

Artificial Intelligence Models to Support Curriculum Development

James Otto, Towson University, United States
Chaodong Han, Towson University, United States
Stella Tomasi, Towson University, United States

The IAFOR International Conference on Education – Hawaii 2020
Official Conference Proceeding

Abstract

This research describes the development of analytical tools to leverage available U.S. Government data to support curriculum development for skills education. Specifically, we apply artificial intelligence neural networks and multiple linear regression models to predict a person's annual wages based on the levels and combinations of skills that they possess. These machine learning models are developed using federal data for 35 general job skills combined with annual wage information for over 960 standard occupations. Given this input data, the resulting neural network trains to above 70 percent accuracy in predicting annual wage levels. The multiple linear regression models provide somewhat lower performance. Curriculum developers and education administrators can use these models to determine what level and mix of occupational skills are most appropriate for meeting student goals and optimizing wage potentials. Job and career seekers can use these models to generate estimates of how well their skills should be compensated by the job market.

Keywords: neural networks, multiple linear regression, occupational skills, salary prediction

iafor

The International Academic Forum
www.iafor.org

Introduction

This paper describes the results of developing machine learning models that relate over 900 different combinations of up to 35 diverse occupational skills to their overall value (in terms of annual wages) in the job market. The predictive analytics leverage both neural network (NN) and multiple linear regression (MLR) models to provide valuable insights for human resource skills management and concomitant wage determinations.

Machine learning is a subset of artificial intelligence and provides information systems with the ability to learn automatically from example data without human intervention. As computers learn directly from data, the need for humans to specifically program is reduced. This is a substantial benefit because the computer can automatically generate models independent of explicit human programming, which can be time-consuming and costly. This machine learning cost advantage, combined with large amounts of freely available government data (BLS, 2019; ONET, 2019), may provide low-cost tools for curriculum developers and education administrators to determine the most appropriate level and mix of occupational skills for meeting student goals and optimizing wage potentials. Job and career seekers can use these models to generate estimates of how well their skills should be compensated by the job market.

To date, there have only been limited human resource (HR) department moves to apply analytics to discover any inherent insights possibly lurking in large HR data stores. According to Cappelli, Tambe, and Yakubovich. (2019), only 22 percent of firms say they have adopted analytics (machine learning or otherwise) in human resources. And while 71 percent of companies see people analytics as a high priority, progress has been slow (Collins, Fineman, & Tsuchida, 2017).

Some organizations, typically third party online recruiting firms, rather than organizational HR departments, have built models to estimate salaries. Their models are based on very specific inputs such as an individual's location, education, experience, and specific skillset.

As can be seen in Table 1, these models accept very specific skills related to each particular occupation (such as 'Apache Kafka' and multithreading skills for Java programmers), detailed qualifications specific to individuals (such as years of experience). These inputs contrast with our model inputs, which use general skills (see Table 2) that are applicable across almost all standardized occupations. And our models do not require detailed attributes tied to specific individuals. Thus, our models are more likely to be better suited for individuals who may not have significant experience in a field or occupation-specific skillsets and are more interested in how combinations of general occupational skillsets impact salaries.

The third-party salary predictors are typically supported by extensive surveys and large big data collections – which can be expensive and maintenance heavy. Our models leverage free, readily available data already collected by the U.S. government. The data collection and maintenance costs are borne by the U.S. government rather than the company, which may incentivize HR departments to dip their toes into the analytics pond.

The third-party models use their own, much more detailed, lists and definitions of occupations. These occupations differ between the various third party models. The government data, on the other hand, uses a list of standardized occupational job codes that can be easily associated and integrated with the same job codes data across the federal government. This makes analysis across multiple government datasets quicker and easier.

Salary Predictor	Sample Input Data	Reference
Dice	Job Title; Location; Years of Experience; Detailed Skills for Job Title (e.g., Multithreading, Apache Kafka, etc...)	Dice, 2019
PayScale	Job Title; Years in Field; City; Foreign Language; Skills Critical to Job Title; Type of employer; Work Venue	Payscale, 2019
Glassdoor	Current Employer; Employer Location; Job Title; Years of Relevant Experience; Base Salary; Additional Compensation; Highest Level of Education; University; Major/Concentration	Glassdoor, 2019
Salary	Job Title; Location; Education; Years of Experience; Direct Reports; Reports To; Performance	Salary, 2019
Indeed	Job Title; Location	Indeed, 2019
SalaryExpert	Job Title; Location; Highest Level of Education; Major/Field of Study; School Attended; Skills or Specialties; Current Employer Name; Years of Relevant Experience; Current Employer; Years of Experience; Salary	SalaryExpert, 2019
Adzuna	Title, FullDescription, LocationRaw, LocationNormalized, ContractType, ContractTime, Company, Category, SalaryRaw, SalaryNormalised, SourceName,	Adzuna, 2019

Table 1: Salary Predictors

Workforce skills are related to earnings, increased productivity, and economic well-being (Prada & Rucci, 2016). Our models can be used to help both organizations and individuals manage their respective skill inventories by predicting the wages of different skillset combinations. For organizations, wage data can help determine what different skill sets will cost in wages, and budget accordingly.

Individual workers must take responsibility for managing their careers and concern themselves with maintaining their skills to adapt to fast-changing work environments. Hall (1996) describes the protean career, which is driven by the individual rather than the organization. The protean career requires the individual to personally acquire and use an identifiable set of skills. Laar, van Deursen, van Dijka, and Haan (2017) argue that the current workplace requires highly skilled workers who must address increasingly complex and interactive tasks. Employees not only need excellent technical preparation; they also need sufficient skills to adapt to the changing requirements of the job (Carnevale & Smith, 2013). For individuals, our models can provide wage information to help them formulate which skill combinations to pursue.

The NN and MLP models developed with these occupational skills and annual wage training and testing data can be used to provide insights to help management make decisions about how to best manage critical skill needs that are driven by the rapid, and constant, advance of technology. These insights can support the management of pertinent employee skills needed to keep apace and stay relevant, as well as supporting the budgeting of salaries for these skillsets.

Methods

We used the data described in this section to develop, train and test a number of different NN and MLS machine learning models that are also described in this section.

Data Description

Our NN and MLS models use the same independent variables data as inputs to predict the dependent variable, annual wages, as described below.

Independent Variables:

The input data are Skill Level Ratings (which range from 0 to 100) associated with over 900 occupations. In addition to the Skill Level Ratings, the related importance of each skill (Skill Importance Rating) to each occupation is also provided, and also ranges from 0 to 100. For example, a Marketing Manager requires mathematic skills at a Skill Level Rating of 45 out of 100, with a Skill Importance Rating of 44 out of 100. Contrast that occupation with the Refuse and Recyclable Material Collectors occupation which requires a mathematics Skill Level Rating of 7 out of 100, with An Importance Rating of 10 out of 100. Thus, the level of mathematics skills required of Marketing Managers is much higher and significantly more important for them, than for Refuse and Recyclable Material Collectors.

Our analysis looked at different ways of using the occupations Skill Level Ratings and their associated Skill Importance Ratings to assess their impact on performance. That is:

- **Skill Level Ratings Only:** Only the Skill Level Rating data are used as inputs to train the models. The associated Skill Importance Rating data was not used.
- **Skill Level Ratings Only (When Skill Importance > 50):** Occupational Skill Level Ratings were only used if their associated Skill Importance Ratings were greater than 50 (out of 100). The associated Skill Importance Rating data was not used.
- **Skill Level * Skill Importance:** the Skill Level Ratings were multiplied by the Skill Importance Ratings for each occupation.

The Skill Level Ratings data and their related Skill Importance Ratings were provided by the Occupational Information Network (O*NET), which provides this data for 35 different occupational skills that are organized into ten basic skills, one complex problem-solving skill, four resource management skills, six social skills, three systems skills, and eleven technical skills (Table 2).

O*NET is the primary source of U.S. occupational information (ONET, 2019). Their database contains hundreds of standardized and occupation-specific descriptors on hundreds of occupations. The database is developed under the sponsorship of the U.S. Department of Labor (DOL).

<p><u>Basic Skills</u></p> <ul style="list-style-type: none"> • Active Learning • Active Listening • Critical Thinking* • Learning Strategies* • Mathematics* • Monitoring • Reading Comprehension* • Science* • Speaking* • Writing* <p><u>Social Skills</u></p> <ul style="list-style-type: none"> • Coordination • Instructing* • Negotiation* 	<ul style="list-style-type: none"> • Persuasion • Service Orientation • Social Perceptiveness <p><u>Technical Skills</u></p> <ul style="list-style-type: none"> • Equipment Maintenance • Equipment Selection • Installation • Operation and Control • Operation Monitoring* • Operations Analysis* • Programming* • Quality Control Analysis • Repairing* • Technology Design • Troubleshooting 	<p><u>Complex Problem-Solving Skills*</u></p> <p><u>Systems Skills</u></p> <ul style="list-style-type: none"> • Judgment & Decision Making* • Systems Analysis* • Systems Evaluation <p><u>Resource Management Skills</u></p> <ul style="list-style-type: none"> • Management of Financial Resources* • Management of Material Resources* • Management of Personnel Resources • Time Management*
--	---	--

Table 2: Occupational Skills

The O*NET data is based on several hundred descriptive ratings from O*NET questionnaire responses from sampled workers, occupational experts, and occupation analysts (Fleisher & Tsacoumis, 2012). Randomly sampled workers answer one of several possible 30-minute questionnaires. For occupations where it would be difficult to sample workers, such as those in remote locations, occupation experts are identified and sampled from professional and/or trade association membership lists. The occupation experts complete multiple questionnaires. In addition to the questionnaires completed by workers and occupation experts, additional ratings are provided by occupation analysts. Responses from all three sources (workers, occupation experts, and occupation analysts) are used to provide complete information for each occupation.

Dependent Variable:

The dependent variable output of the models are estimates of the wages for an employee with the different combinations of Skill Level Ratings. The annual wage data used for training and testing the models is provided by the U.S. Bureau of Labor Statistics (BLS) Occupational Employment Statistics (OES). The BLS is the principal federal agency responsible for measuring labor market activity, working conditions, and price changes in the economy (BLS, 2019).

The BLS OES provides nationwide, average annual wage data for close to 1000 Occupations in the U.S. across numerous industries and job types. Their dataset includes wage data from high-level positions, such as Chief Executives, to low wage employees, such as fast-food workers. The highest and lowest average annual wages listed in the dataset are \$265,990 and \$21,230, respectively.

The BLS OES data is gathered from program semi-annual surveys to estimate wages by national, state, and metropolitan areas. We used the 2017 national data. Wages for the BLS OES survey are defined as straight-time, gross pay, which is exclusive of premium pay. The survey covers all full-time and part-time wage and salary workers in nonfarm industries. The wage data includes base rate, cost-of-living allowances, guaranteed pay, hazardous duty pay, incentive pay, on-call pay, and tips. The wage data does not include back pay, jury duty pay, overtime pay, severance pay, shift differentials, nonproduction bonuses, tuition reimbursements, or data from self-employed workers. The BLS OES surveys approximately 180,000 to 200,000 establishments every six months - so it takes three years to completely collect the full sample of 1.2 million establishments (BLS, 2019).

The BLS occupational annual wage data was integrated with the O*NET occupational skills data via job codes to create a consolidated dataset that links the 35 skillsets to the wage data for 967 occupations. The initial dataset was then pruned to 937 occupations to eliminate missing and incomplete data.

Model Development

We used the open-source RStudio integrated development environment (<https://www.rstudio.com>), along with the R programming language to build several NN and MLR models from the occupational data. There are over 25 different NN architectures available to choose (Mehta, 2019). We selected the backpropagation NN architecture because it is a widely used, classical architecture that is suitable for the type of analysis we are performing. The model development steps are summarized below.

1. Integrate 35 occupational skills data (independent variables) with associated national wage data (dependent variable)
2. Run correlations against all the independent and dependent variables.
3. Run MLR using the 35 occupational skills and their annual wages.
 - Use a variety of combinations of Skill Level Ratings and Skill Importance Ratings for the analysis.
4. Build, train, test, and run backpropagation NNs (using 35 occupational skills and their wages)
 - The NN structure hyperparameters consisted of 35 input nodes, with a 3 node hidden layer, another single node hidden layer, and a single output node. Two hidden layers were chosen because a two hidden layer architecture can approximate any smooth mapping to any accuracy (Heaton, 2017). Training time was not a significant issue given the reasonable size of our dataset. A number of models were tried using different structure hyperparameters. The architecture with a 3 node hidden layer and another single node hidden layer provided

- performance similar to more complex NNs with larger numbers of nodes in the hidden layers.
- The default logistic activation function is used in the nodes.
5. Run stepwise forward and stepwise backward MLR algorithms to select the skills with the highest prediction power. This step reduced the number of inputs from 35 to 19.
 6. Run MLR using the reduced number of inputs (using 19 occupational skills and wage data).
 7. Build, train, test, and run backpropagation NNs against reduced inputs (using 19 occupational skills and wage data).
 - The NN consists of 19 input nodes, with a 3 node hidden layer, another single node hidden layer, and a single output node. The reasons these hyperparameters were selected are the same as those explained in Step 4.
 - The default logistic activation function is used in the nodes.
 8. Perform cross-validation for the NN and MLR models.
 - Train on 90% of randomly selected records and test on the remaining 10% of records.
 - Calculate the accuracy, where accuracy is determined by:
 $1 - \text{MEAN}(\text{ABS}(\text{Actual Wages} - \text{Predicted Wages})/\text{Actual Wages})$
 - Repeat ten times with randomly selected training and testing records.
 - Calculate the average accuracy over the 10 experiments.

Results

The correlation matrix showed that the correlation coefficients between the independent variables and the dependent variable (wages) ranged from -0.26 to 0.73.

MLR and both stepwise forward and backward linear regression techniques were employed to improve the prediction power of the inputs. This reduced the number of independent variables from 35 to the 19 that are marked with asterisks (*) in Table 1.

The results of running the models are provided in Table 3, which provides the accuracy for the NN and MLR models, along with the R-Squared values for the MLR. In each case, these are averages of running the models ten times with random training/testing using 90/10 data splits. These values are provided for inputs of the full complement of 35 Skill Ratings, as well as the reduced set of 19 Skill Ratings.

35 Skills			19 Skills		
	NN	MLR		NN	MLR
	Accuracy	Accuracy R ²		Accuracy	Accuracy R ²
Skill Level * Skill Import	0.764	0.761 0.647	Skill Level * Skill Import	0.786	0.761 0.647
Skill Level Only	0.761	0.715 0.62	Skill Level Only	0.705	0.715 0.62
Skill Level Only (Importance > 50)	0.710	0.671 0.409	Skill Level Only (Importance > 50)	0.728	0.671 0.409

Table 3: Results

Figure 1 visually compares the predicted wage values against the actual wage data for the models trained against the subset of 19 skill rating variables. A perfect fit between actual and predicted salaries would show the dots falling perfectly on the line for a mean square error (MSE) of zero. Figure 2 provides the same visual information for the regression predictions.

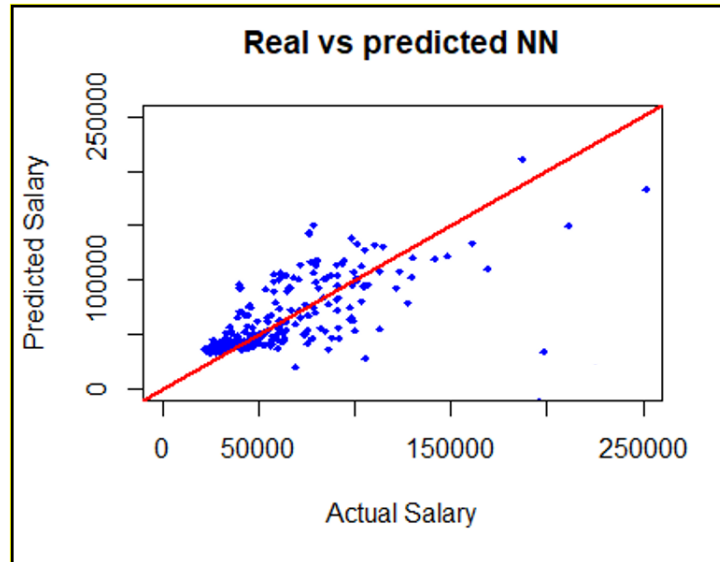


Figure 1: NN Predicted vs. Actual

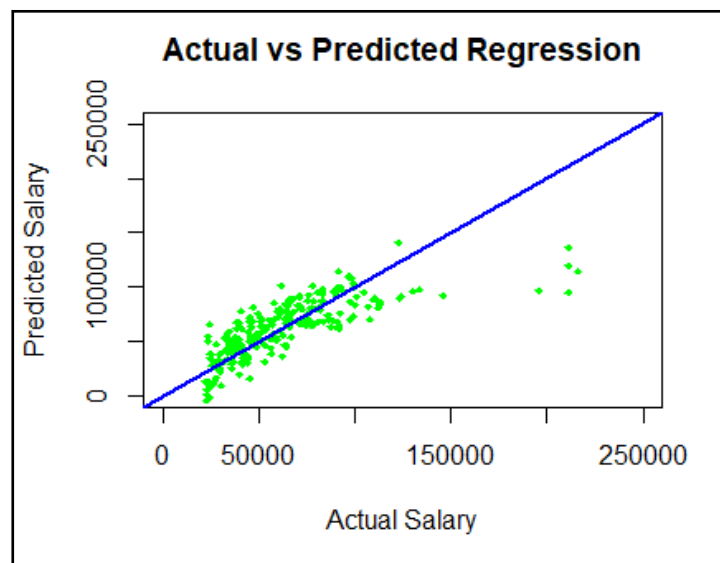


Figure 2: MLR Predicted vs. Actual

Conclusions

We have created initial models that use freely available standardized skills and wage data that can be used by organizations and individuals to determine the wage values of different combinations of general skillsets. Since the data is collected and maintained by the Federal government, this may provide a relatively low-cost way for HR departments to initiate HR analytics. These models may prove useful for the analysis of wages based on general skillsets (see Table 2), that apply across nearly all

occupations rather than the types of very specific occupation-related skills listed in Table 1. These more general models may be better suited for individuals, such as students, who do not have significant work experience or occupation-specific skillsets.

References

- Adzuna 2013 Kaggle Competition. (2013). Job Salary Prediction, Retrieved on August 30, 2019 from <https://www.kaggle.com/c/job-salary-prediction/overview/timeline>.
- Bureau of Labor Statistics (BLS). (2019). Occupational Employment Statistics Overview. Retrieved on June 3, 2019 from https://www.bls.gov/oes/oes_emp.htm.
- Cappelli, P., Tambe, P., & Yakubovich, V. (2019). Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. Retrieved on August 22, 2019 from <https://ssrn.com/abstract=3263878>.
- Carnevale, A., & Smith, N. (2013). Workplace basics: The skills employees need and employers want. *Human Resource Development International*, 16(5), 2013.
- Collins, L., Fineman, R., & Tsuchida, A. (2017). People analytics: Recalculating the route. 2017 Global Human Capital Trends. Retrieved on August 30, 2019 from <https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2017/people-analytics-in-hr.html>.
- Dice.com. (2019). How Much Are Your Skills Worth? Retrieved on May 26, 2019 from <https://www.dice.com/salary-calculator>.
- Fleisher, M. & Tsacoumis, S. (2012). O*NET Analyst occupational skills; ratings: Procedures update. Report FR-11-67.
- Glassdoor.com. (2019). Know Your Worth. Retrieved on June 13, 2019 from <https://www.glassdoor.com/Salaries/know-your-worth.htm>.
- Hall, J. (1996). Protean careers of the 21st century. *Academy of Management Executive*, 10(4).
- Heaton, J. (2017). The Number of Hidden Layers. Retrieved on August 2, 2019 from <https://www.heatonresearch.com/2017/06/01/hidden-layers.html>.
- Indeed.com. (2019). Search and Compare Salaries. Retrieved on June 16, 2019 from <https://www.indeed.com/salaries>.
- Laar, E., van Deursen, A., van Dijka, J., & Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. *Computers in Human Behavior*, 72, 577-588.
- Mehta, A. (2019). A comprehensive guide to types of neural networks. Digital Vidya Data Analytics BLOG. Retrieved on June 29, 2019 from <https://www.digitalvidya.com/blog/types-of-neural-networks/>.
- National Center for O*NET Development (ONET). (2019). About O*NET. Retrieved on June 15, 2019, from <https://www.onetcenter.org/overview.html>.

Payscale.com. (2019). Pay Report. Retrieved on June 22, 2019 from <https://www.payscale.com/my/survey/job>.

Prada, M. & Rucci, G. (2016). Guide to Workforce Skills Assessment Instruments. Inter-American Development Bank Technical Note IDB-TN-1070.

Salary.com. (2019). What Am I Worth? Retrieved on June 14, 2019 from <https://www.salary.com/>.

SalaryExpert.com. (2019). Calculate your Salary. Retrieved on June 20, 2019 from <https://www.salaryexpert.com/salarycalculator>.