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

Brittany Asare [*1] **Hewan Kasie** [*1 2] **Simone Minor** [2] **Firstname8 Lastname8** [3] **Firstname8 Lastname8** [1 2]

Abstract

The current job market in the United States has undergone significant changes in recent decades, creating unique challenges for people seeking employment, particularly for recent college graduates who have suffered significantly. Our project seeks to investigate labor market outcomes with a focus on demographic factors such as race, gender, and field of study shape opportunities and barriers to securing employment. Using data from IPUMS, drawn from national surveys including the NSCG, SDR, NSRCG, and ISDR, we constructed a data set that encompasses decades of labor force experience. This data set allows us to analyze a wide range of variables, including technical skill, highest degree earned, employment status, occupation, employer characteristics, income, job characteristics, job satisfaction, and career trajectory. After cleaning and preparing our data, we applied statistical and visualization methods to identify disparities in employment access and outcomes. Our finding suggests... [TBD].

1. Introduction

Based on our interest in the current economic state of our nation, 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 our data from IPUMS and got our data from a variety of sources, including surveys from across the last few decades, such as NSCG, SDR, NSRCG, and ISDR. In doing so, we were able to get a grasp on the current state of job market. This was a single dataset that we would later be able to derive an analysis from in the ease of securing a job in today's economy.

The main variables that we analyzed from our dataset were the following: technical skill, demographic, highest degree achieved, employment, occupation, employer characteristics, income, job characteristics, job satisfaction, and career path jobs. 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.

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.

[*]Equal contribution [1]Department of XXX, University of YYY, Location, Country [2]Company Name, Location, Country [3]School of ZZZ, Institute of WWW, Location, Country. Correspondence to: Firstname1 Lastname1 <first1.last1@xxx.edu>, Firstname2 Lastname2 <first2.last2@www.uk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

2. Methods

We are predicting a continuous outcome, salary, from a large and mixed-typed survey dataset. To understand the data we were handling, we completed Exploratory Data Analysis (EDA). We used stacked bar chats and sankey diagrams to investigate the data and find patterns and identify potential relationships that we could put to the test with models. After completing this initial exploration, we chose to use four different models, all of which balance our main priorities: handling a big data set, various data types, capturing nonlinear relationships, can produce significant and strong enough results to interpret meaningfully.

The first model ran was a multiple linear regression. Through this model we controlled for categorical variables with dummies to output clean coefficients. This method also accounts for multiple collinearities and dummies. In addition, this method is fast and efficient given there are about one million rows. Ordinary least squares is a start to predicting the relationship between the independent and dependent variables; however least absolute shrinkage and selection operator (LASSO) performs better on this data because it does variable selection and shrinkage, including bringing some coefficients to zero that are less important for a more interpretable model.

The second model ran using decision tree regression. Decision trees are also good for accounting for nonlinear interactions. This model presents a good contrast with the first linear model because the trees do not assume additive effects; instead, it visually shows where the patterns are detected. We fit a shallow tree for better interpretability and then ran a deeper tree for better performance. We compared its out-of-sample interpretability and performance to the previous model.

The third model used was a random forest regression which averaged a multitude of trees and together would have a better predictive performance than one tree. Due to the consideration of many trees, this model would not be sensitive to all the noise in the previous models. This model highlights the strengths of the other models, and we expect that the cross validated performance will demonstrate this.

The final model we used was the k-nearest-neighbor (KNN) regression. We chose KNN because of the pedagogical contrast of our previous models. We anticipated that KNN would not perform as well due to scaling issues for big data. It served as a nonlinear, nonparametric baseline in order to compare against the two tree models. We used this model to demonstrate the bias-variance trade-off when k is tuned.

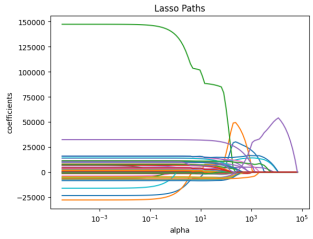

Figure 1. Figure 1: Computing Regularization Path Using the LASSO Turning Most Nonrelevent Coefficients to Zero. This line plot shows the regularization paths for the LASSO model. Each colored line represents the coefficient value for a specific feature as the regularization parameter (alpha) decreases. 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.

3. Results

3.1. LASSO Regression Analysis

The LASSO regression model was applied to the dataset to evaluate the predictive strength of demographic and educational features while performing variable selection. Using an alpha of 2.5, the LASSO effectively shrunk many coefficients to zero as seen in Figure 1 which allowed us to identify the most influential predictors of salary.

3.1.1. Key Observations

We found that only a subset of the features retained nonzero coefficients which highlighting which factors have the strongest association with salary. It is evident that continuous variables such as age and hours worked (after polynomial expansion and scaling) were significant, but many higher-degree interaction terms were shrunk to zero. Regarding the categorical variables, certain race and gender indicators, degree types, and major fields emerged as important predictors. Overall, this model provides a clear picture of which features are consistently impactful while controlling for multicollinearity and overfitting.

3.1.2. Strengths & Limitations

Strengths: Feature selection helps focus on meaningful predictors; interpretable coefficients. Limitations: LASSO assumes linear relationships and may not capture complex interactions that exist in the data.

3.2. Decision Tree

The decision tree classifier was used to explore nonlinear interactions between features and to identify major splits influencing salary. Categorical variables were one-hot encoded, and missing numeric values were median-imputed.

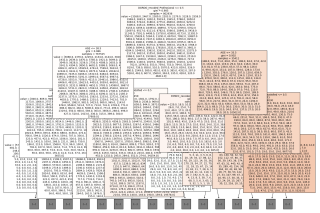

Figure 2. Figure 2: Decision Tree Visualization (Max Depth 3) This decision tree was pruned to a maximum depth of 3 to show the primary splits influencing salary prediction. Each node represents a decision based on a feature (e.g., gender, race, degree type, age, hours worked) which leads to a predicted salary outcome. The color intensity and value at the leaf nodes indicate the average salary within that section, providing insight into the hierarchical relationships and interactions between various demographic and educational factors and salary.

3.2.1. Key Observations

There are major splits in the tree often involved demographic variables like gender and race, as well as educational factors such as degree type and major which can be seen in Figure 2. Root Split (Depth 0): The first split identifies the most impactful feature for predicting salary across the entire dataset which is highest degree earned. From there it splits off into sequentially important variables, like major, race, and gender and through splitting displays the relationships between those variables and its ultimate effect on salary. The interactions were clearly captured, e.g., certain majors had different salary patterns depending on gender or age. In addition, the max-depth-limited tree (depth=3) allowed for interpretable visualization of primary drivers of salary without over complicating the model and tailoring too much to the noise.

3.2.2. Strengths & Limitations

Strengths: Easily interpretable; captures nonlinear relationships and interactions. Limitations: Single trees may overfit; the accuracy is generally lower than other methods used in this project.

3.3. Random Forests Regression

Random Forest regression aggregates multiple decision trees to reduce overfitting and improve predictive performance. Using numeric and one-hot encoded categorical features, the model was trained on 80% of the data and evaluated on the remaining 20%.

Performance Metrics: R^2: 0.3291 — MSE: 935,085,806.87 — MAE: 22,625.00

3.3.1. Key Observations

The feature importance analysis highlighted similar key predictors as LASSO (e.g., major, degree type, race, gender), however in addition it captured complex nonlinear interactions among variables. The Random Forest model outperformed the single decision trees in predictive accuracy while maintaining a degree of interpretability through feature importance metrics. Lastly, the ensemble approach smooths out idiosyncrasies of single trees and provides robust salary predictions across a large dataset.

3.3.2. Strengths & Limitations

Strengths: High predictive power; captures nonlinear interactions; reduces overfitting. Limitations: Less interpretable than individual trees; computationally intensive.

3.4. K-Nearest Neighbors

KNN Analysis The K-Nearest Neighbors model predicts salary based on proximity in the multi-dimensional feature space by using standardized features to ensure that both numeric and categorical predictors contribute appropriately to distance calculations.

3.4.1. Key Observations

KNN achieved comparable predictive accuracy to LASSO and decision tree models but typically lower than Random Forest. The model is less interpretable than the previous models because it does not produce explicit feature coefficients. KNN's predictions are sensitive to the choice of neighbors (k) and the feature scaling hence all the noise seen in Figure 3.

3.4.2. Strengths & Limitations

Strengths: Captures local patterns in the data; can model nonlinear relationships. Limitations: Computationally intensive for large datasets; less interpretable; performance sensitive to hyperparameters and feature scaling. These limitations are displayed via the chaotic graph feature below.

4. Discussion

More to come on this where we will discuss in more depth what the results tell us about salary prediction based on other features

5. Conclusion

More to come later

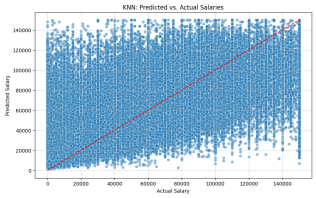

Figure 3. Figure 3: KNN plotting of predicted and actual salaries This scatter plot illustrates the performance of the K-Nearest Neighbors (KNN) regression model in predicting salaries. The x-axis represents the actual salaries and the y-axis shows the salaries predicted by the model. The red dashed line represents a perfect prediction which would occur if the predicted values and actual values match exactly. The spread of points around this line indicates the model's predictive accuracy is low and the distribution of errors is great. This uncoordinated spread means that there is not clear relationship found between salary ranges and its variability features.

5.1. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase.

5.2. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

5.2.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

5.2.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

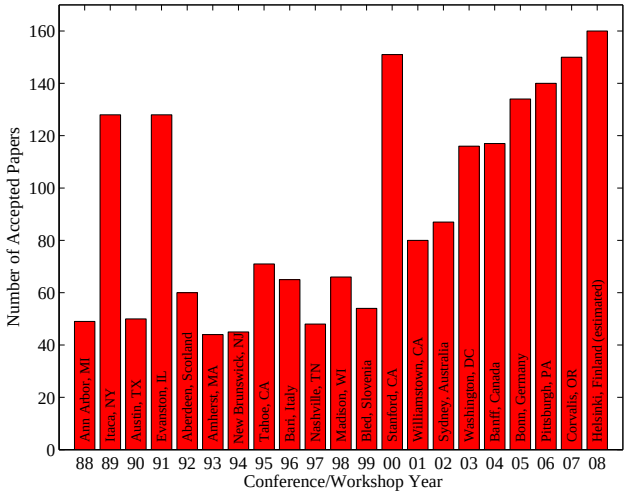

Figure 4. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

5.3. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should

[1]Footnotes should be complete sentences.

[2]Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9± 0.2	96.7± 0.2	√
CLEVELAND	83.3± 0.6	80.0± 0.6	×
GLASS2	61.9± 1.4	83.8± 0.7	√
CREDIT	74.8± 0.5	78.3± 0.6	
HORSE	73.3± 0.9	69.7± 1.0	×
META	67.1± 0.6	76.5± 0.5	√
PIMA	75.1± 0.6	73.9± 0.5	
VEHICLE	44.9± 0.6	61.5± 0.4	√

serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure 4. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX). Always place two-column figures at the top or bottom of the page.

5.4. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

5.5. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 1. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

5.6. Theorems and such

The preferred way is to number definitions, propositions, lemmas, etc. consecutively, within sections, as shown below.

Definition 5.1. A function $f : X \to Y$ is injective if for any $x, y \in X$ different, $f(x) \neq f(y)$.

Using Definition 5.1 we immediate get the following result:

Proposition 5.2. *If f is injective mapping a set X to another set Y, the cardinality of Y is at least as large as that of X*

Proof. Left as an exercise to the reader. □

Lemma 5.3 stated next will prove to be useful.

Lemma 5.3. *For any $f : X \to Y$ and $g : Y \to Z$ injective functions, $f \circ g$ is injective.*

Theorem 5.4. *If $f : X \to Y$ is bijective, the cardinality of X and Y are the same.*

An easy corollary of Theorem 5.4 is the following:

Corollary 5.5. *If $f : X \to Y$ is bijective, the cardinality of X is at least as large as that of Y.*

Assumption 5.6. The set X is finite.

Remark 5.7. According to some, it is only the finite case (cf. Assumption 5.6) that is interesting.

5.7. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2025.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (1959). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (Samuel, 1959). List multiple references separated by semicolons (Kearns, 1989; Samuel, 1959; Mitchell, 1980). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (Michalski et al., 1983).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to **??** for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (Samuel, 1959), conference publications (Langley, 2000), book chapters (Newell & Rosenbloom, 1981), books (Duda et al., 2000), edited volumes (Michalski et al., 1983), technical reports (Mitchell, 1980), and dissertations (Kearns, 1989).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

Please put some effort into making references complete, presentable, and consistent, e.g. use the actual current name of authors. If using bibtex, please protect capital letters of names and abbreviations in titles, for example, use {B}ayesian or {L}ipschitz in your .bib file.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and usually should) include acknowledgements. Such acknowledgements should be placed at the end of the section, in an unnumbered section that does not count towards the paper page limit. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

Impact Statement

Authors are **required** to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. This statement should be in an unnumbered section at the end of the paper (co-located with Acknowledgements – the two may appear in either order, but both must be before References), and does not count toward the paper page limit. In many cases, where the ethical impacts and expected societal implications are those that are well established when advancing the field of Machine Learning, substantial discussion is not required, and a simple statement such as the following will suffice:

"This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here."

The above statement can be used verbatim in such cases, but we encourage authors to think about whether there is content which does warrant further discussion, as this statement will be apparent if the paper is later flagged for ethics review.

References

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Kearns, M. J. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (eds.). *Machine Learning: An Artificial Intelligence Approach, Vol. I.* Tioga, Palo Alto, CA, 1983.

Mitchell, T. M. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

Newell, A. and Rosenbloom, P. S. Mechanisms of skill acquisition and the law of practice. In Anderson, J. R. (ed.), *Cognitive Skills and Their Acquisition*, chapter 1, pp. 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):211–229, 1959.

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one.

The `\onecolumn` command above can be kept in place if you prefer a one-column appendix, or can be removed if you prefer a two-column appendix. Apart from this possible change, the style (font size, spacing, margins, page numbering, etc.) should be kept the same as the main body.