PREDICTING AI/ML SALARIES AND EVALUATING ERROR DIRECTION

COMPARING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

FERNANDO ISCAR

THESIS PROPOSAL
DATA SCIENCE & SOCIETY

STUDENT NUMBER

2103157

COMMITTEE

dr. Bruno Nicenboim

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

September 18th, 2023

WORD COUNT

1499

PREDICTING AI/ML SALARIES AND EVALUATING ERROR DIRECTION

COMPARING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

FERNANDO ISCAR

1 PROJECT DEFINITION, MOTIVATION & RELEVANCE

Research on salary prediction has helped professionals and employers make informed and equitable decisions. However, prior studies have mainly focused on gender pay inequalities and compensation trends across different industries, ignoring exploration into the domains of Artificial Intelligence (AI) and Machine Learning (ML). For more than a decade, these fields have been exponentially developing and are in great demand in the business and technology industries (Alekseeva et al., 2021). While AI and ML algorithms involve sophisticated mathematics, their simplicity in implementation has led to widespread AI adoption and an extension of the labor market to non-experts (Joshi, 2020, p. 247). This study leverages a high-dimensional dataset, which offers a rich variety of features not traditionally accounted for in similar research, to compare both conventional and advanced regression methods and identify the most suitable model for revealing compensation in the AI-ML sector. Moreover, this work seeks to explore bias in prediction error, identifying pertinent features and discrimination patterns beyond the well-researched gender pay gap, aiming to improve results comprehension.

As it facilitates fair and equal employment practices, salary estimation and its bias exploration is an issue that matters to society. This study's findings may prove to be a useful asset by raising awareness of this subject in the AI-ML field, a rapidly growing and already highly competitive job market.

This study is scientifically relevant as it will address the lack of research on salary prediction in the AI-ML field. The employment of both conventional and cutting-edge regression methods will allow for a comparison of each algorithm's performance in this particular scenario. Moreover, by detecting relevant traits and potentially discriminatory tendencies, the results will be more accurate and detailed while also contributing to the body of research on wage prediction.

2 LITERATURE REVIEW

Earlier studies showcased that salary is associated with factors regarding individual demographics, educational background, and professional experience, such as age, gender, education level, country, work experience, or job title. This data is usually collected through surveys (Kablaoui & Salman, 2022; Matbouli & Alghamdi, 2022), which often involve dealing with missing values due to the sensitivity of pay data, the non-symmetric distribution of salaries, and the lack of inflation adjustments. (Özer et al., 2022).

2.1 *Modeling salary prediction*

In the arena of salary prediction, regression analysis serves as a cornerstone, utilizing algorithms such as Multiple Linear Regression (MLR), Lasso and Ridge Regressions, Decision Trees (DT), Random Forests (RF), Extreme Gradient Boosting (XGBoost), Multi-Layer Perceptron (MLP) and Support Vector Regression (SVR). The performance is primarily measured through Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2. While MLR continues to be a robust traditional approach, recent research highlights the superior performance of tree-based methods like RF and XGBoost, especially in capturing non-linear relationships, thereby establishing them as State-Of-The-Art (SOTA) options (Brandwijk, 2021; Jain et al., 2022; Özbalta et al., 2022; Özer et al., 2022). Furthermore, last studies underscore the potential of MLP in yielding reliable results, pointing to its emerging prominence in this sphere (Matbouli & Alghamdi, 2022; Özer et al., 2022; Wang, 2022).

Choosing regression analysis for salary predictions allows us to quantitatively evaluate the influence of various factors on salary outcomes. Hence, this method will be selected over classification approaches, wherein the focus usually shifts to predicting wage brackets or identifying underpaid individuals, potentially overlooking subtle nuances in the data (Kablaoui & Salman, 2022; Martín et al., 2018).

Moreover, while existing research predominantly utilizes traditional analytical methods, this study embraces a forward-thinking approach by continuing the empirical landscape through the incorporation of MLP, tapping into its potential for refined and advanced predictions.

Diverging from previous research that largely centered on the aforementioned factors, this study capitalizes on a richer dataset to explore a broader spectrum of features and reveal more granular insights into the AI-ML salary landscape. While existing methodologies from other sectors and broader IT roles (Martín et al., 2018; Özer et al., 2022) serve as a reference, they underscore the existing gap in research tailor-made for the AI-ML sector, which this research aims to fill.

2.2 *Error direction and potential discriminatory variables*

The bias exploration is often conducted by Exploratory Data Analysis (EDA) (Fry et al., 2021) and frequently centered on the gender pay gap (Brandwijk, 2021). The examination of salaries for IT Developers in the United States by Peslak et al. (2022) motivated the idea of this research to explore biases that can impact salary prediction accuracy. In the mentioned paper, the authors utilized survey data from 2021 and conducted EDA to identify salary biases toward specific variables. The study reveals patterns of discrimination based on gender and mental health status. However, no predictions were made using ML algorithms or additional features concerning physical disability, certification possession, or sexual orientation.

Fry et al. (2021) uncovered how variables like ethnicity or mental health can influence the analysis, with Black and Hispanic women working in STEM jobs earning the lowest wages. Recognizing this gap, our study introduces a "baseline model" using standard data related to demographics, academic and professional information, to initiate salary predictions in the AI-ML domain. This framework will enable us to pinpoint the best-performing algorithm and then iteratively refine it by incorporating lesser-explored features. This approach is designed to unravel nuanced discriminatory patterns, providing a richer comprehension of the dynamics steering salary allocations in the industry.

3 Research Strategy & Research Questions

Hence, the broader research question of this investigation is the following:

> *Given an extensive developers survey dataset, how accurately is it possible to predict AI/ML job roles' salary, identifying error biases,*

by using a baseline and the successive nested models which include factors such as demographics, employment, education, and experience?

This main question can be extended to answer other sub-questions:

RQ1 *To what extent is it possible to explain more variance and reduce prediction errors from a baseline model consisting of the features "Country", "Job type", "Education", "Job title", "Company size", "Age", "Experience" and "Annual Salary" by using different ML and DL algorithms such as MLR, LASSO, RIDGE, RF, XGBoost, and MLP?*

RQ2 *Selecting the best overall performer regression algorithm from the baseline model, how do additional features, including "Remote work", "Certifications", "Coding as a hobby", "Years Coding", "Gender", "Sexual orientation", "Ethnicity", "Physical disability", and "Mental disability", impact the prediction error and variance explained by the aforementioned model?"*

RQ3 *To what extent can patterns of discrimination be detected after including "Gender", "Sexual orientation", "Ethnicity", "Physical disability" and "Mental disability" in the previously introduced baseline model?*

4 METHODOLOGY AND EVALUATION

4.1 *Dataset Description*

The study utilized data obtained from Stack Overflow, a developers' online platform ("Stack Overflow Annual Developer Survey", 2022). The dataset comprised 73,268 responses from developers from 180 different countries, who participated in an annual survey during the period of May to June 2022. Stack Overflow employed measures to ensure the reliability of the responses, such as filtering out responses that took less than 3 minutes and replacing the top 2% of salaries with threshold values. The dataset consists of 81 columns, encompassing demographic factors, education, personal traits, employment characteristics, experience, and coding questions. The survey's specificity facilitates the examination of salary regression using variables not yet further explored, such as physical health or sexual orientation. Despite the dataset's richness, it contains a considerable quantity of outliers in the salary column, numerous null values, and class imbalances, such as only 30% of the respondents being female.

4.2 *Algorithms and Software*

Wolpert and Macready (1997) demonstrated the *"No Free Lunch Theorem"*, which states that we cannot assume which model is best without evaluating all of them. Therefore, this study will include, in the baseline model, a performance comparison among the fine-tuned *RQ1* algorithms, commonly used in salary forecasts, which include conventional and SOTA methods. This will enrich empirical research by examining their efficacy in addressing a high-dimensional, non-linear problem.

To prevent data leakage and tackle class imbalance, we will utilize stratified sampling in the partitioning of data into training and test sets, ensuring a proportional representation of the target variable in both. This process will be carried out before the feature engineering step, with each set undergoing independent and identical processes. Given the significant reduction in rows post-cleaning, a validation set will be deemed unnecessary. Feature engineering steps will encompass merging columns, categorizing, and logically imputing missing values, everything using Python programming language.

4.3 *Evaluation Method*

Drawing from prior regression research, this study will apply RMSE, R^2, and adjusted R^2 for model assessment. RMSE, advantageous for penalizing substantial errors and sharing the target variable's units, takes precedence over MSE in salary predictions. R^2 serves as a common metric in scientific fields, effectively depicting the variance explained by the models (Chicco et al., 2021).

For the baseline model, we will leverage K-Fold Cross-Validation (KF-CV) using R^2 and RMSE to identify the best-performing algorithm. This algorithm will then shape the succeeding nested models. Both baseline and nested models will be tested on a set of unseen data to ensure a robust analysis. As we advance to the nested models, we will enhance the assessment by including adjusted R^2 and residual plots, granting a detailed insight into the error dynamics.

5 MILESTONES AND PLAN

With the hope of this proposal being approved, the final thesis version development will start from week 39th, around the presentation date. From this moment and during the week after, modeling and results evaluation will be carried out, after working in the previous steps time ago. The thesis draft will be written within 41st and 44th week, leaving the rest of the time till draft submission for fine-tuning the content quality.

REFERENCES

Alekseeva, L., Azar, J., Giné, M., Samila, S., & Taska, B. (2021). The demand for ai skills in the labor market. *Labour Economics*, *71*, 102002. https://doi.org/10.1016/j.labeco.2021.102002

Brandwijk, M. (2021). *Analysing the gender pay gap in it through salary prediction: A data driven approach* [Doctoral dissertation, Tilburg University]. http://arno.uvt.nl/show.cgi?fid=157118

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://peerj.com/articles/cs-623/

Fry, R., Kennedy, B., & Funk, C. (2021). Stem jobs see uneven progress in increasing gender, racial and ethnic diversity. *Pew Research Center*, 1–28. https://www.pewresearch.org/science/wp-content/uploads/sites/16/2021/03/PS_2021.04.01_diversity-in-STEM_REPORT.pdf

Jain, A., Jain, S., Pancinovia, N. M., & George, J. P. (2022). A non-linear approach to predict the salary of nba athletes using machine learning technique. *2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT)*, 1–5. https://doi.org/10.1109/TQCEBT54229.2022.10041664

Joshi, A. V. (2020). Machine learning and artificial intelligence, 247. https://link.springer.com/content/pdf/10.1007/978-3-030-26622-6.pdf

Kablaoui, R., & Salman, A. (2022). Machine learning models for salary prediction dataset using python. *2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 143–147. https://doi.org/10.1109/ICECTA57148.2022.9990316

Martín, I., Mariello, A., Battiti, R., & Hernández, J. A. (2018). Salary prediction in the it job market with few high-dimensional samples: A spanish case study. *International Journal of Computational Intelligence Systems*, *11*(1), 1192–1209. https://doi.org/10.2991/ijcis.11.1.90

Matbouli, Y. T., & Alghamdi, S. M. (2022). Statistical machine learning regression models for salary prediction featuring economy wide activities and occupations. *Information*, *13*(10). https://www.mdpi.com/2078-2489/13/10/495

Özbalta, E., Yavuz, M., & Kaya, T. (2022). National basketball association player salary prediction using supervised machine learning methods. *Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation: Proceedings of the INFUS 2021 Conference, held August 24-26, 2021. Volume 2*, 189–196.

Özer, Ş. D. İ., Ülke, B., Daniş, F. S., & Orman, G. K. (2022). Salary prediction via sectoral features in turkey. *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 1–6. https://doi.org/10.1109/INISTA55318.2022.9894130

Peslak, A., Ceccucci, W., & Jones, K. (2022). The effect of mental illness on compensation for it developers. *Proceedings of the Conference on Information Systems Applied Research ISSN*, *2167*, 1528. https://proc.conisar.org/2022/pdf/5736.pdf

Stack overflow annual developer survey [Accessed on February 15, 2022]. (2022). https://insights.stackoverflow.com/survey

Wang, G. (2022). Employee salaries analysis and prediction with machine learning. *2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, 373–378. https://doi.org/10.1109/MLISE57402.2022.00081

Wolpert, D., & Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893