Evaluating Employment and Salary Predictability Based on Various Numerical and Categorical Features Using Machine Learning (DS 3001)

Brittany Asare *[1] **Simone Minor**[2] **Hewan Kasie**[3]

Abstract

The labor market in the U.S. shows persistent disparities in employment access and earnings, especially among recent college graduates. This project investigates how race, gender, and field of study influence two core labor-market outcomes: (1) the likelihood of being employed and (2) annual salary among employed individuals. Using microdata from the IPUMS integrated database, sourced from national surveys including the NSCG, SDR, NSRCG, and ISDR, we analyze a large dataset consisting of demographic, educational, occupational, and income information.

Our analysis addresses two questions. First, examining the extent to which race, gender, and field of study predict employment status (LFSTAT). Second, among individuals who are employed, factors related to their annual salary. To answer these questions, we conducted extensive data cleaning to recode missing values, convert categorical variables into usable formats, and account for logical skips common in large survey instruments. Exploratory Data Analysis, using stacked bar charts and Sankey diagrams, helped identify patterns and potential relationships between demographic characteristics and labor-market outcomes.

We implemented four predictive models: LASSO regression, decision tree regression, random forest regression, and k-nearest-neighbors regression. One-hot encoded and numeric fields were median-imputed and standardized when necessary. LASSO provided linear predictors, decision trees revealed hierarchical interactions, random forest regression delivered the performance, capturing nonlinear relationships, and the KNN model served as a nonlinear baseline, but struggled with the dataset's size, heterogeneity, and noise.

Initial findings indicate meaningful variation by race, gender, and field of study in both salary levels and employment likelihood. Field of study differentiates outcomes, particularly in high-earning STEM disciplines, suggesting structural differences in occupational access and compensation. Despite the strong performance of the random forest model, the moderate predictive power across all models reflects the complexity of labor-market processes, which are shaped by variables not captured in survey data, such as regional labor conditions, employer practices, job search networks, and negotiation behaviors.

Overall, this study illustrates the continued role of demographic and academic factors in shaping economic opportunity. By combining large-scale survey data with machine-learning methods, the project offers a nuanced understanding of employment access and salary inequality among college-educated adults in the United States.

1. Data

1.1. Data Overview

Based on personal interest in the current economic state of our nation and as graduating fourth-years in college, we decided to take a dive into the current job market and the difficulties that it portrays for those who are in search of employment. We extracted the data utilized in this project from Integrated Public Use Microdata Series (IPUMS) and got our data from a variety of sources, including surveys from across the last few decades, such as National Survey of College Graduates (NSCG), Survey of Doctorate Recipients (SDR), National Survey of College Graduates (NSRCG), and International Survey of Doctorate Recipients (ISDR). Then we combined the various variables we selected to compose one larger dataset with both numeric and categorical variables.

1.2. Key Variables

The main variables included in the dataset were the following: technical skill, demographic information, highest degree achieved, employment, occupation, employer characteristics, income, job characteristics, job satisfaction, and career path jobs. Each variable is a part of the larger story of employment and earnings, but we isolated certain variables in order to answer this project's question, like race, gender, field of study. These variables are essential to our research question – what are the labor market outcomes for college graduates by major, race, and gender – because they will better explain the reasoning as to how jobs are found, the capacity in which the job is described, and the feasibility of its accessibility to those who are in search of employment. We are also then able to cross-reference those variables with gender, race, and major of the recent graduate to analyze at which point the labor market is favorable towards certain demographic groups and what is expected from domains. The job market has become increasingly difficult to navigate, especially for new college graduates, as its requirements have become more specific. As a result, many job seekers struggle to find matching criteria that meet their needs. By identifying these variables, we are able to analyze their effects on the labor market selection and requirements, and their search results for those who are using it to find employment.

The population is those featured in the highered_00001.csv dataset pulled from the IPUMS database. The dataset includes relevant variables, such as labor force status (LFSTAT) and earnings (SALARY), which will contribute to our ability to predict outcomes. The two outcomes of focus are employment which is a discrete binary indicator that relays if someone is employed full-time, part-time, or not employed. Salary is the continuous outcome being predicted which is measured in United States dollars earned annually. The key predictors being evaluated are race and ethnicity (RACETH), gender (GENDER), and field of study which includes broad major group (NDGMEMG) and exact major (NDGMED). The fields outlined by the data are STEM, education, health, business, and humanities, and these are the fields of study considered in this paper. The elements we are controlling for to focus on the effects of the predictors outlined previously are age (AGE), highest degree earned (DGRDG), number of hours worked weekly (HRSWK), and sector of occupation (EMSEC).

1.3. Challenges with Data

One of the main issues with reading the data was that the dataset was so large. The size of the dataset continued to be an issue throughout this project because it was a lot of work to clean, prepare, and run the models on. In addition, because of how many categories were in the dataset, any encoding done will make the features high dimensional. In anticipation of this, we looked more into tunning and hyperparameters to ensure to lighten the computational load that would arise for a model, like kNN that is already very computationally intensive.

In processing this data, the issues that we ran into consisted of recoding the missing data per variable. As we know, for many datasets, there are instances where the category goes unanswered per entry, or there was no information provided. In order to address these concerns, we first had to identify the areas of the variable that resulted in either NaN or missing values. We would then go on to recode this area in each variable to ensure that it was no longer a string and rather a numeric value that would then be able to be used as a point in our analysis, while not skewing our data.

Some limitations that may affect our analysis of the data are logical skipping points that were consistent through the data from the questions that respondents bypassed based on their previous responses. The "Logical Skips" are normally coded by a number, and it would be inaccurate and a misrepresentation to drop this data, but we are unsure of how it will affect the relationships we are measuring in our data.

Another issue we anticipated with this data is that the proportion of employed and unemployed people may vary, which would make it difficult to split the training and test sets because they might be drastically different proportions. In anticipation of this, we plan to stratify the split, so the proportion was similar for both sets across training and testing.

Lastly, we noticed part of the picture was missing with this data because none of the variables on IPUMS or in this dataset look at work experience or where these people are located. This is one of the downsides of self-reported surveys. There is critical information that we think would be useful in making predictions but it is not present because it is a dataset compiled from four different surveys none of which had questions that captured this kind of data nor that collected data from employers about their workers resumes and such. With this in mind, we expect and acknowledge that these models will automatically be associative and not causal in regards to the relationship uncovered. In addition, we assume there is probably more to the story and will reflect more on this in the conclusion for further research suggestions.

1.4. Exploratory Data Analysis (EDA)

Before building any models, we spent time cleaning and exploring the dataset to understand how the key variables behaved and what patterns may show up later during the predictive analysis. Using IPUMS codebook, we recoded all special numeric values such as 0, 9,999,998, and 9,999,999, as missing since they would represent "not applicable," "don't know," or top-coded values that would skew our results if treated as real numbers. After, we created an EMPLOYED indicator from LFSTAT and defined two separate samples to analyze, first being for modeling employment, and the second being for modeling salary among respondents were employed and had valid salary information.

Within the salary data that was cleaned, we examined its distribution using boxplots and histograms. The salary variable showed right-skewness, which could be expected with most people's salaries being clustered in the lower ranges and a small number of high earners. We could see a more symmetrical distribution when we transformed the salary by natural logarithm and arcsinh. We also used IQR-based winsorization methods to cap the extreme salary values. We then went on to compute the descriptive statistics for all key variables. The continuous variables, such as age, hours worked, and salary, were summarized by means, median, and percentiles. Categorical variables were calculated by counts and proportions. In discussion of these grouped summaries, patterns of STEM and professional majors had noticeably higher medians, and minorities tended to earn slightly less than non-minority graduates. Through these summaries, we were able to see the growing gaps in salary differences across demographic groups.

Our visualizations helped make these patterns even clearer. Histograms and KDE plots show how transformations affect the shapes directly. Boxplots showed that there were shifts across salary groups based on gender, minority status, and major groups. We also see that within bar graphs and scatterplots, the log of salary against age and hours had a mild positive relationship. We also see that the correlation matrices showed how these continuous predictors correlate.

Overall, the EDA provided several key takeaways. Salary's skewness confirmed the need to work on the log scale. Field of study appeared to be one of the strongest determinants of earnings, with certain majors having an upward association. Degree level and hours worked were associated with higher salaries, but they did not stop disparities from being present. Race and gender differences persisted even within similar degrees, which would suggest that these variables should be interpreted. Lastly, employment rates were not uniform across groups; modeling employment as its own outcome made sense rather than just relying on only salary amongst the employed.

2. Methods

2.1. Methods Overview

Classification is better for employment because employment is a discrete category that has a binary. Whereas regression is better for continuous numerical values, like salary because it can better estimate quantities along the logarithmic scale. Due to the two-fold question, a two-fold methods approach was required. The methods used in this paper were divided into preprocessing training procedure, employment classification, and salary regression. Below are the models that will be used for employment and earnings:

Employment Classification

- Logistic LASSO
- Decision Tree Classifier
- Random Forest Classifier
- k-nearest Neighbors (kNN) Classifier

Earnings Regression (for log salary)

- LASSO regression
- Regression Tree
- Random Forest Regressor
- k-nearest Neighbors Regressor

As noted above, the same general models are used for both outcomes we are predicting for with slight variance of the method as a classifier if it is for the discrete outcome or a regressor if it is for the continuous outcome.

In regards to predictors, all the models will use the same set of predictors. The predictors can be divided into two main categories: demographics and education. Demographics, include gender, race and ethnicity, minority status, and age. Education, included, general/grouped field of study and highest degree obtained. The other variable involved in the regression models is one's number of hours worked a week.

The general pipeline for this machine learning project is as follows:

1. Data collection: gathering raw data from IPUMS database compiled from four different national surveys.
2. Transforming data: using the codebook to recode special and missing values.
3. Preliminary Processing: one-hot encoding categorical variables and standardizing numeric variables to account for discrepancies.
4. Split the data: split the data into training and test sets.
5. Cross-Validate: use cross-validation to tune the model hyperparameters on the training set, including penalty level, tree complexity with depth and leaves, and k in kNN.
6. Compare models: Compare the performance of models based on the testing sets.
7. Interpret the results: interpret the models in the context of the EDA based on the real-world college graduate employment transition.

2.2. Model Training Procedure

To train this data, first all categorical predictors, such as gender, race, minority status, field of study, and degree level, had to be transformed using one-hot encoding to make them into one-hot indicator variables. (James et al., 2023) Whereas, the continuous predictors, like age and hours worked, were standardized to have mean zero and unit variance. This involves Z-score scaling, which is useful because it puts variables with different units on the same scale to allow for direct comparison. (Scikit-Learn, n.d.). In order to do this, the scikit-learn library was used, more specifically, the ColumnTransformer and Pipeline were implemented to preliminarily process everything the same for each model.

The data was split into a 70:30 ratio of 70% training set and 30% testing set for both outcomes. For the employment models, the split was stratified, so the proportion was similar for both sets across training and testing. The model tuning was done as a means of reducing the load on the system and creating a more efficient execution of the models. All of the model tuning and cross-validation were done on the training set and then the test set was used late for the final performance test.

After this preprocessing, the models were then each trained respective to their needs. Both LASSOs were fitted based on optimalization using the L1 penalty. The penalty strength was decided by the cross-validation on the training set. In fact, cross-validation was used throughout the models to choose the best hyperparameters within each model for each outcome. The decision trees were developed with constraints to control for overfitting. Random forests for both were trained as sections or layers of trees and each section was fitted to bootstrapped new samples made from the training data. Lastly, the two k nearest neighbor (kNN) models ingested the training data and made predictions by averaging kNNs in the feature space [Chapter 8, ISLP]. Lastly, each model then preformed on a test set that was held-out to demonstrate an honest performance that could be evaluated.

2.3. Model Validation Plan

As illuded to in the previous method sections, to validate the models we will use cross-validation for hyperparameter tuning within each specific model to make it perform at its best. One way we do this is by using the k-fold cross-validation. Within both of the LASSOs, we search over the grid of penalty strength alpha values and select the value that minimizes median cross-validated error the most. For regression, the lowest RMSE indicates the better-fitting model and will generalize the unseen test data the best from each model is chosen. For the kNN models, we scanned over the grid of k values and selected the k that yielded the best median cross-validated performance relative to the type of outcome the model is trying to predict. For kNN the employment accuracy was used and for salary RMSE was used.

Where possible, for the LASSOs, decision trees, and forests a 5-fold cross-validation (CV) was done, however due to the computational burden of kNN on a large dataset, we used a 10 percent random subsample of the training data to tune the models and instead of a 5-fold CV used a 3-fold. Finally, we refit the selected k on the full training set for the final evaluation. In short, the validation process involves tuning the hyperparameters based on the cross-validation performance of the training set we select the best model for each one and then refit the model on the full training set using those parameters, and then finally evaluate the performance of each model on the held-out test set and compare those metrics.

To compare the various models between classification and regression, and confirm that the trends for the best models were correct, we found the F1 score and Area Under the Curve (AUC). Appendix E details these findings (James et. al, 2023).

2.4. Model Implementation

We implemented all of the models in Python utilizing the scikit-learn library and following the previously mentioned pipeline with two main steps including prepping the data by one-hot encoding (discrete) or standardizing (continuous) and then doing the respective steps for that specific model based on the outcome it is predicting: classifier for discrete, like DecisionTreeClassifier and regressor for continuous, like RandomForestRegressor.

Employment Classification Models

LASSO Regression The first model employed was the logistical LASSO regression with an L1 penalty. The model estimates the probability of being employed while the L1 penalty on the coefficients to force coefficients to zero thus reducing overfitting. Then, we tuned the strength of the penalty using 5-fold cross-validation on the training set and searched over a small grid of values for the inverse penalty parameter C [Regularization, Cross-Validation Lectures].

Decision Tree Classifier We fit a classification decision tree with a constrained maximum depth between six and eight and a minimum leaf size of 200 observations to avoid very deep, high-variance trees. The decision tree classifier allows for nonlinear effects and interactions. For example, we combined field of study and degree level, but it remains more interpretable than other flexible models that may account for more noise.

Random Forest Classifier The third model was a random forest classifier. The random forest tree with 200 trees and a minimum leaf size of 100. The random forest reduces variance by averaging the predictions of multiple trees and provides important measures for variables that will help identify which predictors are most influential for our discrete variable, employment.

k-nearest Neighbors (kNN) Classifier The last model used for the discrete variable is the kNN as a standard of comparison that does not make assumptions about underlying data distribution. As mentioned earlier, the dataset is so large that we refined the tune of k to do a 3-fold cross-validation on a random 10 percent subsample of 70 percent of the data, the training portion, opposed to the original 5-fold cross-validation due to time constraints. To boost efficiency on the running time, the grid was reduced from a range of 1 to 40 to these k values: 1, 3, 5, 7, 10, 15, 20, 30. Then, the classifier was refit with the chosen k on the full training set of 70 percent and evaluated the performance of this k on the test set, 30 percent of data.

Log Salary Classification Models

LASSO Regression The first model employed was the logistical LASSO regression. The model is a linear model fit for log salary. This model also utilizes an L1 penalty on the coefficients. The penalty parameter alpha was selected using a 5-fold cross-validation on the training set from a grid of alpha values on a log scale ranging from 0.001 to 1. By doing this, we identify what the most important predictors are and simultaneously shrink the less influential ones to zero [Regularization CV Lectures].

Regression Tree The regression tree was fit for log salary with a maximum depth equal to eight and a minimum leaf size of 200. This model represents nonlinearities and interactions between predictors, like degree level, minority status, and field of study.

Random Forest Regressor The random forest regressor was fit with 200 trees and a minimum lead size of 100. The model averages across many trees in order to reduce variance and often improves predictive performance relative to a single tree. In addition, it provides a robust and complex benchmark and highlights the measures of importance for earnings, the continuous variable.

k-nearest Neighbors (kNN) Regression Similar to the kNN classification, the k was tuned to a 3-fold cross-validation on a random 10 percent subsample of the training data, opposed to the original 5-fold cross-validation for efficiency. This time, the k was chosen with the lowest median cross-validated root mean squared error (RMSE). The model was then refit with this chosen k and it was used on the full training set before being evaluated for performance on the test set.

Model	Test Accuracy
Naive Model	0.883889
LASSO Classifier	0.890980
Decision Tree	0.902476
Random Forest	0.902262
kNN	0.911791

Table 1. Test-Set Accuracy for Employment Models. This table reports the test-set accuracy for the four employment classification models; It predicts employment status (employed vs. not employed) via logistic LASSO, decision tree, random forest, and k-nearest neighbor. All models use the same set of demographic and educational predictors.

2.5. Evaluation Benchmarks

The benchmark is meant to demonstrate if the more complex built models outperform the baseline models and to indicate if fitting is an issue based on variance in accuracy between training and test sets. That said, we compared each of our build complex models or ensembles to a naive Bayes model which tends to oversimplify feature independence (James et. al., 2023). If the naive model performs better than the complex one, it is an indicator of the fact that the features may be more independent than we thought.

The evaluation for the employment classification was based on test-set accuracy and confusion matrices. Accuracy represents the measure of correct predictions on the test set. Confusion matrices are utilized to comprehend how many true positive and negatives and false positives and negatives are present. Cross-validation was performed on the training set to choose hyperparameters. At the end, the test set was used to perform an evaluation of performance.

All of the salary regression models used root mean squared error (RMSE) as the scoring metric for the test set. RMSE works well for the log-salary scale and is directly interpretable and comparable across models. The lower the RMSE, the better the accuracy of the prediction for the model.

3. Results

3.1. Results Comparison

The test-set accuracy ranges from 0.883889 to 0.911791, with the Naive Model and the kNN respectively indicating that all models perform better than a naive majority-class classifier. Because our kNN model performed better than the Naive Model, the benchmark measurement, this shows how well our data did despite the differences across data sets being modest. What Figure Y above does not show is that this difference is marginal with such a small range of accuracy scores. This suggests that the relationship between the predictors and employment status can be captured rea-

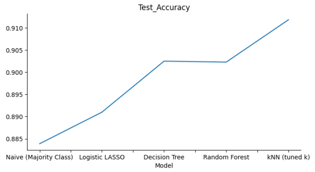

Figure 1. Comparison of Employment Classification Models' Test Accuracy. This line chart height represents a test-set classification of accuracy for each employment model. The figure complements Table X by visually highlighting the the improvement in accuracy from the benchmark to tuned kNN.

Model	Test RMSE log salary
Naïve	0.740185
LASSO	0.580871
Regression Tree	0.534525
Random Forest	0.522468
kNN	0.527075

Table 2. Test-set RMSE for log-salary models. This table reports test-set root mean squared error (RMSE) on the log-salary scale for the four regression models done in this project: LASSO, regression tree, random forest, and tuned k-nearest neighbors. Note: the lower the RMSE, the better predictive performance.

sonably well with feature independence. Nonetheless, the other models are more valid competitor with flexible methods like random forests and kNN, which suggests that the relationship between the predictors and employment status can be captured reasonably well by an approximately linear decision boundary with appropriate regularization. This is consistent with the exploratory data analysis (EDA), which showed clear but not extremely complex separations in employment rates across fields and demographic groups. In fact, these trends were in alignment with the general schema based on history already formed in our heads about who is employed and who is awarded with higher paying jobs.

Table 1 and Figure 2 show the predictive performance of the five log-salary models. Test-set RMSE ranges from 0.522 to 0.740, indicating that while the models capture a meaningful portion of the variation in earnings, there's still a lot of individual variation they don't explain. The Random Forest model performs best, with the lowest RMSE of 0.522, slightly outperforming the other approaches. The regression tree and kNN models are close behind, with RMSEs of 0.535 and 0.527, respectively, while LASSO does a bit worse at 0.581. As expected, the naïve model has the highest RMSE at 0.740. This pattern fits with the bias–variance tradeoff: regularized linear models and ensemble methods like Random Forest strike a balance between flexibility and stability, whereas a single regression tree can be either too simple (if pruned) or too noisy (if deep), and kNN may struggle when dealing with high-dimensional one-hot encoded features.

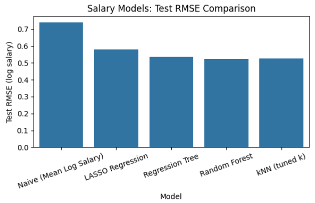

Figure 2. Comparison of log-salary model using the test RMSE metric. The bar graph represents a test-set RMSE for each of the five models, including the benchmark on the log-salary outcome. The figure highlights which of the model best balances bias and variance in predicting earnings among employed respondents.

	Outcome	Model
Employment (Accuracy)	Naive Model	0.883889
	Logistic LASSO	0.890980
	Decision Tree	0.902476
	Random Forest	0.902262
	kNN	0.911791
Log Salary (RMSE)	Naïve	0.740185
	LASSO Regression	0.580871
	Regression Tree	0.534525
	Random Forest	0.522468
	kNN	0.527075

Table 3. Compiling the Performance Metrics of the Employment & Log Salary The employment metric is accuracy with higher numbers indicating more accuracy. The log salary metric is root mean square error (RMSE) and the lower this metric is, which means the model is more accurate because there is smaller error in the variance between predicted in actual values. Thus, here, it can be seen that kNN preforms the best for employment with the highest accuracy metric and random forests performs the best for salary because it has the lowest RMSE, meaning the least difference.

Based on this comparison, it is evident that generally the test accuracy is for the employment models. Even though there is a displayed increase in performance from the naive model at 0.883 accuracy out of 1, the fact that the naive model performs at around 88 percent accuracy means this data is more predictable. However, the predictive accuracy gets up to approximately a 91.2% for kNN classifier and all the other employment classification models were more accurate than the benchmark which demonstrates that the complexity of the models help improve accuracy.

Regarding the log salary regression models, the metric used for these models is root mean square error (RMSE) and for this metric, the lower the value the more accurate the predictive model is. That said, the naive benchmark model got an RMSE of 0.740 which was nearly 0.20 higher than most of the other models. Clearly, the more complex and fitted models for the continuous outcome made significant improvements in the predictability for that outcome. The model with the lowest RMSE and thus the best predictive accuracy was random forests with a 0.522. This difference between the naive and the other models, especially the random forest, means that including personal demographics (e.g., race), field of study, and number of hours worked a week can create a significant reduction in the prediction error. Although this is a large jump from the benchmark, it is quite mediocre, which means that there is a large part of individual salary variation that cannot entirely be explained by the variables we evaluated in this survey.

3.2. Results Analysis (Discussion)

Regarding the employment classification models, all four of them achieved high and tightly clustered test-set accuracy. Logistic LASSO classifier correctly classified about 89 percent of the respondents; the decision tree and random forest each reached about 90 percent, and the tuned kNN model performed the best at about 91 percent accuracy. These numbers suggest that employment status is fairly predictable from the observed characteristics independently, and that more flexible nonparametric methods only marginally improve a well-regularized logistic regression. From the classification employment models it is evident that field of study and degree level are the most important predictors of employment. In addition, race and minority status are both influential but the models indicate that they are not as predictive as the other two.

On the other hand, the earnings regressions based on log salary of the employed in the dataset, the naive baseline that always predicts the mean log salary had an RMSE of about 0.74. All of the fitted models improved by a large margin as discussed earlier. The LASSO regressor reduced RMSE to about 0.58, the decision regression tree to 0.53, the tuned kNN to about 0.53, and the random forest to about 0.52. Here we are seeing a better return on investment for conditioning on demographics, field of study, degree level, and weekly number of hours worked yields a meaningful reduction in the prediction error. However, 0.52 is the lowest RMSE and it is still middle of the line which means there is still a significant amount of individual earnings and the variations between them that are not explained by any of our variables.

Lastly, we are moderately to highly confident in these results. For one, there is consistency between the exploratory data analysis and the trained models used. The simple boxplots showed variance in salary based on race and general field. The model results show these same disparities in their predictions. In addition, it seems all the models are telling the same story about the variables overall. In addition, there is always an improvement from the naive model. Together, this suggests that these findings are not one-offs but account for a general trend in employment, salary and their predictors. Some of the high confidence also comes from all of the cross-validation and tuning that was done in addition to doing a hold-out test set to train the models to their finest and then create an opportunity for the most honest evaluation.

There are still things that make our confidence reserved. For one, all of the data is self-reported so there is the possibility that individuals self-report incorrectly. For instance, it is easy to guess how many hours one works in a week in the moment of a survey, but this is different from checking a time sheet or payroll strip. Another issue with self-reporting is that there may be sample bias, like who reports their salaries. Lastly, the feature set is limited by the variables with sufficient responses available in the IPUMS and thus there are unexplored dimensions of this question and the dimensions of this story.

4. Conclusion

4.1. Project Summarization

The overarching goal of the project was to examine how personal demographics, like race and academic characteristics, like field of study shape employment outcomes after graduation. Using a large and mixed-type survey dataset the probability of being employed was modeled along with the amount of earnings from those who were employed. In this project, we followed the pipeline outlined in the methods section.

From these methods we found a few key findings:

1. Employment status is very predictive from the available covariates. The models, including the naive benchmark all achieved test accuracies between 88 and 91 percent which means that this differences between them were marginal. This relationship in particular does not seem to be complex given the high predictive power of the benchmark.

2. Complex and fitted models are better predictors of earnings. The native baselines was 0.740 for RSME, but all the other trained models preformed better. This signifies that there is a level of interdependence between variables that there is a level of interdependence between variables that the benchmark could not account for. The best preforming model was the random forest model with a 0.522. Despite, the improvement seen for the continuous outcome. It is clear that these variables do not explain the whole story with salary earnings and there are more high influence drivers over salary prediction of an individual.

3. The patterns are consistent in regards to the strongest drivers of outcomes of employment and salary. Field of study and one's degree level are the most powerful predictors of both employment and salary. More specifically, those within the STEM field of study are predicted to have a higher employment probability and are estimated to make more when employed,

4. Another consistent pattern in the models is that personal demographics do matter. For instance, minority graduates have lower predicted employment probabilities and lower predicted earnings than their non-minority, white, peers. These demographic disparities are present in employment and earnings. These findings point to a joint system of inequality between the education system and the labor force. It is said that education is the great equalizer, but not when those with the same education but different races or genders still have disparities in employment and earnings.

4.2. Strength of Results

There are two layers of evidence for the results provided. For one, the exploratory data analysis (EDA) was extremely fruitful in the way it preliminarily highlighted patterns in the distribution of employment and salary across the predictors explored in this study, such as race, gender, and field of study. An amalgamation of histograms, boxplots, grouped summaries, Sankey diagrams, and kernel density plots displayed the differences we would expect to see with the models and their predictions. For instance, it was clear from the box plots that salary tended to be higher for men rather than women and non-minorities opposed to minorities. In addition, despite a clear breadwinner not being identified from the first level of analysis, it is evident from the EDA that humanities and social science discipline majors tend to make less than their peers. The second level of evidence and support for the results is that the models were trained with cross-validation, creating a more accurate performance estimate with hyperparameter tuning and using different data folds to prevent overfitting or underfitting (James et. al., 2023). In addition, the held-out test data approach was taken to create an unbiased estimate for how the model would perform on unseen data. This is especially important because, since the overturning of affirmative action, the influence of personal demographic factors in employment decisions has been a contentious topic. From a machine learning perspective, it provides fair benchmarks for performance evaluation. Lastly, the models' ability to perform on never seen data is best comparable to its practicality on real-world data when this approach is taken, which means the models in this project are suitable for real-world applications. Together, this research strategy of understanding the data from EDA and then creating realistic estimates with the models using cross-validation, tuning, and test-set evaluations.

4.3. Defense Against Criticism

Several critiques can be addressed from this modeling project. One of the more common points of criticism centers on the moderate values of RMSE metrics across the models. This leaves room for one to suggest that there is a lack of predictive power present. However, salary is a variable that is majorly influenced by immeasurable forces. The presence of these things, such as economic conditions and industrial differences, does not stop the validity of the findings, but rather it introduces the terms of complexity when it comes to modeling human wages. Another critique may be the concern of representation within these demographic categories and the potential bias that can be introduced by self-reported surveys. We argue that the data will always contain a measure of error, but by the size of our dataset, those fears are mitigated by the data offering a broad scale of representation across demographic groups. Finally, some may argue that simple models cannot capture nonlinear effects. Our project goes directly against that rhetoric and incorporated tree-based and ensemble methods that would prove that our reliance on just a single model and its assumption shows that patterns shown in nonlinear models can be missed by linear models alone.

4.4. Further Research

In the future, we would like to see this project expanded and strengthened through deeper analysis. The next step would be to incorporate geographic variables and how they impact the labor market may differ across the country. Another space for expansion would be to investigate the differences within narrower occupational groups rather than an overall categorization. From the modeling perspective, we could see the exploration of improved predictive accuracy methods and the ability to handle complex interactions efficiently. Finally, qualitative approaches could contextualize the quantitative findings by exploring how graduates interpret their labor market experiences, with a greater focus on the fields with persistent pay gaps.

To conclude, our research strategy strengthens the reliability of our results and supports their relevance for real-world applications. Though there is no perfect model that predicts salary or employment status accurately without a failure of doubt, the consistent pattern across methods reinforces the broader conclusion that demographic and educational characteristics continue to shape employment opportunities for college-educated adults. By analyzing large-scale survey data with machine learning methods, this project provides a nuanced understanding of the labor market and its disparities, as well as laying the groundwork for further exploration.

5. References

Importance of feature scaling. scikit. https://scikit-learn.org/stable/auto$_{e}xamples/preprocessing/plot_{s}caling_{i}mportance.html$

James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J. (2023). An introduction to statistical learning: With applications in Python. Springer International Publishing Springer. The scikit-learn developers. (2011).

Li, L. (2025). Visualizations [Class lecture]. University of Virginia, School of Data Science, DS 3001: Machine Learning.

Li, L. (2025). Exploratory data analysis (EDA) [Class lecture]. University of Virginia, School of Data Science, DS 3001: Machine Learning.

Li, L. (2025). Regularization [Class lecture]. University of Virginia, School of Data Science, DS 3001: Machine Learning.

Li, L. (2025). Decision trees [Class lecture]. University of Virginia, School of Data Science, DS 3001: Machine Learning.

Li, L. (2025). K-nearest neighbors (KNN) [Class lecture]. University of Virginia, School of Data Science, DS 3001: Machine Learning.

Minnesota Population Center. IPUMS Higher Ed: Version 1.0 [dataset]. Minneapolis, MN: University of Minnesota, 2016. https://doi.org/10.18128/D100.V1.0

Srivastava, T. (2018, March 27). Introduction to k-neighbours algorithm (clustering). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

USE OF AI Consulted AI to explore alternative analysis approaches after initial linear regression and PCA attempts yielded low accuracy.

Used AI to improve grammar, LaTeX formatting, and overall clarity.

Leveraged AI, alongside textbook and lecture guidance, to explain techniques and optimize data processing (e.g., KNN tuning).

Utilized AI for brainstorming, refining research questions, organizing ideas, and finding relevant information.

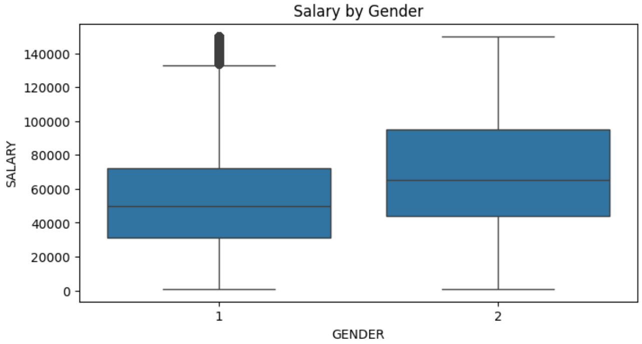

Figure 3. Boxplot of Salary by Gender (woman=1 and man=2)

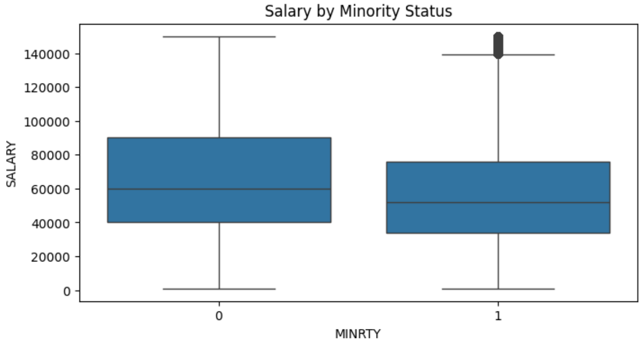

Figure 4. Boxplot of Salary by Minority Status (0= not a minority, 1= minority)

A. Appendix

Appendix A. Data Summary

Dataset contains nearly one million observations.

Predictor categories included:

- **Demographics:** AGE, RACETH, GENDER
- **Education:** highest degree (DGRDG), broad major group (NDGMEMG), detailed major (NDGMED)
- **Labor characteristics:** employment status (LFSTAT), annual salary, hours worked per week (HRSWK), occupational sector

Target Outcomes:

- Employment status (discrete)
- Salary (continuous)

Appendix B. Data Exploration (EDA)

Appendix C. Data Cleaning and Feature Preparation Missingness Handling

- For continuous variables, medians were taken.
- For categorical variables, the elements with the most were considered

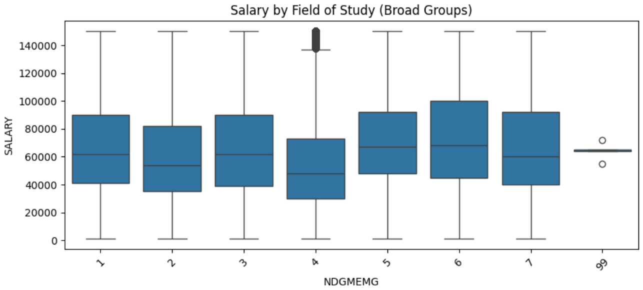

Figure 5. Boxplots of salary by general field of study with the corresponding codebook below.

01	Computer and mathematical sciences
02	Biological, agricultural and environmental life sciences
03	Physical and related sciences
04	Social and related sciences
05	Engineering
06	Science and engineering-related fields
07	Non-science and engineering fields
99	Missing

Table 4. Codebook for reference for Figure 5

Variable	Special Codes	Meaning of Code	Action Taken
SALARY	0, 9999998, 999999999	Logical skip, no response	Recode to NA, drop from salary sample
HRSWK	98, 99	Not applicable, missing	Recoded to NA

Table 5. Examples of what was recoded to handle missing data

Model	Hyperparameter	Grid Searched
LASSO Classifier	C	{0.01,0.1,1,10}
LASSO Regression		0.001 to 1 (log spaced by ten values)
kNN Classifier	k	{1,3,5,7,10,15,20,30}
kNN Regressor	k	{1,3,5,7,10,15,20,30}
Trees/forests	depth, leaves	Max depth of 6-8 Minimum leaves of 100-200

Table 6. The hyperparameter and grids of each model that was tuned

C	Median CV Accuracy	Mean CV Accuracy
0.01	0.890	0.890
0.10	0.891	0.891
1.00	0.891	0.891
10.0	0.891	0.891

Table 7. Output from the LASSO classifier model's cross-validation based on accuracy to select the best model

- Did not remove any missing data or delete rows or columns, simply recoded and filtered

Encoding and Scaling

- Applied one-hot encoding to all categorical features.
- Scaling (StandardScaler) applied only for LASSO and KNN.

Appendix D. Model Implementation Train–Test Breakdown

Used an 70/30 train–test split, with preprocessing fit only on training data to avoid leakage.

Models Implemented via scikit-learn:

- Naive
- Linear Regression
- LASSO Regression
- Decision Tree Regressor
- Random Forest Regressor
- KNN Regressor

Hyperparameter Tuning

Cross-Validation (CV)

- 5-fold CV on full training data for LASSOs and trees/forests
- 3-fold CV on 10% subsample from training data for kNN for computational efficiency

Appendix E. Model Outputs & Comparison

Using LASSO as an example of the iterations that were run to cross-validate the best model.

LASSO Classifier (Employment)

LASSO Regression (Salary Model)

(LASSO penalty)	Median CV RMSE	Mean CV RMSE
0.0010	0.602	0.597
0.0022	0.602	0.598
0.0046	0.603	0.599
0.0100	0.608	0.604
0.0215	0.619	0.617
0.0464	0.633	0.630
0.1000	0.646	0.644
0.2154	0.690	0.689
0.4642	0.742	0.742
1.0000	0.742	0.742

Table 8. Output from the LASSO regressor model's cross-validation using the RMSE metric to identify the best model for accuracy

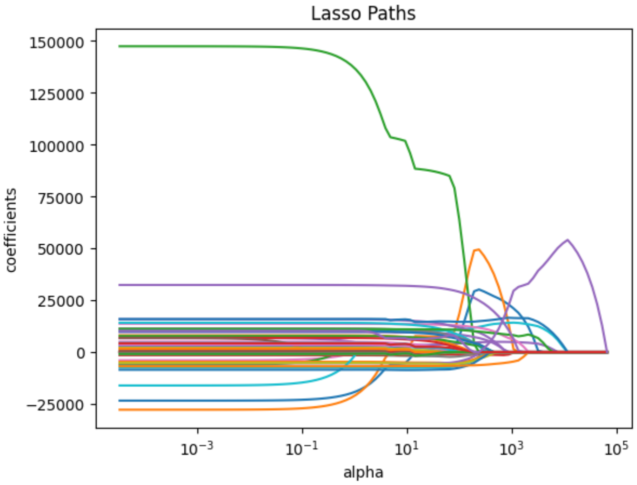

Figure 6. Computing Regularization Path Using the LASSO Turning Most Nonrelevent Coefficients to Zero. This line plot shows the regularization paths for the LASSO model. Each colored line represents the coefficient value for a specific feature as the regularization parameter (alpha) decreases. As alpha decreases, more features enter the model thus their coefficients become non-zero and their magnitudes increase. It is displayed here how LASSO performs continuous feature selection and coefficient shrinkage.

Model	Accuracy	F1 (Employed)	ROC–AUC
Logistic LASSO	0.891	0.942	0.770
Decision Tree	0.902	0.946	0.850
Random Forest	0.902	0.946	0.850

Table 9. Employment Classification Performance Metrics. Attempted to compare the prediction quality of the employment classification models with other metrics. The F1 and ROC-AUC metrics are consistent with the metric used throughout the paper, accuracy. Could not compute for knn due to computational cost.

Model	RMSE (log)	MAE (log)	R^2
Naive (Mean Log Salary)	0.725	0.524	0.000
LASSO Regression	0.581	0.409	0.357
Regression Tree	0.535	0.367	0.456
Random Forest	0.522	0.355	0.480
kNN (tuned k)	0.527	0.359	0.471

Table 10. Salary Regression Performance Metrics. Attempted to compare the salary prediction models based on different metrics where RMSE and MAE are both computed on the log salary scale, thus the closer these are to 0, the better they preform and the closer to 1 the R-squared is, the better the model predicts. All the metrics show that the predictive power improves with complexity and the best model slightly varies depending on the metric.

Employment Classification Metrics

F1 Score Metric

- F1 formula is: F1 = 2 x (Precision x Recall) (Precision + Recall)
- The higher the F1 score the better the precision and recall

ROC-Area Under the Curver (AUC) ROC AUC measures the employment classification model's ability to distinguish between classes and the closer the metric is to 1, the better its predictive power.

Salary Regression Metrics

Mean Absolute Error (MAE) Average magnitude of errors in the predictions of salary and is expressed in the same units as the data normally with a lower a number metric (on a 0 to 1 scale) representing better predictive power.

R-squared

- R-squared Formula: R-squared = 1 - [(Unexplained Variation)/(Total Variation)]
- This metric ranges on a scale from 0 to 1 with 0 representing none of the variance being explained by the model, and 1 meaning all of the variance in salary can be explained by the model. Thus, any coefficient of determination can be converted into a percentage of variation explained, like Random Forests explain 48.0 percent of variation in salary which is the highest R-squared but is quite mediocre and reinforces one of our main takeaways that salary requires more variables outside of our dataset to predict.

ROC-Area Under the Curver (AUC) ROC AUC measures the employment classification model's ability to distinguish between classes and the closer the metric is to 1, the better its predictive power.