Multi Regressor Based User Rating Predictor for ImageCLEF Aware 2022

Aarthi Suresh Kumar[1], Anirudh A[1], Jeet Golecha M[1], Karthik Raja A[1], Bhuvana Jayaraman [1] and Mirnalinee T T [1]

[1]*Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India*

Abstract

Every one of the public nowadays have their presence in social media networks. The profile information of the social media account helps to understand nature of the user. Images, that are part of the profile information mostly characterizes the user and reveals much more about the user than the textual information. Such information extracted are used in many applications namely the employers, credit scoring, etc. This work has proposed Random forest regressor, Extra tree regressor and a dense neural network model for online user data scoring. Three submission using these models were made to the ImageClef Aware 2022 [1] task and has obtained 0.139 as Pearson Correlation Coefficient for testing.

Keywords

Multi Output Regressor, Random Forest, Extra Trees, Neural Network, User Rating

1. Introduction

According to a recent report, people are uploading data online at the rate of 1.8 billion images per day. This statistic adds up to around 657 billion photos [2] every year [3]. Most of these image files are in social networking platforms which can be accessed publicly. However, the owners of these digital images are often unaware of the fact that third parties could access them for a plethora of unethical reasons. Examples include the practice of obtaining information of potential employees by employers and using a user's online data to obtain an automatic credit score.

Existing methods rate the information a user uploads online. For instance, Bargh et. al. [4] explored the implications of public user data in the area of user privacy. The paper outlined how user data could be used to derive sensitive information about a user. It also introduced a feedback system from the data recipients to the data disseminators to curb the issue of leaking private information. Other similar approaches focus on inferring user characteristics and their practical utility is rather limited.

This paper aims to develop a more data-centric approach to solving the problem of online user

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

aarthi19003@cse.ssn.edu.in (A. S. Kumar); anirudh19015@cse.ssn.edu.in (A. A); jeetgolecha19043@cse.ssn.edu.in (J. G. M); karthikraja19048@cse.ssn.edu.in (K. R. A); bhuvanaj@ssn.edu.in (B. J.); mirnalineett@ssn.edu.in (M. T. T.)
https://www.ssn.edu.in/staff-members/dr-j-bhuvana/ (B. J.);
https://www.ssn.edu.in/staff-members/dr-t-t-mirnalinee/ (M. T. T.)
0000-0002-9328-6989 (B. J.); 0000-0001-6403-3520 (M. T. T.)

data scoring. It explores the efficacy of two classes of models, namely, regression models and deep learning models to predict the pertinence of a user's data to the following situations [5]:

1. Bank Loan
2. Accommodation
3. Applying to a job as a waitress/waiter
4. Applying to a job in IT

The regression based models include the Random Forest Regressor, Extra Trees Regressor and the Mutli-Output Regressor. A dense neural network was the deep learning model used for the user data feedback system. Of these models, the Random Forest Regressor performed the best, with a validation error of 0.49. The regression class on models performed better than the deep learning model.

2. Task and Dataset

ImageCLEF Aware 2022 [6]deals with developing model to predict the user ratings [7] for four distinct situations given the scores of different visual concepts. The models are expected to provide rankings for user test profiles that are as close as possible to the human rankings.

The dataset has 1000 user profiles, each having 100 photos that were annotated along with an appeal score via crowd sourcing for the real life scenarios listed earlier. Each profile is rated globally [8] for every situation using a Likert scale of 7 that ranges from strongly unappealing to strongly appealing.

Ground truth was created after averaging and normalizing the appeal score, which was then used for ranking the users in situation that are modeled. Prediction files, which contain visual concepts associated with each user, constitute the training data. Gt_files, which contain the the appeal score for each user for each real-life situation. A file with the score for each visual concept was provided as well. Incorporating the scores of each visual concept did not change the result.

3. Methodologies

3.1. Data Preprocessing

Prior to applying the machine learning and deep learning techniques, some preprocessing techniques were applied. The location of the visual concepts and the scores for each real-life situation were concatenated and made into a stacked matrix for each user. The cases involved not adding some of this features to reduce diverging, but all patterns gave similar results on training accuracy.

3.2. Regression Models

3.2.1. Random Forest Regressor

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training

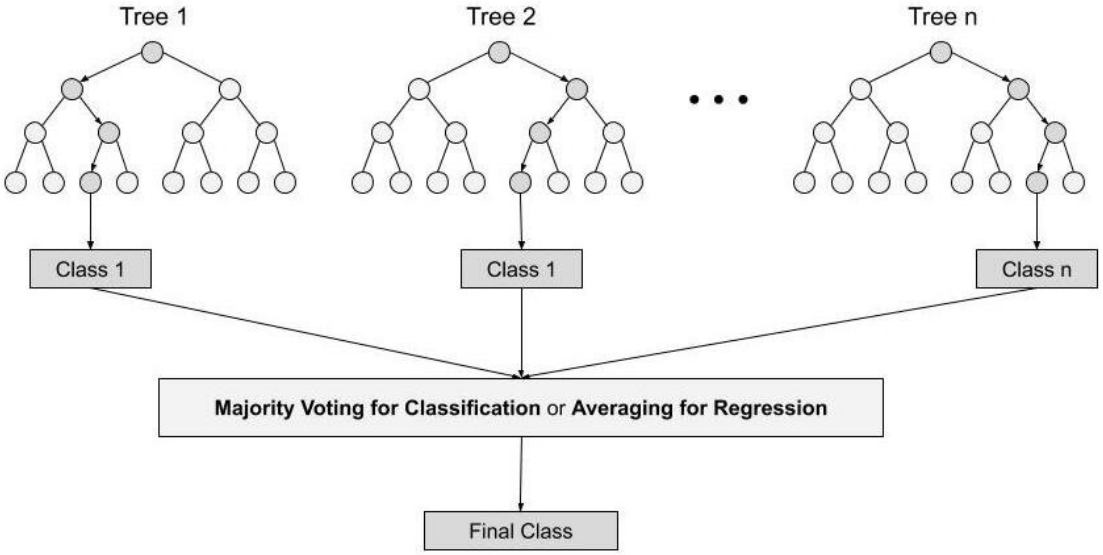

Figure 1: Random Forest

time. For classification tasks, The class chosen by majority of the trees will be the output class. A Random Forest [9] as an ensemble approach of decision trees, constructs as many trees in a random fashion as shown in in Figure 1. Each and every tree is constructed with different feature samples for splitting and at each node with different set of rows. Predictions are made by each tree which are combined / averaged together to give a single prediction for classification.

3.2.2. Extra Trees Regressor

Extra Trees is an ensemble machine learning algorithm that combines the predictions from several decision trees. It is a commonly used random forest algorithm. Although it uses a simpler approach where the invidual members are the decision trees, it can often yield similar or better results than the random forest algorithm.

Both the Random Forest Regressor and the Extra Trees Regressor are tree algorithms. The difference is that the Random Forest Regressor uses resampling and the Extra Trees Regressor uses original data to create the random forest of decision trees.

3.2.3. Multi Output Regressor

The two models discussed above output only a single real value. Hence, they are modified to produce multiple outputs using the Multi Output Regressor(MOR) function. The MOR function runs the Random Forest and Extra Tree Regresors 4 times to get the values for each of the real-life situations.

3.3. Neural network Model

A dense neural network was also explored. A dense neural network consists of dense layers. A dense layer is one that is connected to every neuron of its preceding layer. The dense neural network used for this task consists of 7 dense layers. The input is flattened into a 3000 point vector before passing it into the first layer of the dense neural network. The output of the dense neural network is a 4-point vector. The architecture of the deep learning model is shown in Figure 2.

```
Model: "sequential_13"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_66 (Dense)            (3000, 3000)              15000

 dense_67 (Dense)            (3000, 2048)              6146048

 dense_68 (Dense)            (3000, 1024)              2098176

 dense_69 (Dense)            (3000, 512)               524800

 dense_70 (Dense)            (3000, 128)               65664

 dense_71 (Dense)            (3000, 64)                8256

 dense_72 (Dense)            (3000, 4)                 260

=================================================================
Total params: 8,858,204
Trainable params: 8,858,204
Non-trainable params: 0
```

Figure 2: Deep Neural Learning Model

3.4. Training and Validation Set

In this section, we present a concise anaylsis of the two best models: Random Forest Regressor and Extra Trees Regressor. The training accuracy of the former was less than the latter. This can be attributed to the fact that the Random Forest Regressor uses the concept of bootstrap re-sampling, bringing in new data that can diverge from actual data for training.

On the other hand, the validation accuracy of the Random Forest Regressor was better than that of the Extra Trees regressor by approximately 0.01%.

4. Tested Models

The regression and deep learning models were tested. The regression class of models performed better than the dense neural network model. The 7 layer dense neural network had a validation accuracy of 0.15. It followed the same preprocessing techniques as the regression models. We suspect that lack of data can attributed to this poor accuracy. Hence, we had to alter our model to a much simpler neural network that can work with smaller amount of data.

5. Hardware used

A Google Colab notebook was used to train the model. A general purpose RAM size of 8GB was alloted with a 2.3GHz Intel Xenon CPU.

6. Code

The resources used by JBTTM for CLEF aware, including the research papers, exploratory data analysis, and code can be found here: https://github.com/AAnirudh07/CLEF-2022

7. Result

The Pearson Correlation Coefficient is a measure of linear correlation between two sets of data. The formula is provided Equation 1.

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \tag{1}$$

Team JBTTM had a maximum observed Pearson Correlation Coefficient of 0.139 for the two out of three submissions made. This correlation with Random Forest Regressors and the dense neural network.

8. Conclusion

An attempt was made to score the user profile using the visual contents availble in the social network account. Two regressor methods and one dense neural network were used for this

purpose . Of the 3 submissions that team JBTTM made, the regression model had the best accuracy based on the metrics proposed by CLEF. The accuracy of the 7 layer dense neural network model was inferior to the machine learning models and, no further improvements were made to it. In conclusion, machine learning models are more suitable for the task of user data rating than deep learning models. These results may also be attributed to the lack of training data.

References

[1] B. Ionescu, H. Müller, R. Peteri, J. Rückert, A. Ben Abacha, A. G. S. de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. Kozlovski, Y. D. Cid, V. Kovalev, L.-D. Ştefan, M. G. Constantin, M. Dogariu, A. Popescu, J. Deshayes-Chossart, H. Schindler, J. Chamberlain, A. Campello, A. Clark, Overview of the ImageCLEF 2022: Multimedia retrieval in medical, social media and nature applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy, 2022.

[2] V.-K. Nguyen, A. Popescu, J. Deshayes-Chossart, Unveiling real-life effects of online photo sharing, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2898–2908.

[3] T. Atlantic, How many photographs of you are out there in the world?, 2014. URL: https://www.theatlantic.com/technology/archive/2015/11/how-many-photographs-of-you-are-out-there-in-the-world/413389/.

[4] M. Bargh, P. Conradie, S. Choenni, R. Meijer, Privacy protection in data sharing: Towards feedback based solutions, volume 2014, 2014. doi:10.1145/2691195.2691279.

[5] P. Li, Z. Wang, Z. Ren, L. Bing, W. Lam, Neural rating regression with abstractive tips generation for recommendation, in: Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, 2017, pp. 345–354.

[6] A. Popescu, J. Deshayes-Chossart, H. Schindler, B. Ionescu, Overview of the imageclef 2022 aware task, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy, 2022.

[7] P.-Y. Hsu, Y.-H. Shen, X.-A. Xie, Predicting movies user ratings with imdb attributes, in: International Conference on Rough Sets and Knowledge Technology, Springer, 2014, pp. 444–453.

[8] N. Armstrong, K. Yoon, Movie rating prediction, Technical Report, Citeseer, 1995.

[9] A. Ajesh, J. Nair, P. Jijin, A random forest approach for rating-based recommender system, in: 2016 International conference on advances in computing, communications and informatics (ICACCI), IEEE, 2016, pp. 1293–1297.