Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230910

Comparison of deep-learning and conventional machine learning algorithms for salary prediction

Ziyuan Feng[1, †], Zixian Liu[2, †] and Yibo Yin[3, 4, †]

[1]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China
[2]College of Engineering, Hong Kong Polytechnic University, HongKong, 999077, China
[3]School of Economics and Management, Tongji University, Shanghai, 200092, China

[4]dolado@tongji.edu.cn
[†]These authors contributed equally.

Abstract. Salary is an integral part of contemporary life. With the large-scale use of machine learning, it has become possible to predict salaries with machine learning. Previous researchers have used random forest algorithms to solve this problem, however, there is a research gap in using a neural network to solve this problem. Therefore, the research topic of this paper is to use convolutional neural networks (CNN) and datasets on Kaggle to predict salary. The research methodology of this paper is as follows. First, the Kaggle dataset is divided into the train-dataset and the test-dataset. After preprocessing the data, two kinds of features are obtained. The features will be transformed into two two-dimensional matrices. Next, the matrices were used to train two CNNs separately. These two CNNs will be connected together to get the predicted salary by fully connected layers and Relu activation functions. After training the CNN, the study called the test dataset to verify the accuracy of the model. Similarly, the study used a random forest model for prediction. Finally, the comparison of the two results showed which algorithm was better. The study found that the error rate of the CNN was 0.0732 and its variance was 0.1899. The error rate of the random forest was 0.2437 and its variance was 0.8285. From the results, CNN is better than random forest in terms of accuracy and stability. Therefore, using CNN for salary prediction has a high probability of getting better results.

Keywords: salary prediction, deep learning, convolutional neural network.

1. Introduction

Salary is what people get for their mental or physical work. For the individual, salary can sustain personal or family expenses. For society, salaries have the function of redistribution of labor resources. The social system achieves the optimal allocation of labor resources through the regulation of salary [1]. Therefore, it is important for contemporary workers to get the expected salary based on the job posting.

With the promotion of machine learning algorithms, machine learning has implicitly taken root in modern society. It is widely used for face recognition [2], speech-to-text [3], predicting traffic

DOI: 10.54254/2755-2721/6/20230910

conditions [4], product recommendations in online shopping [5], and self-driving of cars [6] et al. All of these machine learning-based products have deeply influenced people's lives and work. According to previous related works [7, 8], most salary prediction projects use a random forest model to predict salary. Random forest is a classifier with multiple decision trees, the output of which is decided by multiple classes of the output of the individual trees. The forest consists of decision trees, and every decision tree is unrelated to the other. After obtaining the forest, the new input samples need to be classified and let each decision tree in the forest determines which class the samples belong to. Finally, all the judgments are combined to arrive at the final result.

The random forest model has many advantages in that it can handle very high dimensional data and does not have to do the feature selection. Because Random Forest is not sensitive to missing values and outliers, it maintains accuracy even when a large proportion of features are missing. However, random forests have some disadvantages, they can be over-fitted for problems with high noise levels.

But can convolutional neural network (CNN) be used for this purpose? CNN is a class of feedforward neural networks containing convolutional computation with deep structure [9]. CNN is built in a way that mimics the visual perception mechanisms of living things. Research into CNN began in the 1980s and 1990s. CNN is one of the representational algorithms for deep learning. It has a representational learning capability, which is why it is also known as a "Shift-Invariant Artificial Neural Network". The earliest CNN to emerge were time-delay network and LeNet-5. After the twenty-first century, CNN has been applied to computer vision and Speech recognition with the improvement of numerical computing devices.

The advantages of CNN are also significant, as it shares convolutional kernels and can easily handle high-dimensional data. CNN does not require manual selection of feature values, and feature classification is effective. But, the demand for sample size and GPU is huge. Once neural networks are proven to have better performance in the project of salary prediction, the related field will get a faster and more effective research avenue.

This study was based on the dataset of "Job Salary Prediction" on the Kaggle website [10]. First, the Kaggle dataset was divided into training and testing datasets. After some data processing of the dataset, two types of features were obtained. These two types of features were processed as two two-dimensional matrices that can be accepted by CNN. Next, the study used the previous matrixes to train two CNNs separately. The two CNNs were connected together to get the predicted payoffs through the fully connected layer and the Relu activation function. After training the CNN model, a test dataset was taken to verify the accuracy of the model. Similarly, the study used a random forest model for prediction and obtained the accuracy of the model. Finally, the prediction results of CNN and random forest were compared to conclude which model was better.

2. Method

2.1. Dataset

The dataset was collected from various job search websites in the UK with information about individuals seeking employment by web crawlers. It is mainly composed of a large number of rows representing individual job ads, and a couple of fields about each job ad. All of the data is collected from the real world, and therefore it is obviously subject to lots of noise in the real world, including ads that are doesn't focus on the UK, and salaries that are falsely stated.

So, data cleaning is of crucial importance before the author begins to build their models to solve this problem. In the data cleaning, the author has done the following: standardize the case of characters, remove the rows with duplicate and missing data, split the original data into text variable and categorical variable, and finally show the cleaned data using a bar chart. The data distribution is displayed in Figure 1.

DOI: 10.54254/2755-2721/6/20230910

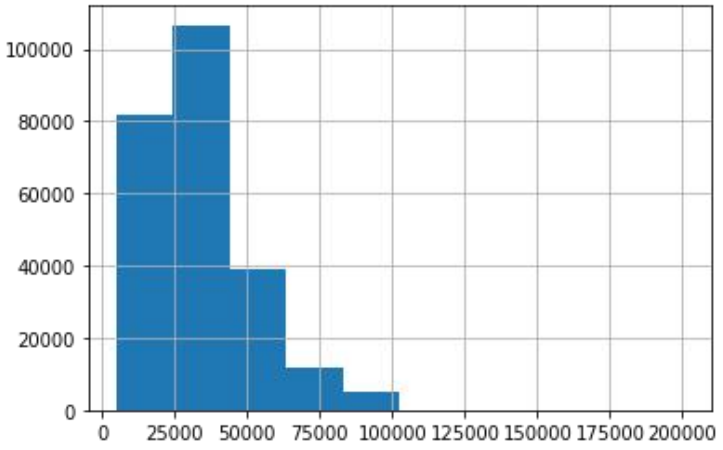

Figure 1. salary distribution of original dataset.

For text data collected in the questionnaire, the One-hot encoding was used to transform it into a two-dimensional matrix, which was then handed over to the subsequent convolutional neural network for training. In the original dataset, different methods are adopted to transform the text into vectors.

The attributes are separated into categorical variables and text variables. For those categorical variables that have only a limited number of types, such as contract type and contract duration, dummy variable classification is used to transform it into a 2-d matrix. Furthermore, for those text variables with more attributes, more text is combined into one full attribute text, using the text-vectorization method for processing.

The text vectorization is meant to make natural language input acceptable to a neural network. In the real practice, it takes a set of irregular strings and transform them into a low-dimensional matrix (usually a two-dimensional matrix) by some algorithm (e.g., one-hot coding, word2vec, Neural Network Language Model). that can be processed by a neural network.

2.2. Theoretical foundations

A convolutional neural network is primarily a network that specializes in processing data structures similar to meshes (i.e., two-dimensional data structures, matrix data structures). A review of convolutional neural network models is presented below.

2.2.1. Convolutional layer. In its general form, convolution is a particular mathematical operation on two functions of real variables for which the form is generally defined as follows. Let f(x), and g(x) both be integrable functions on R and calculate the following integral:

$$\int_{-\infty}^{+\infty} f(t)g(x-t)dt \quad (1)$$

In the process of theoretical research and practical application of machine learning, the input is generally a multi-dimensional array of data sets, and the kernel function is usually the parameter of the multi-dimensional array learned by convolutional neural network through the input dataset and the output dataset with its parameters called "tensor".

2.2.2. Pooling layer. The typical convolutional network structure is generally a three-stage structure in a general convolutional neural network. The first layer of the structure computes multiple convolutions at the same time to produce a set of linear activation responses. The second layer uses the output of the first level as input through a nonlinear activation function. In the last layer, a pooling function is called to adjust the output of this convolutional layer more accurately. The pooling function uses the overall statistical characteristics of the neighboring outputs at a given location to substitute the output of the convolutional layer at that location, also known as undersampling or downsampling. Theory summarizes that its central role is to compress the number of data. Pooling functions that are

DOI: 10.54254/2755-2721/6/20230910

commonly used in practice include: the maximum pooling function, minimum pooling function, average pooling function, and random pooling function.

2.2.3. Activation layer. The convolution operation is a linear operation, i.e., y=ax+b. If simply stack convolutional layers (or fully connected layers) to increase the depth of the network, the whole network is still a linear network, which means that the whole network can be reduced to a linear operation of convolutional layers (or fully connected layers), so the effect is the same no matter how many layers are stacked. It is then necessary to perform nonlinear operations between the convolutional layer and the convolutional layer. The activation layer is the one that performs the nonlinear operations between convolutional and convolutional layers so that the network is nonlinear. The activation layer operates on the exact size of the input data as the output data.

The Sigmoid function is the most frequently used activation function at the beginning of the traditional neural network and deep learning field, and its mathematical expression is

$$f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

2.2.4. Random forest. Random forest model is a kind of supervised learning algorithm. Its main idea is to create multiple decision trees by selecting samples and features randomly. At the end of the algorithm, each decision tree is allowed to vote independently on the final result, and the classification result is based on the classification result of the majority of the decision trees, thus achieving the effect of allowing multiple decision trees to collaboratively participate in the classification decision. Compared with the traditional single decision tree classification, random forest can effectively reduce the risk of overfitting and reduce the error brought by outlier, and it can find the key features quickly because of its great randomness. In addition, the data needs to be more comprehensive based on real scenario where there is a large amount of interference and huge amount of duplicate and missing data. Compared with those more complex algorithms (e.g., genetic algorithms), random forest can handle this type of data at a relatively lower cost.

2.3. Model constructing

In this paper, two models are constructed, convolutional neural network and random forest model, to predict the salary. The accuracy of the models is judged by comparing the prediction accuracies obtained after feeding the previously processed datasets into the two different models separately.

2.3.1. Convolutional neuron networks. Based on the previous processing of the dataset, the variables are classified into text variables and categorical variables. Furthermore, two neural networks are constructed to process two types of variables separately and merge at the end. Finally, the outputs of the two networks are combined together and pass through a fully connected layer to obtain the final prediction result.

In the process of constructing the convolutional network, an embedding layer is firstly used to downscale the text matrix previously encoded with unique heat, then stack three convolutional layers with three maximum pooling layers, and finally perform an average pooling operation to output the result to the fully connected layer. In contrast, the convolutional network II is directly stacked with three fully connected layers. Finally, the outputs of the two separate neural networks are combined, and then two consecutive fully connected operations are performed to obtain the final prediction results. Moreover, the neural network structure is shown in Figure 2.

DOI: 10.54254/2755-2721/6/20230910

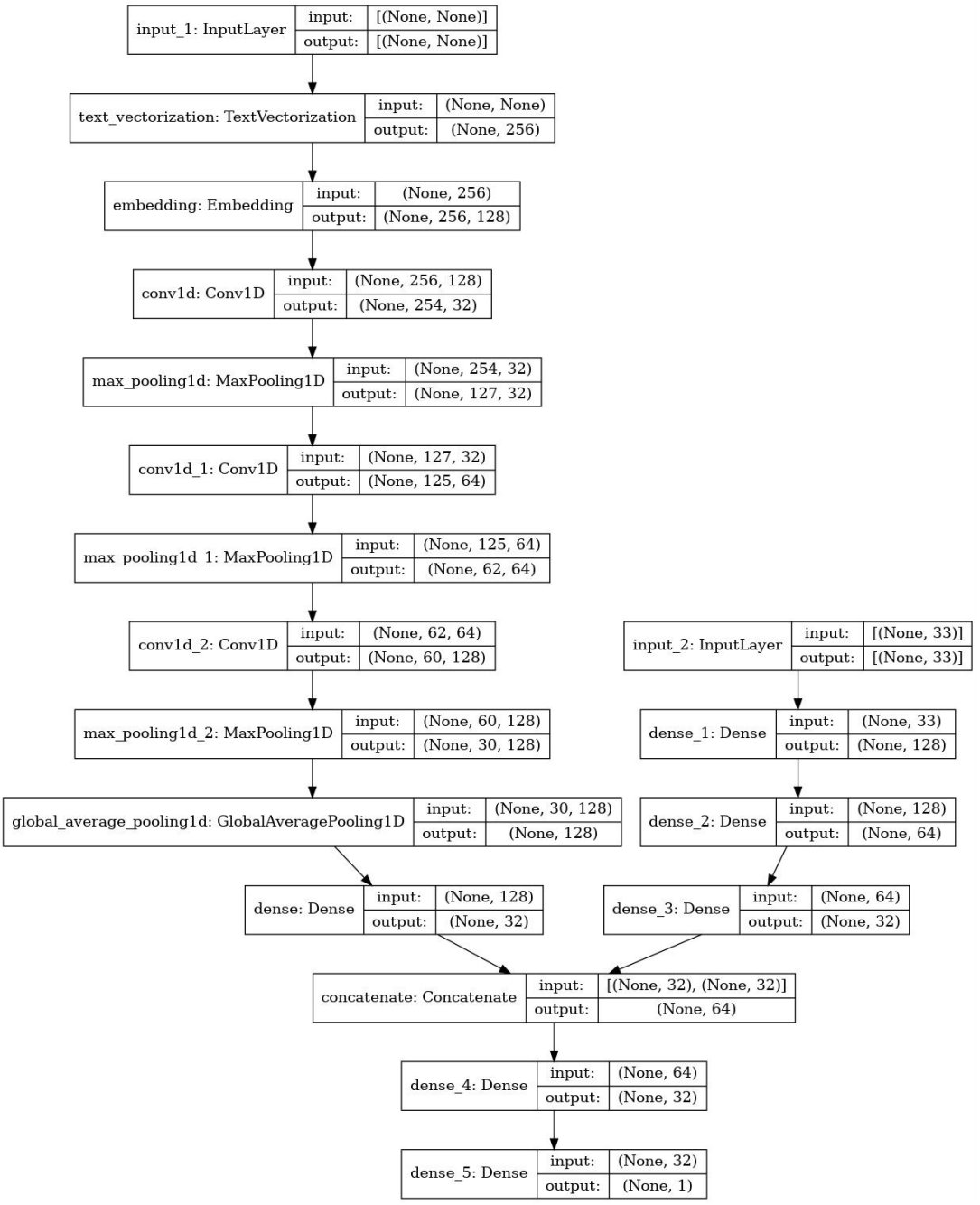

Figure 2. The architecture of the Convolutional Neural Network.

2.3.2. Random forest. The standard "RandomForestClassifier" from the SKlearn library is used as the classifier here, and the author tune the random forest model's parameters using GridSearchCV to get the best prediction, which mainly involving the following parts.

1)Adjusting the min_samples_split in the random forest

2) Adjusting the max depth in the random forest.

3) Adjusting the max feature in the random forest.

4) Adjusting the random_state of the model

DOI: 10.54254/2755-2721/6/20230910

2.4. Evaluation matrix

This paper set a real salary above 30,000 as a high salary and below 30,000 as a low salary. Finally, the following metrics are used to test and compare the prediction results of the constructed CNN model and random forest model for the above two groups.

To verify the accuracy of the CNN Salary prediction model, the Author use mean absolute percentage error Ae and variance absolute percentage error Ve to evaluate the model prediction accuracy, which is calculated as follows:

$$e = \left|\frac{S_0 - S_p}{S_0}\right| \times 100\%$$

$$Ae = \frac{1}{n}\sum_{i=1}^{n} e_i$$

$$V_c = \frac{1}{n}\sum_{i=1}^{n} (A_e - e_i)^2$$

In the formula, S_o is the real salary and S_p is the predicted salary, Ae is the absolute percentage error, n is the total number of test sets.

3. Result

After predicting the same test dataset using the CNN model and the classical random forest model, the predicted salary dataset for the two models is obtained separately. The author will divide the higher and lower salary ranges, and use the mean absolute percentage error and the variance fundamental percentage error to represent the prediction accuracy and stability of the two models for the higher and lower salary groups, respectively.

The test set contains data from 0 - 100000 real salary (henceforth abbreviated as RS). 30,000 is used as the dividing line and classifies those with fewer than 30,000 as a lower salary and those with more than 30,000 as a higher salary. The data distributions of both are demonstrated in Figure 3 and Figure 4 respectively.

Figure 3. Data distribution of high salary.

DOI: 10.54254/2755-2721/6/20230910

Figure 4. Data distribution of low salary.

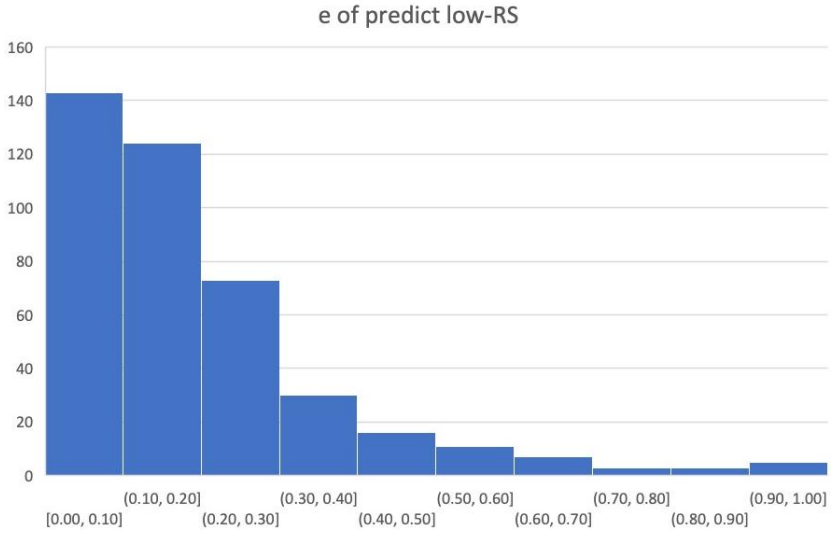

Figure 5. Error of low salary.

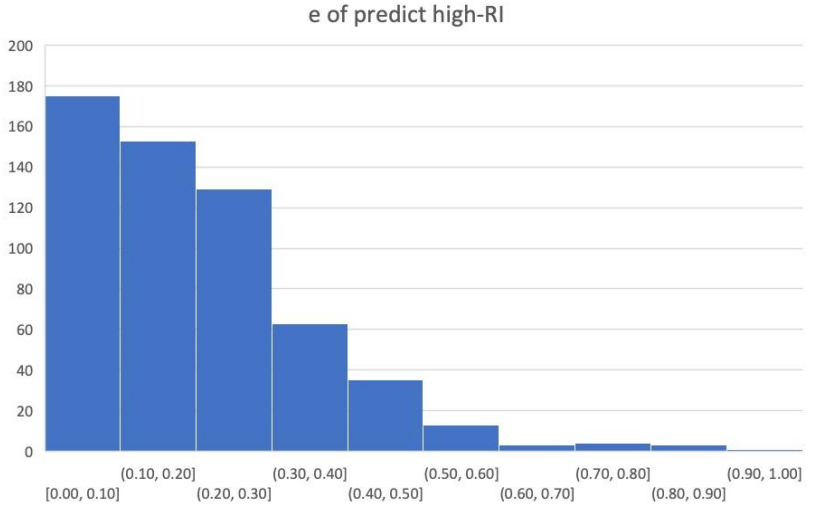

Figure 6. Error of high salary.

DOI: 10.54254/2755-2721/6/20230910

Table 1. Result comparison of CNN and random forest.

	Low RS		High RS	
	Mean of e	Variance of e	Mean of e	Variance of e
CNN	0.2129	0.054	0.1986	0.023
Random forest	1.1294	1.283	0.3249	0.08

The result comparison of CNN and the random forest is shown in Figure 5 and Figure 6 and further summarized in Table 1. The authors found that the results of the random forest model are acceptable for the high-salary group, and the error for the low-salary group is far beyond the acceptable level, which may be due to the low value of RS for the low-salary group and the enormous absolute value of the error of the second random forest model, resulting in such a result.

The CNN model has a mean absolute percentage error of about 20% for both low and high-salary groups, while the random forest model predicts a significant error rate for the low-salary group, and for the high-salary group, his mean absolute percentage error is also more than 10% higher than the CNN model.

Moreover, it can also be seen from the variance absolute percentage error that the CNN model has more stable prediction results.

4. Discussion

This work used CNN & random forest model to predict salaries and compare the results generated by these models. The author can find that the random forest model's average error rate and variance are very high for the low-income group. There are two possibilities: this may be because the random forest model needs a large number of samples to achieve the desired performance, and the lack of samples for the low-income group causes this. Another possibility is that the formula for calculating the error rate causes the low-income group to be more prone to higher error rates. Namely, deeper research is needed to explain the generation of such high errors. It can be compared by increasing the sample size or introducing new comparison models, such as other CNN models. In addition, the CNN model uses a different number of layers in processing category class attributes and complex text class attributes, which makes the work more focused, and the data block passing is used when resuming the Transflow dataset, which improves the efficiency of data processing, and it could be observed that these two decisions improve the efficiency of the model work.

As for the data, it should be noted that because the dataset on which this study is based is only from Kaggle, there are limitations in the kinds of features in the dataset. First of all the data in the dataset only has job-related information such as location, job category, etc. The data is missing the requirement for the personal information of the applicant. In job advertisements, the requirement of personal ability also determines the salary level. For example, if a company requires that an applicant for a position must be a graduate of a good university, then the corresponding salary should be higher. Second, there are some garbled codes in this dataset that have not been processed into recognizable information, so some useful data may be missing. In addition, using only a single dataset may not show the superiority of the models and may blur the differences between models. All in all, there is still room for more development of this research in terms of data. If the research could find datasets with individual requirements and without garbled data, a model may be improved even better.

5. Conclusion

This experiment initially understands and unparsed the dataset's properties through EDA, subsequently processes the classification features, and divides the processed data into three TensorFlow datasets. Finally, the attributes were processed using text-vectorized NLP techniques to prepare for the subsequent convolutional neural network (CNN) model building. The authors built two neural network models with the following structural features. The authors use 30,000 as the dividing line and classify

DOI: 10.54254/2755-2721/6/20230910

those with fewer than 30,000 as lower salaries and those with more than 30,000 as a higher salaries. From the results, it could be seen that CNN has a lower prediction bias rate and higher prediction results stability. The significance of this research lies in the combination of convolutional neural networks and wage prediction, which may be useful for economics research and sociological research. One of this model's significant challenges is how to eliminate extreme values. There are two possibilities; one is that the model's logic will be abnormal in the face of severe cases. The other is that the model has not been trained enough, requiring further research and improvement.

References

[1] Khongchai, P., & Songmuang, P. (2016). Implement of salary prediction system to improve student motivation using data mining technique. In 2016 11th International Conference on Knowledge, Information and Creativity Support Systems, 1-6.

[2] Hu, G., Yang, Y., Yi, D., Kittler, J., Christmas, W., Li, S. Z., & Hospedales, T. (2015). When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In Proceedings of the IEEE international conference on computer vision workshops, 142-150.

[3] Bahar, P., Bieschke, T., & Ney, H. (2019). A comparative study on end-to-end speech to text translation. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop, 792-799.

[4] Li, Y., & Shahabi, C. (2018). A brief overview of machine learning methods for short-term traffic forecasting and future directions. Sigspatial Special, 10(1), 3-9.

[5] Shahbazi, Z., & Byun, Y. C. (2021). Improving the product recommendation system based-on customer interest for online shopping using deep reinforcement learning. Soft Computing and Machine Intelligence, 1(1), 31-35.

[6] Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. Social studies of science, 48(1), 25-56.

[7] Khongchai, P., & Songmuang, P. (2016). Random forest for salary prediction system to improve students' motivation. In 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems, 637-642.

[8] Dutta, S., Halder, A., & Dasgupta, K. (2018). Design of a novel prediction engine for predicting suitable salary for a job. In 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks, 275-279.

[9] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al. (2018). Recent advances in convolutional neural networks. Pattern recognition, 77, 354-377.

[10] Adzuna (2013) Job Salary Prediction. URL: https://www.kaggle.com/competitions/job-salary-prediction/overview/description