Salary Predictions Using Machine Learning Models

By Unknown unkown

WORD COUNT 1409
TIME SUBMITTED 19-MAR-2023 02:33PM
PAPER ID 97778155

Salary Predictions Using Machine Learning Models

<Student Name>

University of North Texas

Professor <Professor Name.>

CSCE <FULL COURSE NAME>

18th March 2023

Introduction

Career is an important part of our lives in this modern day civilization, more so that it can affect us at various levels of our lifestyle which are influenced by the money we earn. Most of us, choose financial stability while looking for work, or earning a living but how would one exactly know, how much money they could expect from taking up a specific role in a specific field?

Today for reference, one might look up the expected compensation or salary on various popular web pages like Glassdoor, where the users are actual employees who post their pays online to help others have a reference of pay at a specific firm while being in a specific role. Primary problem is that the pay posted has high variance, based on Location, and Firm they work at, it would be pretty had to narrow down or even have a clear expected salary by just looking at those numbers.

I propose a solution to this by providing a Machine Learning model that predicts expected salary of a candidate based on their field of work and experience. This could be useful for people of various fields not pertaining to just the tech sector in order understand the pay scale of roles in different fields. This project would act as a good reference to people searching for a reference pay scale and don't want to keep searching for websites, and signing up by providing their emails in order just get a quote.

My solution would be to select a rich dataset, and train various machine learning models like Linear Regression, Lasso Regression, Decision Trees, and Random Forest. These models would be trained on the same training set in order to obtain the best working model for the dataset, based on their metrics like MSE, r2_score, hyper parameters used etc., I will then decide the model that is the best fit, and then probably provide how that model is able to perform a good job. My reference model will be chosen from one of the models mentioned above to compare my predictions later.

Literature Review

❖ *Using Linear Regression to Predict Employee Salary*

This is a research paper with a self-explanatory title authored by *D. M. Lothe, T. Prakash, P. Nikhil, P. Sanjana and P. Vishwajeet*. As the title suggests the authors used a Linear Regression model to fit on a dataset and then make the predictions for the target variable i.e. Salary. They also plot a graph to make it easier for the user to see where their pay stands.

The authors conclude that they found the MSE of 357, and accuracy to be 76%. Personally I think these numbers could be better. And also the fact that the authors did not try and elaborate how various models would have performed on the data set they had with them.

However this paper will act as a good reference, to try and decide my baseline model to be Linear Regression.

❖ *Using job description to predict probable salaries*

Although this is different approach and cannot be used upon the dataset I have, as it does not have any job description field within it, I found this work really interesting. I see many employers posting roles on LinkedIn but rarely ever post a quote for expected compensation. The authors of this paper, solve this very interest problem by using Regression models like Lasso Regression, Neural Network Regression and Random Forest Regression. Which puts me on the right track of model selection. I personally found the paper to be well written, they did provide a lot of useful insights, and one such interest conclusion is the fact that for their dataset, the Random Forest model outperformed the Neural Network, even though they used Drop out method for generalizing this network i.e. dropping a neuron to avoid over dependence on one of the weight.

This shall act as a good reference as to understand how my Regression Model might work. And test it with and without hyper-parametric tuning.

❖ ***Salary prediction using Machine Learning Regression Models***

I found this paper by *D. Sayan, B. Rupashri and M. Ayush* to be very similar to the first paper mentioned within my Literature Survey, this is because both of these papers have the same motivation and similar conclusions. The authors used Linear Regression for linear plots, and tried to use polynomial regression for the other, and then predict the user salary.

Even this project as a graph output for user to understand where their probable earning stand. Within conclusion the authors also mention that K nearest neighbors can be used for the same task to be able to obtain better accuracy.

To my opinion, these papers seemed pretty simple of an approach to the same task at hand. I would like to dwell into this model analysis a bit more. The authors did not provide any model statistics like MSE, Accuracy. However their Methodology was well explained.

❖ ***Salary Prediction in the IT job Market]***

This has to be the closest and best fit paper to the dataset I have, the authors *M. Ignacio, M. Andrea, A. H. Jose* did a wonderful job on describing their methodology, while mentioning how they performed data collection, cleaning and describing for the set. They performed automatic feature selection on this dataset, and then later performed a grid search to find the optimal hyper-parameters, this was such a good approach to the problem.

I am really interested in this approach as I had the same plan while I was going through my dataset. To first clean, then transform, train and then perform a grid search to each of my models. The dataset that the authors create by crawling also considers an important feature that we have in common, i.e. 'Education level'. They also used most of the models I mentioned in my high level approach to this task in addition to a MLP (Multi-Layer Perceptron.)

Their dataset description gave me a thorough idea of what I can look for within my dataset and what features might have a greater influence on predictions that could be made.

Data Exploration

The dataset I decided to work with is titled as, 'Employee-Salaries' which is available on Kaggle. This data set has more than 1 Million samples within it. And has a usability score of 7.6/10 which is pretty moderate. This dataset was posted by a user named *Mukesh Manral*. And the user has not provided the source of this set or the methods used to compile this dataset.

Upon initial review of the dataset, here are my observations:

- This dataset has total 9 features (including the target feature).
- Here are the 8 features that are used to find the target variable: 'jobId', 'companyId', 'jobType', 'degree', 'major', 'industry', 'yearsExperience', 'milesFromMetropolis'
- Out of these nine features, 6 of the features are categorical in nature. Which would need us to perform some encoding before training our models.
- One of the advantage of having such a large dataset is that I can divide this dataset into multiple sets and compare results in each of the test set, which would be more optimal in determining the best performer.
- There are no null values. Which makes it easier by not having to discard samples.
- I assume that the features 'yearsExperience', 'industry' would have higher correlation as they do in real life while determining pay.

With these metadata insights, I aim to proceed with some preprocessing and creating a baseline model to see how it would work and make observations. On the whole this dataset is more than just sufficient to train a well performing model.

References

Lothe, D. M., Tiwari, P., Patil, N., Patil, S., & Patil, V. (2021). Salary prediction using machine learning. *International Journal of Advance Scientific Research and Engineering Trends*, 6(5), 199-211. https://doi.org/10.51319/2456-0774.2021.5.0047. Retrieved from http://ijasret.com/VolumeArticles/FullTextPDF/842_47._SALARY_PREDICTION_USING_MACHINE_LEARNING.pdf

Jackman, S. D., & Reid, G. J. (2013). Predicting Job Salaries from Text Descriptions. *University of British Columbia. STAT 540*. https://doi.org/10.14288/1.0075767

Das, S., Barik, R., & Mukherjee, A. (2020). Salary Prediction Using Regression Techniques. *Social Science Research Network*. https://doi.org/10.2139/ssrn.3526707

Martín, I., Mariello, A., Battiti, R., & Hernández, J. L. (2018b). 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. https://doi.org/10.2991/ijcis.11.1.90

M. (2022, June 2). *Salary_Prediction*. Kaggle. https://www.kaggle.com/code/mukeshmanral/salary-prediction

Salary Predictions Using Machine Learning Models

ORIGINALITY REPORT

8%
SIMILARITY INDEX

PRIMARY SOURCES

1	arno.uvt.nl Internet	33 words — 2%
2	www.researchgate.net Internet	25 words — 2%
3	download.atlantis-press.com Internet	19 words — 1%
4	Emine Kambur, Cüneyt Akar. "Human resource developments with the touch of artificial intelligence: a scale development study", International Journal of Manpower, 2021 Crossref	13 words — 1%
5	core.ac.uk Internet	8 words — 1%
6	journal.50sea.com Internet	8 words — 1%
7	www.coursehero.com Internet	8 words — 1%
8	www.doria.fi Internet	6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF