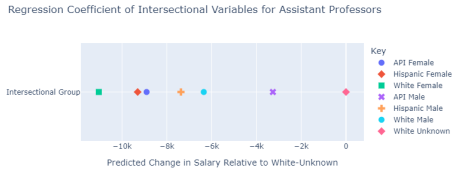

Figure 13: Ast. Professor Intersectional Coefficients

their Course/Instructor Opinion Survey rating, their salary is predicted to increase by nearly $7,000.

Work Division

Tyler was responsible for the gender and race prediction and evaluation, modeling, model selection, coefficient visualization. Edmund was responsible for data aggregation and processing, rating normalization, data partitioning, and data pipelining. We both worked on all written deliverables for the project.

Conclusion

Overall, looking at our analysis, our goals going in were to identify differences in salary due to gender, race, or intersectional factors. Based on our findings it seems that we can confirm the findings of other studies that female professors get paid less, and we can add the fact that this is consistent across various subdivisions of the professor title. Further discussing the intersectional data, our most apparent finding was that white females specifically get paid less across all three professor roles. We also found with our Georgia Tech data that Course/Instructor Opinion Survey rating did have a significant, positive predictive effect on salary.

Again it is important to take the findings of this study with some skepticism as the models were trained on data with race and gender data only being about 80% accurate. Other potential sources for bias in our data include biases in the consolidation process of reducing the varied role titles to our restrictive categories. Additionally with only 5 factors, the model may have been too rank-deficient to create effective models, and some of the variance we found to be explained by race or gender could in actuality be explained by one or many factors that we were unable to include in our model.

Come directions future research may be able to follow would be to use real race and gender data to eliminate the need to put qualifiers on our results. Additionally they could collect additional data to help reduce the issue of rank deficiency. Further other methods could be used to evaluate significance, or a better more meaningful normalizing term could be used in the regression. We hope this paper can lay a groundwork for investigating intersectional groups with respect to salary.

References

Cheng, Y.; Mohanty, A.; Ogunyemi, O.; Smith, C.; Leroy, G.; and Zeng, Q. 2019. 2018 salary survey of amia members: Factors associated with higher salaries. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2019:275–284. Publisher Copyright: ©2019 AMIA - All rights reserved.

Claypool, V. H.; Janssen, B. D.; Kim, D.; and Mitchell, S. M. 2017. Determinants of salary dispersion among political science faculty: The differential effects of where you work (institutional characteristics) and what you do (negotiate and publish). *PS: Political Science & Politics* 50(1):146–156.

Hanson, M., and Checked, F. 2022. College enrollment statistics [2022]: Total + by demographic.

Hoyt, D. P., and Reed, J. G. 1977. Salary increases and teaching effectiveness. *Research in higher education* 7(2):167–185.

Konsor, K. J. 2010. *Determinants of Professor Salaries at Elite Liberal Arts Colleges*. Ph.D. Dissertation, Baylor University.

Schrouder, S.; Allen, M.; Rhodd, R.; and Jones, T. 2019. Evidence of faculty salary differences across business disciplines and employment contracting systems. *International Journal of Accounting and Financial Reporting* 9:469.

Serengil, S. I. 2019. Race and ethnicity prediction in the perspective of ai ethics. https://sefiks.com/2019/11/10/race-and-ethnicity-prediction-in-the-perspective-of-ai-ethics/. [Online; accessed 2022-04-29].

Tsikliras, A. C.; Robinson, D.; and Stergiou, K. I. 2014. Which came first: the money or the rank? *Ethics in Science and Environmental Politics* 13(2):203–213.

Webster, A. L. 1995. Demographic factors affecting faculty salary. *Educational and Psychological Measurement* 55(5):728–735.