Predictive Analysis of Salary Classifications: The Influences of Demographics and Professions

Authors
Renee Reddy, Rhea Johnson, Richa Juvekar, Saisha Ketkar

Affiliations
Northeastern University
7 December, 2023

01

Abstract

This project aims to analyze income determinants, exploring the interplay of various demographic, personal, and professional factors. With a dataset consisting of 32,561 records and 15 features, our goal is to unravel the complexities of salary classifications—categorized as either surpassing $50,000 or falling at or below this threshold. Income disparities are a major concern, influenced by a myriad of factors spanning personal demographics, educational backgrounds, and experiences.

02

Introduction

Definition of Problem

Every human experience is unique, and each individual's life and career is shaped by a variety of environmental factors, such as age, education level, familial relations, and gender. The connections between such environmental variables and income levels highlights the complexity of societal structures and individual opportunities. The problem lies not only in recognizing the impact of these factors but also in comprehending how they interact and contribute to shaping financial outcomes.

Motivation

Our motivation for this project lies in recognizing the uniqueness of every individual's life and career journey, molded by diverse environmental factors. Income inequality is a pervasive concern, and the project is motivated by a genuine desire to unearth deeper insights into the factors driving these disparities. As a team comprised of female-identifying members, we are accutely awareness of the institutional wage gap, a critical issue that can significantly impact potential incomes. Through this project, we aim not only to contribute to the broader understanding of income determinants but also to shed light on gender-specific dynamics that play a role in shaping economic outcome

03

Goals and Objectives

Our main objective is to recognize what disparities are prevalent in salaries across a variety of occupations and demographics across the US. From our data, we aim to spot what demographics are facing notable differences in income distribution. Our analysis will aim to highlight social justice issues in the way salaries are distributed amongst the individuals part of our dataset and perhaps provide important insight into institutional issues regarding income inequality.

04

Related Works

Data Science Salary Prediction using Streamlit (Medium.com)
This work delves into the user's experience in utilizing data science tools in order to predict their own salary. The user utilized a tool called Streamlit in order to create a user-friendly interface that allows anyone to utilize their model and predict their own salaries. The salary predictions are based on factors such as current salary, work sectors, years of experience, job title, location, and education level. We found this work to be an incredible reference point in our own research, especially with the step-by-step nature of the user's posted article.

05

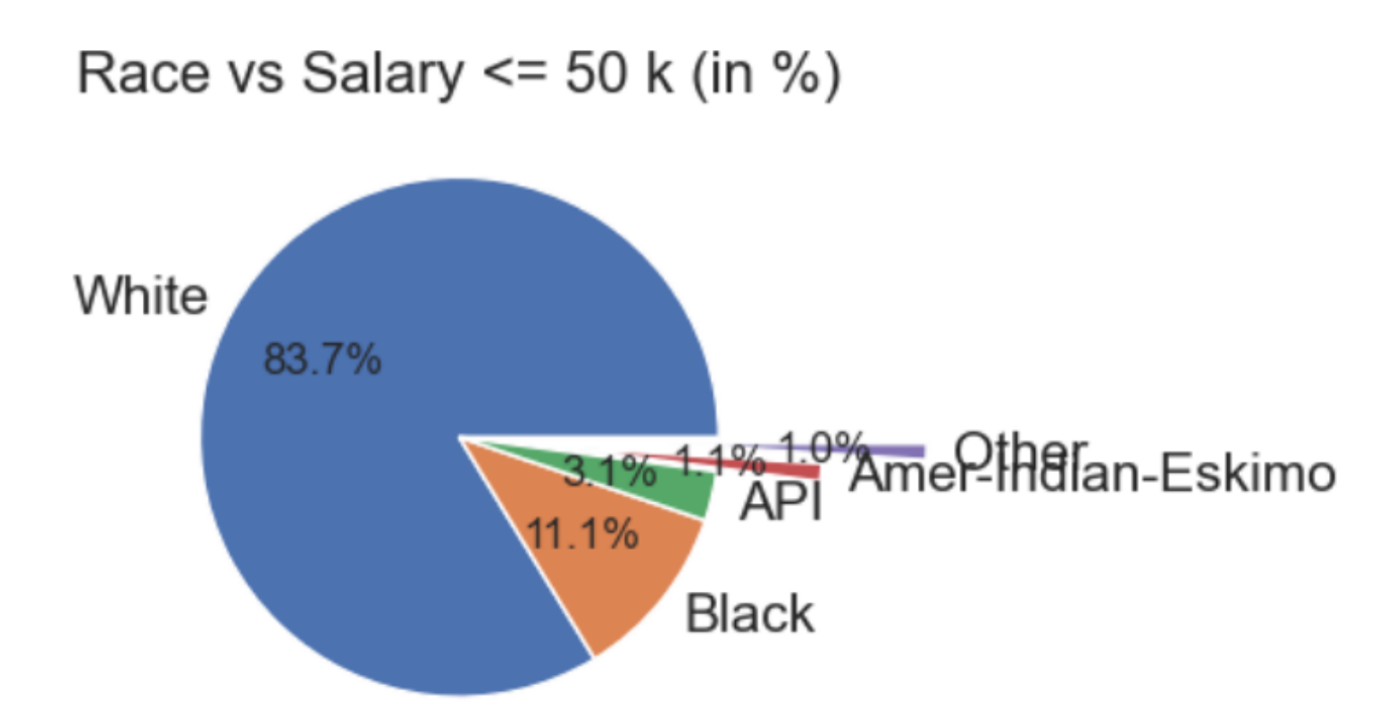
Methodology

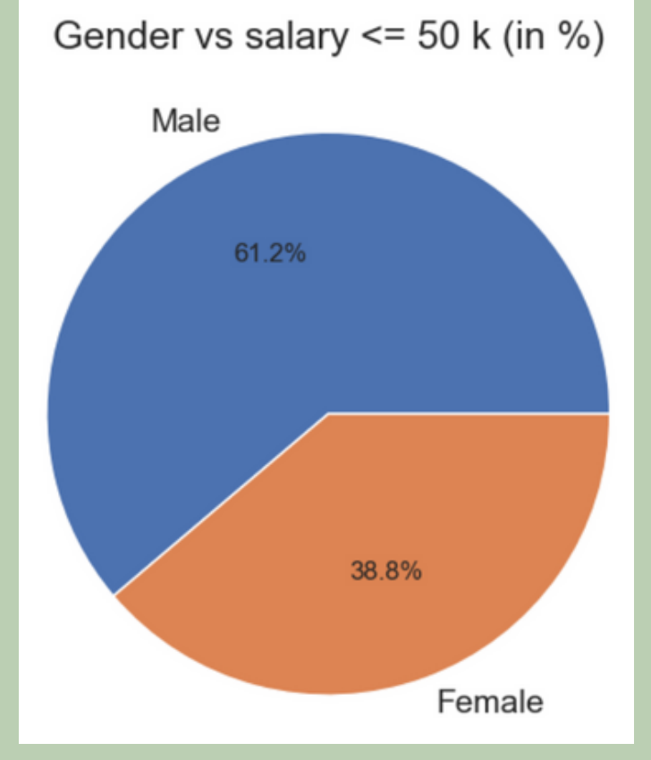
Data Acquisition/Preparation

We utilized the Salary Prediction Classification dataset from Kaggle.com to begin our data acquisition and preparation. This data set includes age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, and salary. We utilized EDA practices to make our raw data useful for the project.

- Dropped any duplicated rows.
- Replaced all the '?' within the dataset with 'None', then replaced them with a String value so that, it would be easier to identify in the next step.
- Replaced names within the dataset to be easier to read and condensed some names as they could be a part of a broader category.
- Utilized bar charts and pie charts to visualize the distribution of the different variables.
- Encoded all variables to be numerical.

A pie chart representing the proportion per race with a salary over 50k





A pie chart representing the proportion of males vs females with a salary over 50k

Model Selection

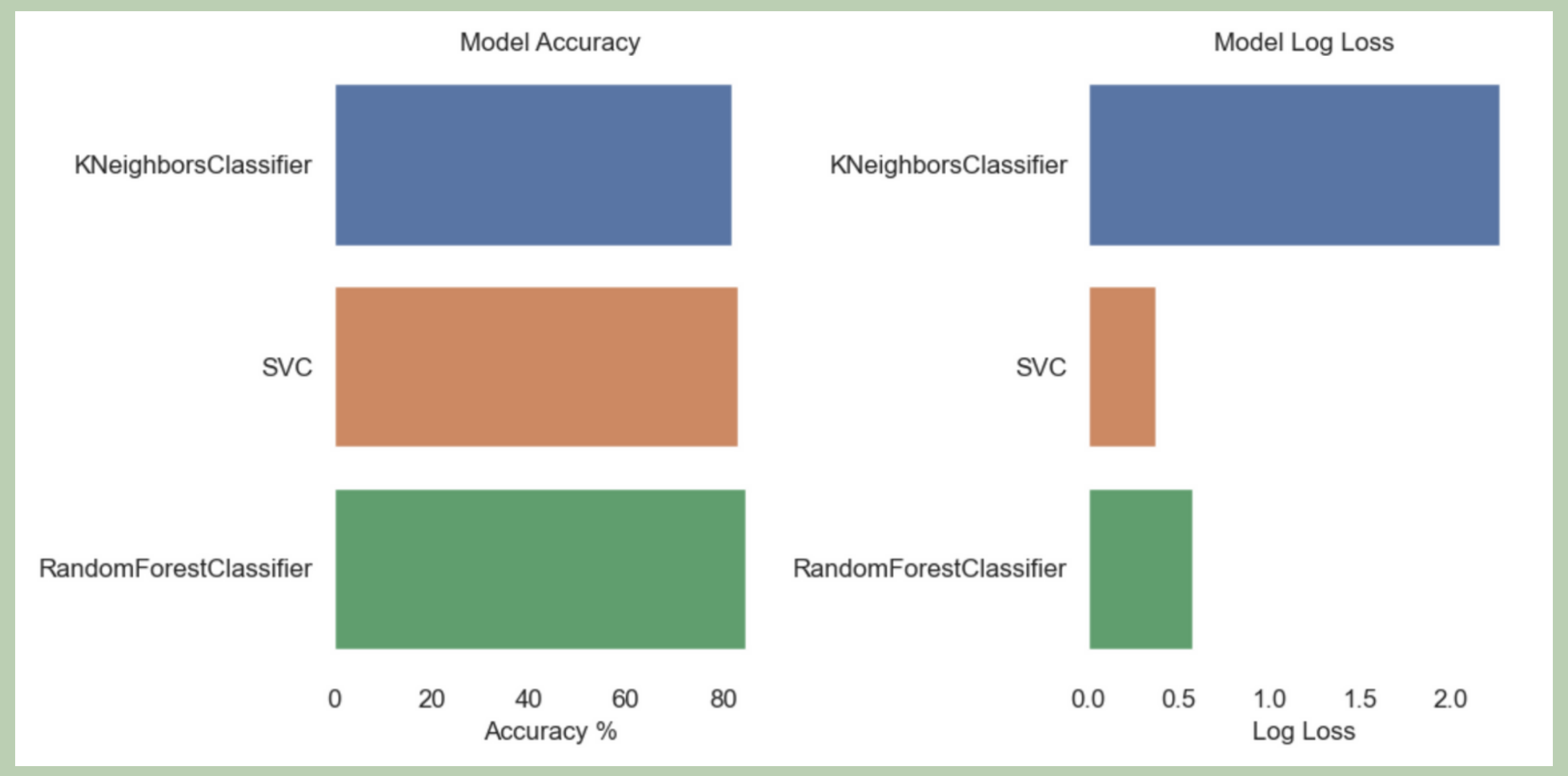
After preparing our data, we then selected our three models of choice: KNeighbors Classification, Random Forest Classification, and Support Vector Machines (SVM). To train our models, we split our dataset into training and testing sets using the train_test_split function from scikit-learn and scaled the data. In terms of evaluating models, we planned to evaluate the testing set to assess the models' generalization capabilities to new, unseen data. We chose to utilize accuracy and log loss to further analyze our results.

Ultimately, the best model decision is dependent on our personal project goals. For higher accuracy, RandomForestClassifier is the best fit. Meanwhile, for well-calibrated predictions, SVC's lower log loss score makes that the best contender. For the purpose of our project, as we are more focused on higher accuracy as a measure, we believe the RandomForestClassifier is the best fit of the three. However, it is important to note that even though it has the highest accuracy levels, it is still relatively low and indicates that these models might not provide the best predictions possible.

06

Results and Evaluations

- RandomForestClassifier: This model had an accuracy of 84.50% and a log loss of 0.5702. While the log loss is higher than then SVC, this model had the highest accuracy measure out of the three.
- Support Vector Classifier (SVC): This model had an accuracy of 82.90% and a log loss of 0.3714, indicating good calibration.
- KNeighborsClassifier: This model had an accuracy of 81.58% and a log loss of 2.3079. Being the model with the lowest accuracy and highest log loss, we can deduce that this model is not the best suited for predictions.



Bar charts representing the initial results of each model's accuracy and log loss

After hyperparameter tuning, while each model's accuracy measures increased between 1-2% for each of the models, all still had considerably low accuracies. This told us that perhaps due to issues such as overfitting and the complexity of our chosen dataset, none of the models are the best suited to give consistently accurate predictions—something to continue exploring,

07

Impacts

The implementation of our predictive analysis model on various salary classifications can bring about positive impacts across various stakeholders within the workforce, including...

- **Job Seekers**: Job seekers will not only be more empowered to make informed decisions about their careers but also allow them to more effectively negotiate their pay. In specific, those who are part of marginalized groups could benefit from understanding how a myriad of factors could contribute to the disparities seen in income.
- **Employers and Human Resource Departments**: By adopting these insights, employers and HR departments can modify their hiring processes and policies. This can improve diversity within the workforce and incite a significant positive impact on wage gaps.
- **Research Communities**: By understanding the deeper societal issues leading to income disparities between gender and income, researchers can build upon existing research to explore additional findings on the issue. Researchers can focus on interdisciplinary studies that integrate demographic, personal, and professional factors to analyze income disparities, allowing collaboration between various fields.
- **General Public and Advisory Groups**: Advocacy groups (such as women or people of color organizations) that focus on social justice can use the findings to contribute to initiatives regarding welfare and social justice. This would help allow transparency within public discourse as well and help promote equity in the workforce.

06

Conclusion

We began our project with the goal of unraveling the intricacies of salary classifications by trying to explore the combination of demographic, personal, and professional factors. Through our utilization of our three models, KNeighbors Classification, Random Forest Classification, and Support Vector Machines (SVM), we feel we were successful in garnering a better understanding of what factors impact the various distributions of salary across a large set of individuals.

Future Work and Improvements

We hope to incorporate more nuanced features into the dataset to offer a deeper understanding of salary determinants such as work-life balance, or industry-specific factors and job satisfaction could be impactful on the model. We would also be interested in reworking our models to utilize less complex datasets for future work in order to create more accurate models. Since the landscape of the workforce is constantly changing, updating our model would be a crucial aspect. We believe that our work is crucial to understanding and interpret institutionalized issues in regards to income disparities in everyday life.

References:

Wage inequality in the United States - Statistics & Facts. https://www.statista.com/topics/3453/wage-inequality-in-the-united-states/#dossier-chapter4. (n.d.).

Office, U. S. G. A. (n.d.). Women in the workforce: The gender pay gap is greater for certain racial and ethnic groups and varies by education level. Women in the Workforce: The Gender Pay Gap Is Greater for Certain Racial and Ethnic Groups and Varies by Education Level | U.S. GAO. https://www.gao.gov/products/gao-23-106041

Kochhar, R. (2023, March 1). The Enduring Grip of the Gender Pay Gap. Pew Research Center's Social & Demographic Trends Project. https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/#:~:text=Gender%20pay%20gap%20differs%20widely%20by%20race%20and%20ethnicity,-Looking%20across%20racial&text=In%202022%2C%20Black%20women%20earned,%2C%20making%2093%25%20as%20much.