aware

by Bhuvana J

Submission date: 26-May-2022 10:02PM (UTC+0530)
Submission ID: 1844718284
File name: clefaware_main.pdf (183.65K)
Word count: 1207
Character count: 6267

Multi Regressor Based User Rating Predictor (ImageCLEF Aware)

Aarthi Suresh Kumar[1], Anirudh A[2], Jeet Golecha M[3], Karthik Raja A[4], Bhuvana Jayaraman[5], and Mirnalinee T T[6]

SSN College of Engineering, India
`(aarthi19003,anirudh19015,jeetgolecha19043,karthikraja19048,bhuvanaj,`
`mirnalineett)@cse.ssn.edu.in`

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 are part of the profile information t
Analysis of images is a vitally important task in any field of domain. Images constitute a large part of the content shared on social networks. Their disclosure is often related to a particular context and users are often unaware of the fact that, depending on their privacy status, images can be accessible to third parties and be used for purposes which were initially unforeseen.

Keywords: Multi Output Regressor, Random Forest, Neural Network, Scikit Tree, User Rating

1 Introduction

According to a recent report by Mary Meeker's annual Internet Trends, people upload an average of 1.8 billion digital images every day. This statistic adds up to around 657 billion photos every year [1]. Most of these image files are concentrated in social networking platforms and 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 have been introduced to rate the information a user uploads online. For instance, Bargh et. al. [2] 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 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:

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

ImageCLEFAware 2022 deals with developing models to predict the user ratings for 5 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.

A data set of 1000 user profiles with 100 photos per profile was created and annotated with an appeal score for a series of real-life situations via crowdsourcing. A global rating was provided for each profile in each situation modeled using a 7-points Likert scale ranging from strongly unappealing to strongly appealing. An averaged and normalized appeal score was used to create a ground truth composed of ranked users in each modeled situation. 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 accuracy.

3.2 Techniques and Models used

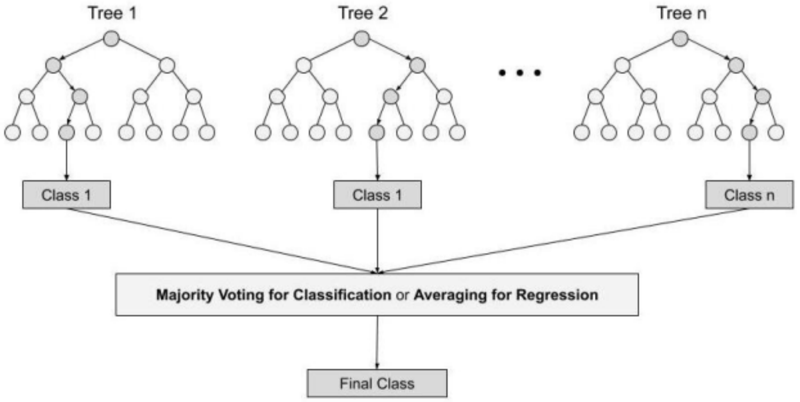

Tree 1
Tree 2
Tree n
Class 1
Class 1
Class n
Majority Voting for Classification or Averaging for Regression
Final Class

Fig. 1. Random Forest

Random Forest Regressor Random forest is an ensemble of decision trees. This is to say that many trees, constructed in a certain “random” way form a Random Forest.

Each tree is created from a different sample of rows and at each node, a different sample of features is selected for splitting. Each of the trees makes its own individual prediction. These predictions are then averaged to produce a single result.

Extra Trees Regressor Extra Trees is a machine learning approach that combines the predictions of multiple decision trees into a single forecast.

It is a commonly used random forest algorithm. Although it uses a simpler approach to create the decision trees used as members of the ensemble, it can often yield similar or better results than the random forest algorithm.

Both the techniques discussed above are tree algorithms, with random forest regressor using resampling and the latter using only the original data to create the random forest.

MultiOutput Regressor The above linear out put models are run more times to get multiple values to match with the target size, in this case to match with the 4 visual concepts.

3.3 Deep Learning model

3.4 Training and Validation set

As discussed in the above section, the random forest regressor brings in a concept of Bootstrap resamplaing bringing in new data that can diverge from actual data for training, and hence decreasing its training accuracy compared to the Extra trees regressor. Whereas the validation accuracy of the random forest regressor faired better than Extra trees by approximately 0.01

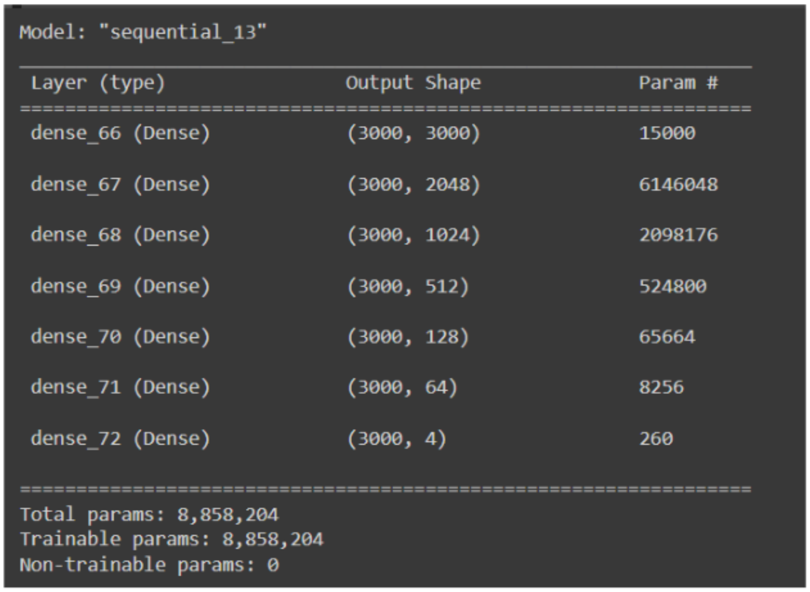
4 Tested Models

We tested on the above stated training models, and additionally a DL model with the architecture as shown below. But the error value was too high, hence was discarded at the end. 3.2

This followed the same pre-processing techniques from the proposed model. The average validation accuracy measured was only 0.15 .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

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.



```
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
```

Fig. 2. DL model

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 proposed models had an accuracy of 0.137 pearson correlation value , after submitting to the aicrowd submissions.

8 Conclusion

References

1. Atlantic, T.: How many photographs of you are out there in the world? (2014), `https://www.theatlantic.com/technology/archive/2015/11/how-many-photographs-of-you-are-out-there-in-the-world/413389/`

2. Bargh, M., Conradie, P., Choenni, S., Meijer, R.: Privacy protection in data sharing: Towards feedback based solutions. vol. 2014 (01 2014). https://doi.org/10.1145/2691195.2691279

aware

ORIGINALITY REPORT

22%	20%	21%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	WWW.IMAGECLEF.ORG Internet Source	12%
2	neptune.ai Internet Source	6%
3	Hemal Patel, Neha Soni. "Machine Learning Based Approach For Prediction Of Suicide Related Activity", 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021 Publication	2%
4	Aakansha Dijendra, Jignesh Sisodia. "Telecomm Churn Prediction Using Fundamental Classifiers To Identify Cumulative Probability", 2021 International Conference on Communication information and Computing Technology (ICCICT), 2021 Publication	2%
5	www.theatlantic.com Internet Source	1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		