Unveiling Salary Trends: Exploring Machine Learning Models for Predicting Data Science Job Salaries

Jiayi Zhu

School of Information Management, University of International Business and Economics, Beijing, 100029, China
202212016@uibe.edu.cn

Abstract. The challenging employment landscape is characterized by a significant disparity between high job expectations and intense competition, often resulting in a discrepancy between applicants' self-assessment and enterprise standards. Within the vast array of job information available, salary emerges as a crucial concern for job seekers. This paper delves into the Kaggle salary prediction dataset, a rich repository that serves as a valuable resource for understanding trends and patterns in salary expectations across various industries, with a specific focus on data science job predictions. This paper systematically introduces three prominent models of machine learning—deep learning, decision trees, and random forests—and elucidate their applications in the context of salary prediction. By providing a detailed overview of each model's workings and methodologies, author aims to offer readers a comprehensive understanding of their potential utility in predicting salary outcomes. Through rigorous analysis, this paper meticulously evaluates the strengths and weaknesses inherent in each model, shedding light on their respective performance metrics and predictive capabilities. In addition, this article outlines the future prospects of machine learning in the field of salary prediction and emphasizes trends and potential avenues for further research. The importance of a comprehensive approach is emphasized, which combines the insights of multiple models and can significantly improve the accuracy and effectiveness of predictions in real-world scenarios.

Keywords: Employment Landscape, Salary Prediction, Machine Learning, Data Science.

1 Introduction

The continuous expansion of enrollment in China's colleges and universities has gradually transformed elite education into mass education for the benefit of every ordinary citizen, but this has led to the increasingly severe employment situation of fresh graduates. One of the key reasons for the increasing difficulties of graduates' employment lies in the high expectation of employment, which is mainly manifested in the high expectation of salary, but this is a strong contradiction with the employment market environment is getting more and more competitive. There is

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_20

often a deviation between the self-competence model that students think and the selection standards of enterprises, which leads to the situation that fresh graduates pile up in high-paying enterprises and often ignore the ordinary enterprises that are more likely to be successful in applying for jobs, which contributes to the social phenomenon of difficult employment to a certain extent. Therefore, it is in the interest of most of method (sweatshops excepted) to know the fair market rate for a position so that it can minimize mismatches and unsuccessful interviews [1]. And it is important to accurately assess the employment situation and set reasonable salary expectations.

Basically, machine learning (ML) is a branch of computer science that enables computer systems to understand and process data in a manner similar to humans [2]. In short, ML represents a form of artificial intelligence that leverages algorithms or techniques to identify patterns within raw data [3]. The core objective of ML is to empower computer systems to learn from past experiences without direct human intervention. It has a wide range of applications, including finance, healthcare, education, natural language understanding, image recognition, fault diagnosis, network information services, and many other industries and fields. Among the typical algorithms for machine learning are Decision Tree Algorithm (DTA), Random Forest Algorithm (RFA), Boosting and Bagging Algorithms (BBA), Support Vector Machine Algorithm (SVM), Neural Network Algorithms (NNA), Convolutional Neural Networks (CNN), Deep Learning Algorithms (DLA), Gate Recurrent Unit (GRU) and so on.

Recently, in the field of salary prediction research, there has been a number of studies using techniques in ML. Pornthep used data mining techniques to compare both decision trees and random forests to generate a graduate salary prediction model for individuals with similar training attributes. The system exhibited remarkable efficacy in bolstering students' motivation and fostering a positive outlook towards their future [4]. Notably, Sayan Das, Rupashri Barik, and Ayush Mukherjee employed regression methodologies to predict an individual's salary after a specific period [5]. To achieve this objective, the Biz Reach Artificial Intelligence (AI) Technology Group introduced a hybrid deep architecture, Bidirectional-GRU-CNN, which enables them to attain superior accuracy compared to other leading models. Utilizing Deep Learning techniques, Viroonluecha has achieved a significant milestone in developing a predictive model that accurately forecasts the monthly salary of job seekers in Thailand, efficiently tackling a regression challenge with numerical outcomes [6]. This system analyzed personal profile data spanning five months from a prominent job search website. Consequently, the Deep Learning model exhibited robust performance.

The primary objectives of this study are to summarize existing concepts and backgrounds related to salary prediction, analyze core technology—ML—in detail, focus the research on data science by introducing a dataset, and discuss the strengths, weaknesses of machine learning in salary prediction, and its future development prospects.

2 Methodology

2.1 Dataset Description and Preprocessing

This dataset contains job postings from 2017-2018 on Glassdoor [7]. The comprehensive dataset includes details such as the job title, estimated salary, job duties, overall rating, company name, its geographical location, headquarters, the number of employees, years in operation, ownership structure, the specific industry, the broader sector, financial revenue, and a list of its competitors. After careful consideration, a comprehensive list of the most significant features within this dataset has been compiled. Using this dataset, it was possible to determine which factors have the greatest impact on salaries in the data science field and to identify which states and cities offer the highest paying data science related jobs. Finally, the salaries posted for data science jobs were predicted based on job descriptions.

2.2 Proposed Approach

In response to the employment challenges due to high salary expectations, this research plans to explore and enumerate a series of research methods based on the field of machine learning using existing datasets that are designed to analyses and predict salary levels in data science related industries. Through an in-depth study and analysis of three commonly used techniques, namely decision trees, random forests and deep learning, readers will be able to better understand the effectiveness of these techniques in salary prediction and compare their respective strengths and weaknesses.

In decision trees, the structure of the tree makes it clear which features have the greatest influence on salary prediction. However, decision trees may face the problem of overfitting, so appropriate pruning strategies are needed to avoid this. Random forests, an integrated learning technique, can be used to further improve prediction performance. By constructing multiple decision trees and integrating them, random forest can make full use of the information present in the data to enhance the stability and generalization of the model. Deep learning is also widely used in salary prediction. Deep learning has powerful feature extraction and representation learning capabilities, which can automatically extract features useful for salary prediction from raw data. After completing the presentation of the models, a detailed comparative analysis of the three techniques will be presented. This helps researchers (see in Fig. 1) in gaining deeper insights into the practical performance of these techniques, offering invaluable guidance for subsequent research endeavors and applications.

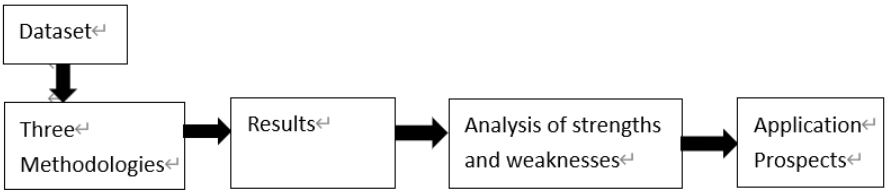

Fig. 1. The flow chart of research.

Decision Trees. Decision trees, proposed in the 1960s, remain an efficient method for data mining, widely adopted across disciplines for their user-friendliness, clarity, and resilience to missing values [8].

A decision tree methodology segregates the population into hierarchical subdivisions, resembling the structure of an inverted tree that consists of a root, intermediate internal nodes, and terminal leaf nodes. This algorithmic approach is non-parametric, meaning it does not rely on predefined parametric structures, enabling it to proficiently manage vast and intricate datasets. With a sufficiently large sample size, the research data can be bifurcated into two distinct sets: a training dataset, which serves as the cornerstone for constructing the decision tree model, and a validation dataset, aiding in determining the optimal tree structure to yield an efficient final model [9]. The Fig. 2 provides a visual representation of the decision tree's structural composition.

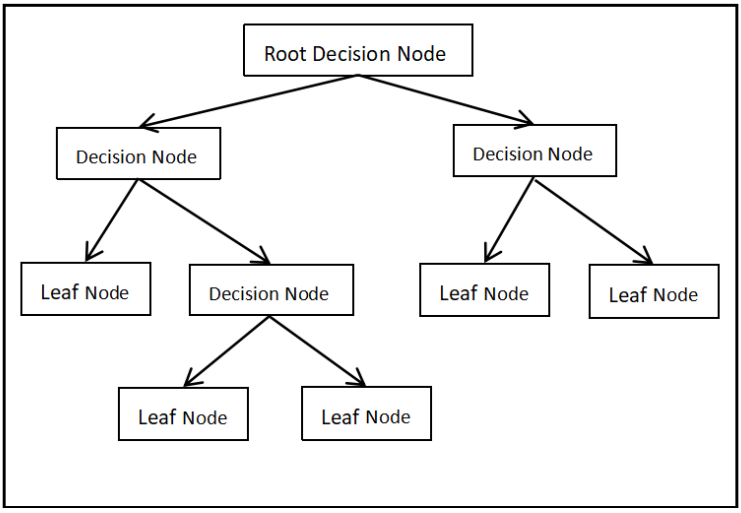

Fig. 2. A Decision Tree.

Generally, the construction of a decision tree comprises three key stages: feature selection, tree generation, and pruning. Prior to feature division, it's crucial to introduce the concept of entropy, which in information theory and probability statistics, quantifies the uncertainty of random variables. Higher entropy indicates greater uncertainty, while lower entropy signifies less uncertainty and purer information.

The following details several indices employed in selecting features for decision trees. Information gain is used to quantify the extent to which the uncertainty of information about category Y is reduced after information about feature X is known. Generally, a higher information gain indicates a greater "purity enhancement" achieved through partitioning by a specific attribute. However, the magnitude of

information gain is relative to the training dataset and lacks absolute significance. Difficult classification problems, characterized by high empirical entropy in the training data, tend to yield higher information gain values, and the reverse holds true. This poses a paradox with the original intent of information gain and entropy. Random forests, on the other hand, utilize the "Gini Index" to choose partitioning attributes, where a lower Gini index signifies a higher dataset purity.

Tree generation is typically a recursive process. However, during this recursive generation of decision trees, overfitting can occur. Overfitting occurs when an excessive focus is placed on enhancing the accuracy of classifying training data during the learning process, leading to excessively intricate decision trees. To address this, the generated decision tree is often considered for simplification, a technique referred to as pruning. Some straightforward pruning algorithms include those based on loss functions, cost functions, and similar metrics. Decision tree models can be used by following steps: First the dataset is imported and the file is read to obtain the data; second data cleaning is performed to preprocess the data, this may include dealing with missing values, outliers, and converting categorical variables (e.g. job title, industry, etc.) to numeric variables. Next the construction of a decision tree model proceeds by recursively partitioning the dataset into progressively smaller subsets, where each split is determined by the values of a selected feature, and the resulting subsets are then utilized to train the decision tree model using the training data. During training, the model learns how to predict salary based on input features. Subsequently, test data is employed to assess the model's performance, leading to optimizations based on the evaluation outcomes, which involve fine-tuning the model's parameters, such as the maximum depth allowed for trees and the minimum number of samples required for leaf nodes. Finally, the optimized decision tree model is used to make salary predictions on the new data. The features of the new data are input into the model, which makes predictions based on the learned rules and outputs the predicted salary values. The decision tree model is also susceptible to overfitting during use, so it is essential to exercise caution and regulate the model's complexity when employing it.

Random Forest. Leo Breiman introduced random forests in the 2000s as a methodology for creating an ensemble of predictors utilizing a collection of decision trees, where each tree is grown within randomly chosen subsets of the overall data [10]. This method is a commonly used tool for classifying high-dimensional data, which can rank candidate predictors by a built-in variable importance measure [11]. Random forests can be effectively utilized in a diverse range of regression tasks, encompassing both nominal (categorical), metric (numerical), and survival analysis response variables. Random forests are capable of both classification and regression tasks. In the ensemble algorithm, random forests utilize the concept of bagging, implying that at each iteration, a new training set is constructed by randomly selecting n samples from the original training set. Train M sub-models using this new set. For classification, employ a voting mechanism and assign the final category as the one with the most votes. For regression, average the predictions of the sub-models to obtain the final predicted value.

The random forest employs the decision tree as its fundamental building block. By aggregating a significant number of decision trees, the random forest framework is established. In the salary prediction experiment, it is necessary to first confirm that the data has been cleaned and preprocessed, and then carry out feature engineering according to the data set to enhance the prediction ability of the model, and use the processed data set to train the random forest model. During the training phase, the random forest constructs numerous decision trees, where each tree undergoes training on a randomly sampled subset of data and features. In this way, each decision tree learns different aspects of the data and improves overall predictive performance through integration. Random forests have some parameters that can be adjusted, including the number of trees, the number of features considered on each node, the maximum depth of the tree. Suitable evaluation metrics can serve as a means to assess the predictive capabilities of the model.

Deep Learning Algorithms. Deep learning is comprised of computational models that incorporate numerous processing layers, which facilitate the learning of data representations across varying levels of abstraction. These techniques have propelled significant advancements in areas like visual object recognition, speech recognition, object detection, and numerous other domains, including genomics and drug discovery. Deep learning uncovers intricate patterns in extensive datasets through the utilization of backpropagation algorithms, which offers direction on the adjustment of internal parameters within a machine.

As depicted in Fig. 3, notable advancements have been made in image, video, speech, and audio processing, whereas recurrent networks have excelled in processing sequential data, including text and speech [12].

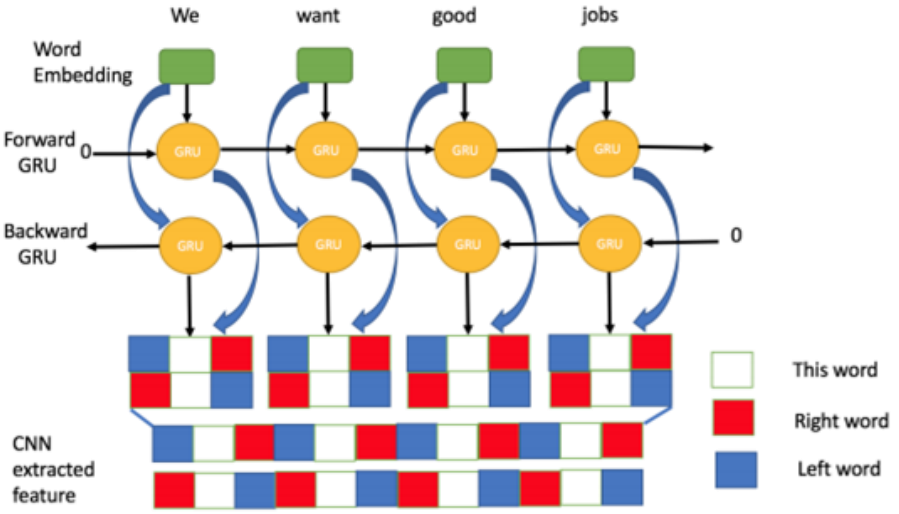

Fig. 3. Bidirectional-GRU-CNN model.

3 Results and Discussion

In the conducted study, by comparing the capabilities of deep learning, decision tree model and random forest model in salary prediction experiments, it is found that each method has its unique advantages and limitations. Hence, selecting the most suitable model based on the specific dataset and requirements is imperative.

3.1 Results

In the experiments conducted to implement a salary prediction system aimed at enhancing student motivation through data mining techniques [4], the prediction accuracy of each technique is depicted in Table 1. Among the evaluated metrics of Recall, Precision, and F-measure, K-Nearest Neighbors (KNN) emerged as the technique with the highest overall accuracy, achieving 84.69%, while Multilayer Perceptron (MLP) recorded the lowest at 38.08%. Decision Trees (J48) secured a percentage of 73.96%, while SVM garnered 43.71%, Naive Bayes (NB) attained 43.63%, and MLP registered 38.08%. Notably, KNN excelled in overall prediction accuracy, outperforming the other models. In terms of Recall, KNN delivered the best results for salary classes exceeding 18,000 baht, scoring 87.10%. Similarly, KNN topped the Precision metric for salaries above 18,000 baht with 86.30%, and achieved the highest F-measure, scoring 86.70% for the same salary class. This indicates that KNN performed particularly well in predicting salaries exceeding 18,000 baht.

Table 1. Summary of salary prediction models.

Class(baht)	Recall (%)				
	KNN	MLP	J48	NB	SVM
<13500	84.90	45.30	75.80	50.90	37.00
13501-15300	83.40	41.00	71.80	49.70	58.00
15301-18000	83.70	0.50	71.80	21.70	23.60
>18000	87.10	68.70	76.90	53.40	56.70
	Precision (%)				
	KNN	MLP	J48	NB	SVM
<13500	84.90	45.30	75.80	50.90	37.00
13501-15300	83.40	41.00	71.80	49.70	58.00
15301-18000	83.70	0.50	71.80	21.70	23.60
>18000	87.10	68.70	76.90	53.40	56.70
	F-measure (%)				
	KNN	MLP	J48	NB	SVM
<13500	84.90	45.30	75.80	50.90	37.00
13501-15300	83.40	41.00	71.80	49.70	58.00
15301-18000	83.70	0.50	71.80	21.70	23.60
>18000	87.10	68.70	76.90	53.40	56.70
Accuracy (%)	84.69	38.08	73.96	43.63	43.71

3.2 Analysis of Strength and Weakness

In terms of accuracy, decision trees perform better in salary prediction. By building a tree structure, decision trees are able to visualize the key factors that affect salary and make salary predictions based on these factors. Random forests usually achieve higher accuracy than decision trees in salary prediction. By building multiple decision trees and integrating them, it is able to make full use of the information in the data and reduce the risk of overfitting, thus improving the accuracy of the prediction. Deep learning, on the other hand, is able to automatically learn complex feature representations in the data by constructing deep neural network models, thus predicting salary levels more accurately.

In terms of stability, the decision tree model performs generally well. Since it is susceptible to noise and outliers in the data, it sometimes suffers from overfitting, leading to a decrease in the model's performance on the test set. Random Forest performs well in terms of stability. As it adopts the idea of integrated learning by constructing multiple decision trees and voting or averaging them to get the final prediction, it can effectively reduce the influence of noise and outliers on the model and improve its stability. Deep learning performs generally in terms of stability. Due to its high model complexity, it is prone to overfitting. In order to lower the risk of overfitting, it is usually necessary to use some regularization techniques, data enhancement and other technical means.

In terms of computational efficiency, decision trees have a higher level of efficiency. Decision trees only need to traverse the data set once, so their training time is relatively short. The computational efficiency of random forest is relatively low. Since it needs to construct multiple decision trees and integrate them, its training time is usually longer than that of decision trees. In addition, when the dataset is very large, the training time of random forest may increase further. Since deep learning its model complexity is high, it requires a large number of computational resources to train the model. In addition, the training time of deep learning models is usually longer than that of traditional machine learning models, especially when dealing with large-scale datasets. In summary, these three models have their own advantages and disadvantages in the field of salary prediction. In practical applications, it is necessary to choose the appropriate model according to the specific dataset and needs. For example, when it is necessary to quickly construct models and make salary predictions, decision trees or random forests can be chosen; when it is necessary to obtain higher prediction accuracy, deep learning models can be considered. It is also necessary to pay attention to the complementarity between different models, and multiple models can be used in combination through methods such as integrated learning to obtain better prediction results.

3.3 Application Prospects

The deep neural network model has strong feature learning and generalization capabilities, but requires large data resources for training; the decision tree model is easy to understand and interpret, but is prone to overfitting; and the random forest

model excels in prediction accuracy, but is computationally expensive and not easy to interpret. Different salary prediction methods are applicable to different datasets and problem scenarios. In practical applications, it is necessary to choose the appropriate prediction model based on the characteristics of the data and the needs of the problem. For complex problems such as salary prediction, integrated learning models (e.g., Random Forest) usually provide better prediction results, but there is a trade-off between computational cost and model interpretability. In practice, the prediction results of multiple models can be combined and integration strategies can be used to improve the overall prediction performance.

4 Conclusion

The primary aim of this study is to outline a series of research methodologies rooted in the field of machine learning, to introduce and elucidate existing datasets, and to deepen comprehension of the efficacy of three commonly employed techniques—decision trees, random forests, and deep learning—in wage prediction. Through thorough investigation and analysis, the study seeks to compare their respective strengths, weaknesses, and future development prospects. The experimental findings demonstrate that, in practical applications, combining prediction results from multiple models and employing an integration strategy can enhance overall prediction performance. Given the contemporary challenge of employment difficulties arising from high wage expectations, the enthusiasm for research in salary prediction is expected to persist in the future. The next stage of research objectives will likely involve integrating existing research techniques and exploring the interconnection between salary prediction and other domains.

References

1. Nguyen, D.P.T., Shinsuke, S., Zhongsheng, W.: Salary prediction using bidirectional-gru-cnn model. Assoc. Nat. Lang. Process, 2019.
2. Mullar, A.: Introduction to Machine Learning using Python: A guide for data Scientist, 2019.
3. Buczak, Anna, L., and Erhan, G.: A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications surveys & tutorials 18(2), 1153-1176 (2015).
4. Khongchai., Pokpong, S., and Pornthep.: Random forest for salary prediction system to improve students' motivation. In 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). IEEE (2016).
5. Ayush, M., Das., Rupashri, B., and Sayan.: Salary prediction using regression techniques. Proceedings of Industry Interactive Innovations in Science, Engineering & Technology (2020).
6. Phuwadol., Thongchai, K., and Viroonluecha.: Salary predictor system for thailand labour workforce using deep learning. In 18th International Symposium on Communications and Information Technologies (ISCIT). IEEE (2018).

7. Jerome, F., Robert, T., Trevor, H.: The elements of statistical learning: data mining, inference, and prediction. 2009.
8. Salary prediction, https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor/data, last accessed 2022/5/3.
9. Yan-Yan, S., and Ying, L.U.: Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130 (2015).
10. Biau, Gérard.: Analysis of a random forests model. The Journal of Machine Learning Research, 13(1), 1063-1095 (2012).
11. Breiman, Leo.: Random forests. Machine learning, 45, 5-32 (2001).
12. Cun, L., Geoffrey, H., Yann., and Yoshua, B.: Deep learning. nature 521(7553), 436-444 (2015).