Classification of Salary By Occupation, Gender, and Other Metrics

Eric Nguyen

April 28, 2022

Abstract

In an educational venture, we investigate the construction of a classifier which is used to classify salaries. For simplicity, the k-Nearest Neighbors (kNN) classifier is selected in this case. We use a sample of 29,000 records of U.S. Americans working full-time and reduce the feature set to have the following features: survey year ("year"), inflation-adjusted salary ("realrinc"), age ("age"), occupational prestige ("prestg10"), number of children ("childs"), gender ("gender"), education level ("educcat"), and marital status ("maritalcat"). Then, to deal with missing data, we apply kNN imputation and deletion methods. The salaries are categorized into three brackets: low income ($227–$30,600), middle income ($30,600–$117,000), and high income ($117,000–$480,000) using k-means clustering method of discretization on the numerical salary data. After the data is processed, we produce a kNN classifier ($k = 10$) that is able to achieve $(78 \pm 1.5)\%$ accuracy on the test data where 90% of the data (26,100 samples) is used for training and the other 10% is used for testing (2,900 samples). Compared to various to kNN classifiers with smaller k, our classifier with $k = 10$ performs better.

1 Introduction

Everyone wants to have a successful career. Many people are told at a young age that to have a successful career, they must go to school and study in a respectable, professional field, such as engineering, law, or medicine whereas other, nonprofessional occupations such as retail or sanitation work may be looked down upon. To what extent is it true that successful careers only come from the pursuit of professional occupations? Is it possible for those working in nonprofessional occupations to make salaries similar to those working in professional occupations?

Occupation may not be the only factor to take into consideration when predicting career success. We may want to consider other variables could affect such a prediction, such as education, age, gender, etc. For example, we might infer that those who have completed higher education (i.e., graduated from college) will tend to have more successful careers since jobs requiring higher education are often highly respectable and well paid, in addition to many other benefits associated with them. Or, we might guess that older people are going to be more successful in their career than younger people simply due to them having more experience, on average, than younger people. How might these different variables (education, age, gender, occupation, etc.) affect career success for an individual?

Before such questions may be investigated, we need to define what success means in terms of career development. The problem is, everyone will have their own definition of a successful career. For example, a successful career to one person may mean working in a very respected profession, while a successful career to another person simply may just mean achieving the highest salaries possible for their target occupation. We may also consider that not everyone will be working full-time, salaried jobs, but perhaps are only working part-time or are retired, for example.

The General Social Survey (GSS) is a project from the University of Chicago which has been collecting survey data on Americans since 1972 to investigate societal changes. The GSS data is publicly available online and includes a vast collection of records with many variables

(over 5,000 variables) including the variables that may affect salary levels but also including variables that have no correlation with salary and also has a great amount of missing data [3].

While according to Abele et al. [1] it is recommended to evaluate career success based on multiple measures (e.g., position/status, promotions) in addition to salary, for simplicity's sake we will only quantify career success in terms of salary. In this paper, our goal is to apply data mining and machine learning (ML) techniques to predict salaries of full-time workers given in the GSS data based on the variables previously mentioned, including education, age, gender, occupational prestige, number of children, and marital status. Indeed, many studies agree that at least some of these variables are indicators of salary level and career success [5], [7], [9]. The goal is not to achieve the best model for salary prediction, but merely to produce a working prototype for experimental and educational purposes. We use a curated subset of the GSS data from the `stevedata` package in the R programming language called `gss_wages` which retains the dimensions relevant to our investigation, drastically reducing the dimensions from over 5,000 variables to only 11 variables [8]. We only focus on full-time workers (as opposed to part-time workers, students, retirees, etc.) since salaries generally do not apply to those other categories. For the salary prediction, we specifically use the inflation-adjusted variable for income which takes care of the need for us to account for inflation—the details on how these variables are calculated are explained in [4]. Given the structure of the data and for practical purposes (e.g., compared to artificial neural networks which may require more computational resources), the models we consider in this case are either the decision tree classifier or kNN classifer. We choose the kNN classifier over the decision tree classifier due to the greater flexibility in the boundary conditions. Ultimately, we are able to produce a working model that accurately predicts salaries (78% $\pm$ 1.5% accuracy) based on the relevant variables.

2 Related Work

Salary prediction has been investigated in various recent studies, however few studies were found that used machine learning techniques in particular. Nonetheless, most recents studies suggest that occupation, education, and gender are all indicators of salary levels.

In one study by Matz et al. [7], they collected a sample of 2,623 U.S. Americans from Facebook including data on age, gender, education, occupation, and other variables. They used a machine learning technique called ridge regression on Facebook likes and status updates which yielded high accuracy with a correlation of up to $r = 0.49$. Their findings showed that industry had an correlation of $r = 0.23$ and education level had a correlation of $r = 0.3$ which imply that these variables are strong indicators of salary.

In another study by Martín et al. [6], they use machine learning techniques to predict salaries in the IT job market in Spain. Specifically, they predict salaries from 4,000 jobs listed by 488 different companies across various job boards and social networks such as those listed on websites such as *Monster* or *LinkedIn* based on features such as *dedication*, *incentives*, *education*, etc. While the domain slightly differs from that of this paper (e.g., this paper examines salaries in the U.S. based on general population metrics) it still maintains a similar approach, that is, using machine learning. In their study, they consider various solutions including linear models (LM), logistic regression (LR), kNN, multi-layer perceptrons (MLP), support vector machines (SVM), random forests (RF), and adaptive boosting with decision trees (AB), and they also consider using ensemble methods of the previous models to improve classification performance. They decide to use two different methods of ensemble learning using so-called "voting classifiers": one that includes all classifiers considered (called *Vote*), and the other using only the top-3 best performing models (called *Vote3*). Out of all classification methods, the ensemble models generally performed the best including outperforming the kNN ($k = 16$) method. However, still, the kNN method was on the higher end of the spectrum in terms of model performance compared to the other classifiers with only 5% less accuracy than the ensemble methods, well outperforming the LR, MLP, and SVM methods.

Considering the simplicity of implementing the kNN classifier and that it only marginally underperforms other techniques such as ensemble methods (which may be more complicated to implement), the kNN is indeed not a bad choice of model for the purposes of this paper.

3 Methodology

3.1 Data collection

We use the `gss_wages` dataset as found in the `stevedata` R package. Obtaining the dataset is as easy as installing the `stevedata` package in R using the command *install.package("stevedata")* and then using the package via the command *library(stevedata)* in R. Alternatively, the `gss_wages` dataset can be found in Vincent Arel-Bundock's collection of 1884 datasets known as *Rdatasets*, downloadable as a CSV file [2]. However, this method will require additional processing of the data, since the method of importing the data in R will have the data already processed for manipulation in using R whereas the CSV file will only provide string values which must be parsed.

3.2 Data cleaning and manual feature selection

The `gss_wages` dataset originally contains 61,697 entries of survey results of U.S. Americans with 11 columns (variables): "year", "realrinc", "age", "occ10", "occrecode", "prestg10", "childs", "wrkstat", "gender", "educcat", and "maritalcat". First, we select only the records where "wrkstat" variable corresponds to full-time employment which reduces our dataset to 30,491 entries. Then, we drop the columns that are either redundant or not a strong indicator, in this case those are "occ10", "occrecode", and "wrkstat". The inflation-adjusted income column has a significant number of missing entries (23,810 missing entries). To account for this, we apply kNN imputation methods ($k = 3$) to fill in the gaps. After the income is imputed, we drop the remaining records that still have missing values in other variables, which happens to be a small percentage of the data ($\sim 2\%$), reducing our dataset to 29,783 entries

with 8 variables remaining. Because the number of variables are low to begin with, I decide to skip dimensionality reduction methods as the model training is still reasonably fast. In a real-world scenario, there would likely be many more dimensions involved thus requiring dimensionality methods such as singular value decomposition (SVD) or autoencoders to be applied.

3.3 Data processing for training

We discretize the income data using k-means clustering to form three categories: low income ($227–$30,006), middle income ($30,006–$117,000), and high income ($117,000–$480,000); and then we drop the original column for income.

We then normalize every data point such that the numbers range from 0 to 1 so that we can apply kNN. To do this for categorical variables, we employ a so-called "dummy coding" technique which converts each category into a separate, binary column in which 0 represents absence of the category and 1 represents presence of the category.

For training, we take 29,000 samples of the 29,783 entries. 90% of the sample data is used for training while the other 10% is used for testing.

3.4 kNN classification

The general idea of the kNN classifier is that each prediction is based on the predictions of the k-nearest neighbors where the neighbors are calculated using some distance measure. For simplicity, the Euclidean distance measure is used in this case, with $k = 10$.

4 Results

To evaluate the model performance I produced a confusion matrix (cross-tabulation of observed and predicted classes) as shown in Table I. The confusion matrix tells us that our kNN classifier has a $(78 \pm 1.5)\%$ accuracy. The error of $\pm 1.5\%$ is found in comparing the

accuracy 78% against its calculated 95% confidence intervals (76.45%, 79.5%).

When comparing the kNN ($k = 10$) classifier to other kNN classifiers with less neighbors (smaller k), we see that $k = 10$ achieves better performance as shown in Figure 1.

In investigating the prediction indicators for each variable, which get interesting results. To evaluate the predictive strength for each variable, we simply make use of visualizations in the form of histograms. While we will not display visualizations for each metric for the sake of space, we will display a select few that are interesting, as shown in Figure 2 and Figure 3. In our visualizations, we find that people in the higher income brackets tend to: (1) be older, (2) work in more prestigious occupations, (3) be male, (4) have attained at least a bachelor's degree, and (5) be married. Other variables such as survey year and number of children appear to not be as strong predictors.

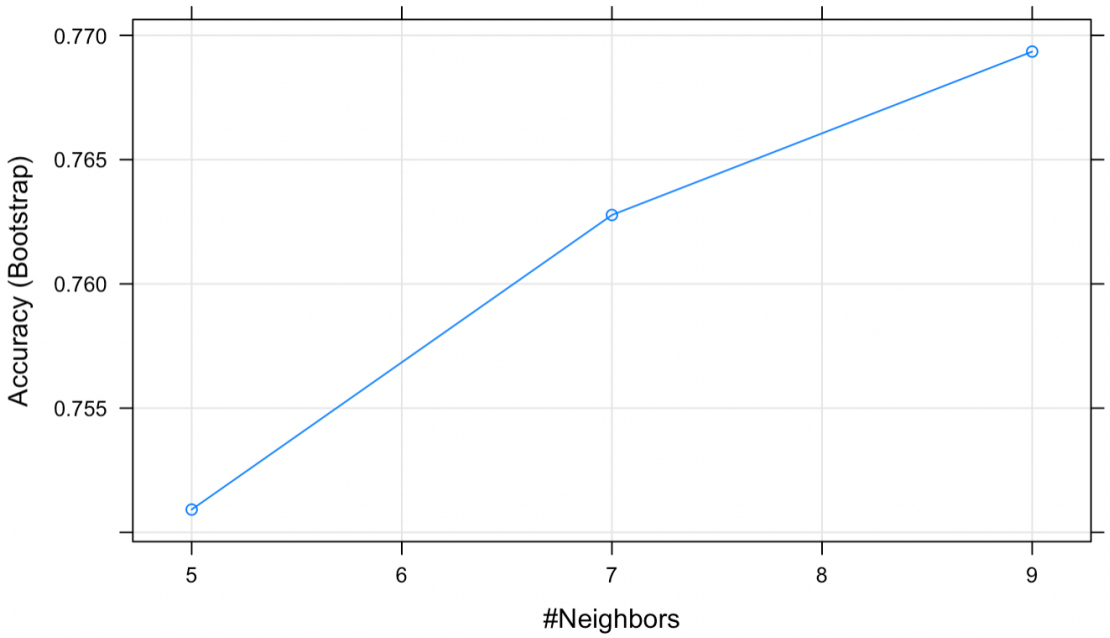

Figure 1: Plot of kNN classifier accuracy for $k = 5$, $k = 7$, and $k = 9$ where $k = 9$ achieves the highest accuracy out of all of them—just less than 77%—however falls short of the accuracy for $k = 10$ which gives an accuracy of 78%.

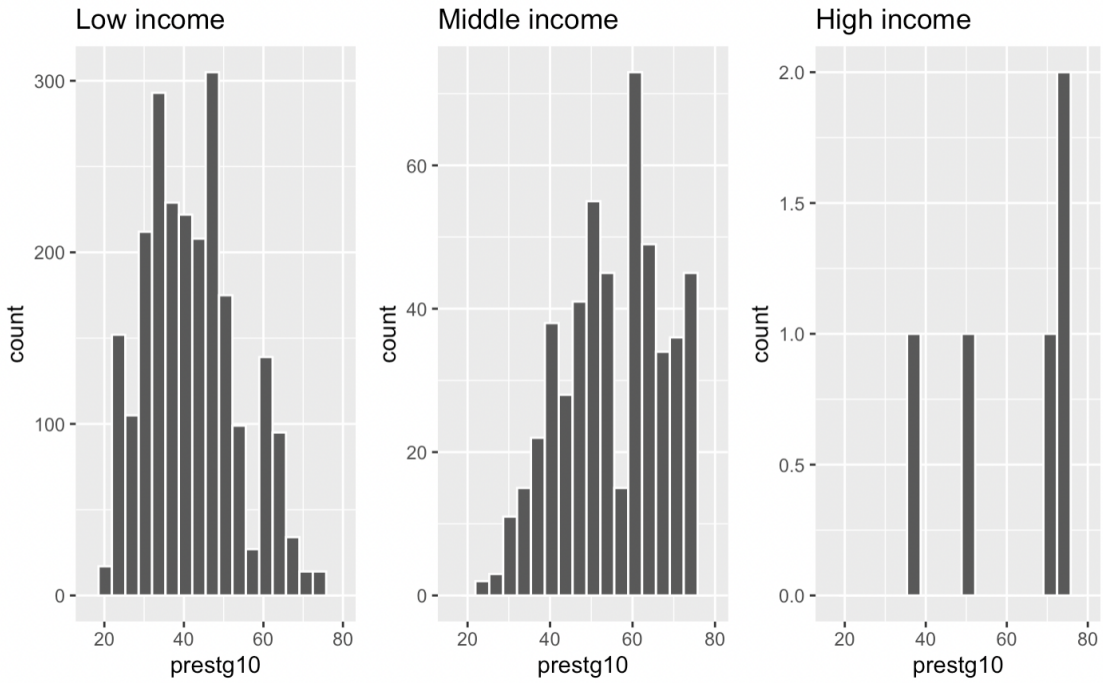

Figure 2: Occupation distributions based on model predictions.

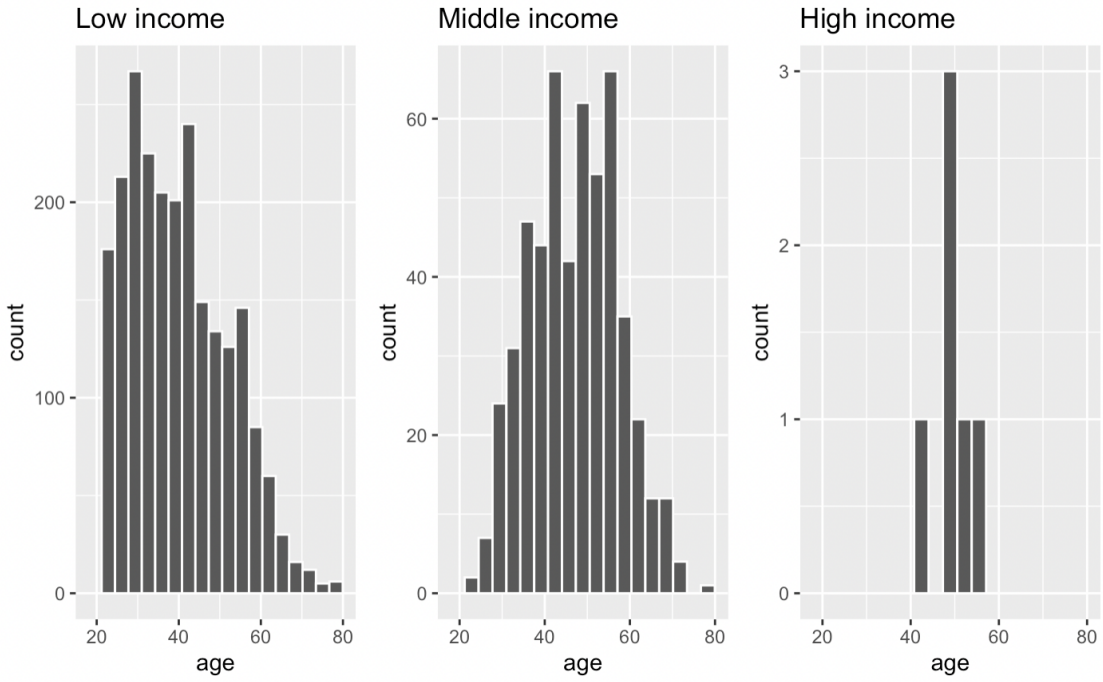

Figure 3: Age distributions based on model predictions.

Table I: Confusion matrix of observed vs. predicted salaries based on $k = 10$ kNN classifer, yielding accuracy of $(78 \pm 1.5)\%$.

		Observed		
		Low income	Middle income	High income
Predicted	Low income	1972	390	21
	Middle income	185	522	37
	High income	3	2	2

5 Discussion

This study focuses on the prediction of salaries based on occupation, age, gender, survey year, number of children, education status, and marital status using the kNN classifier model. We achieved a $(78 \pm 1.5)\%$ accuracy on our kNN classifier model using a 90/10 ratio of training and testing data on 29,000 samples of full-time workers. Ultimately, we were successful in constructing a classifier that can reasonably predict salaries based on various metrics.

Through visualizations, we found that the variables of occupation, age, gender, education level, and marital status are indeed strong predictors of salaries. In general, people with higher education levels and with higher occupational prestige are more likely to have higher salaries (and thus, more successful careers) however the occupational prestige is not a strict requirement. To clarify, the "occupational prestige" variable is a number ranging from 16-80 which is determined by sociologists for various occupations, where lower occupational prestige tends to be associated with service occupations such as retail or restaurant work while higher occupational prestige tends to be associated with professional work such as engineering, legal, or medical professions.

There were many limitations in this study that were not addressed, including feature selection, dimensionality reduction, model comparison, model evaluation, ensemble methods, and more. For example, we may have used a decision tree classifier and compared its accuracy to the kNN classifier, but we did not. In a real-world setting, we would take the time to investigate these areas in more depth; however because this study is only for educational purposes and given the time constraints on the study, we decide to leave these efforts for

possible future work.

6 Acknowledgements

This study was possible thanks to the data provided by the GSS and the curated `gss_wages` dataset from Steve Miller. Also, I would like to acknowledge the QuantDev team for their guidance on using R for kNN classification.

References

[1] Andrea E. Abele, Daniel Spurk, and Judith Volmer. "The construct of career success: measurement issues and an empirical example". In: *Zeitschrift für ArbeitsmarktForschung* 43.3 (2011), pp. 195–206. DOI: 10.1007/s12651-010-0034-6. URL: https://doi.org/10.1007/s12651-010-0034-6.

[2] Vincent Arel-Bundock. *Rdatasets*. URL: https://vincentarelbundock.github.io/Rdatasets/articles/data.html.

[3] *GSS Data explorer: NORC at the University of Chicago*. URL: https://gssdataexplorer.norc.org/.

[4] Michael Hout. "Getting the Most Out of the GSS Income Measures". In: *GSS Methodological Report* 101 (2004). URL: https://gss.norc.org/Documents/reports/methodological-reports/MR101%20Getting%20the%20Most%20Out%20of%20the%20GSS%20Income%20Measures.pdf.

[5] Jakob Mainert et al. "The Incremental Contribution of Complex Problem-Solving Skills to the Prediction of Job Level, Job Complexity, and Salary". In: *Journal of Business and Psychology* 34.6 (2019), pp. 825–845. DOI: 10.1007/s10869-018-9561-x. URL: https://doi.org/10.1007/s10869-018-9561-x.

[6] Ignacio Martín et al. "Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study". In: *International Journal of Computational Intelligence Systems* 11.1 (2018), p. 1192. DOI: 10.2991/ijcis.11.1.90. URL: https://doi.org/10.2991/ijcis.11.1.90.

[7] Sandra C. Matz et al. "Predicting individual-level income from Facebook profiles". In: *PLOS ONE* 14.3 (Mar. 2019). Ed. by Jaap Denissen, e0214369. DOI: 10.1371/journal.pone.0214369. URL: https://doi.org/10.1371/journal.pone.0214369.

[8] Steve Miller. *stevedata: Steve's Toy Data for Teaching About a Variety of Methodological, Social, and Political Topics*. R package version 0.7.0. 2022. URL: http://svmiller.com/stevedata/.

[9] Erica S. Weisgram, Rebecca S. Bigler, and Lynn S. Liben. "Gender, Values, and Occupational Interests Among Children, Adolescents, and Adults". In: *Child Development* 81.3 (2010), pp. 778–796. ISSN: 00093920, 14678624. URL: http://www.jstor.org/stable/40599133 (visited on 04/27/2022).