SSN CSE at ImageCLEFaware 2022: Contextual Job Search Feedback Score based on Photographic Profile using a Random Forest Regression Technique

Aarthi Nunna[1], Aravind Kannan Rathinasapabathi[2], Chirag Bheemaiah P K[3] and Kavitha Srinivasan[4],

[1,2,4] *Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam – 603110, India.*

[3] *Department of IT, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam – 603110, India.*

Abstract

Social networks have become increasingly popular with millions of users where the digital presence has become more crucial to a person's character judgement. Employers tend to screen through their candidates' profiles on social media to understand their personalities and infer the knowledge about the candidate's eligibility for a specific job. To address this issue, the ImageCLEF forum is conducting a task to quantify the effect of the photographic profile from 2021 onwards and we participated this year. Therefore, an algorithm was developed to score the images of a user and provide them with comprehensive feedback on the consequences of the images on their selected professions. The approach used to develop the algorithm uses Random Forest Regression which resulted in a Pearson Correlation Coefficient of 0.544.

Keywords 1

Supervised learning, Random Forest Regression Algorithm, Social photographic profile-based score, Pearson Correlation Coefficient

1. Introduction

In today's world, the digital presence of humans has become more pivotal than their physical presence. With the ease of internet access, everybody is more digitally conscious than social. As a result, social networking is undoubtedly an integral part of life. Every day, millions of users upload content such as images, posts, and stories on platforms like Instagram, Twitter, Facebook, etc. The actions and posts on social media can have real-time effects on the physical world, it can even affect a person's ability to acquire a job. So, it has become crucial to learn the effects of visual media uploaded on various social platforms as explained by Van-Khoa Nguyen et al. [1]. If users are digitally responsible and disciplined when uploading various visual media, it benefits the society as a whole because their digital presence wouldn't have any adverse effects on their career and its growth.

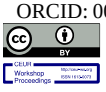
The ImageCLEFaware 2022 is the second edition of the aware task conducted by the CLEF Initiative. The task asked participants to provide a global rating of each profile in each situation using a Likert Scale. In the edition held in 2021 [2], 500 user profiles were provided in the dataset opposed to 1000 user profiles in this edition. This forum has taken this socially significant issue into their hands again, for the second time, and put together a dataset of various users along with the pictures they posted in an anonymized format for us to analyze the real-world effect on four selected situations, namely bank loan, accommodation, jobs as a waitress/waiter, and jobs in IT. The final objective of the task would beto integrate the model with a mobile application for users to obtain their feedback efficiently and easily.

Our team strived to develop an algorithm that provides feedback to the users that resembles feedback given by humans. The dataset used is a subset of the YFCC100M [3] dataset. It comprises various user profiles. Each profile constitutes a maximum of hundred images. A thousand user profiles were used to train our model. The objects present in the images are initially detected by a Faster-RCNN model, resulting in a confidence score for each object detected. Thus, our models have taken as input a JSON file that comprises the object detected along with its confidence score and its bounding box. We experimented with

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

EMAIL: aarthi19002@cse.ssn.edu.in (A. 1); aravindkannan19022@cse.ssn.edu.in (A. 2); chiragbheemaiahpk19025@it.ssn.edu.in (A. 3); kavithas@ssn.edu.in (A. 4);

ORCID: 0000-0002-5944-5712(A. 1); 0000-0003-3783-372X (A. 2); 0000-0003-3315-2994 (A. 3); 0000-0003-3439-2383 (A. 4);



the algorithms explained in Section 3 and 4 and found Random Forest Regression to be the best performing algorithm.

The following sections of the paper are: Section 2 describes the dataset provided by ImageCLEF, the various models that were used to obtain the required outputs in Section 3, the comparison between the models and the inferences obtained are discussed in Section 4. Finally, we have summarized the findings in the conclusion and future work Section.

2. Dataset

The dataset was given with a split of three categories, namely training, validation, and testing data by the ImageCLEF forum. The testing data was used to obtain the final output file which was submitted for evaluation. Table 1 provides a comprehensive understanding of the dataset and additional information about the same.

Table 1
Dataset description

File Provided	Data Provided	Observed
Class_scores.json	Each visual concept detected has a score depicting its influence on the four professions	Scores for visual concepts 80 and 215 are unavailable.
Prediction_train.json, Prediction_val.json, Prediction_test.json	The three folders are pertaining to the input files for the three respective dataset categories. The train and validation input files are used to train our various models. The test input file was used to obtain the ground truth output file to be submitted for evaluation.	The folder contains the users' photographic profiles that comprises of each user, their respective images and the objects detected.
Gt_train.json, Gt_val.json	The final output ranks of each user's profile with respect to the four professions chosen.	The file comprises of each user and four values that determine how the social profile of the user would affect his/her career choice.

3. System Design

The system design of the developed model is visually represented in Figure 1. As observed, there are multiple components that serve as the input for the model which are obtained from the dataset explained in Table 1

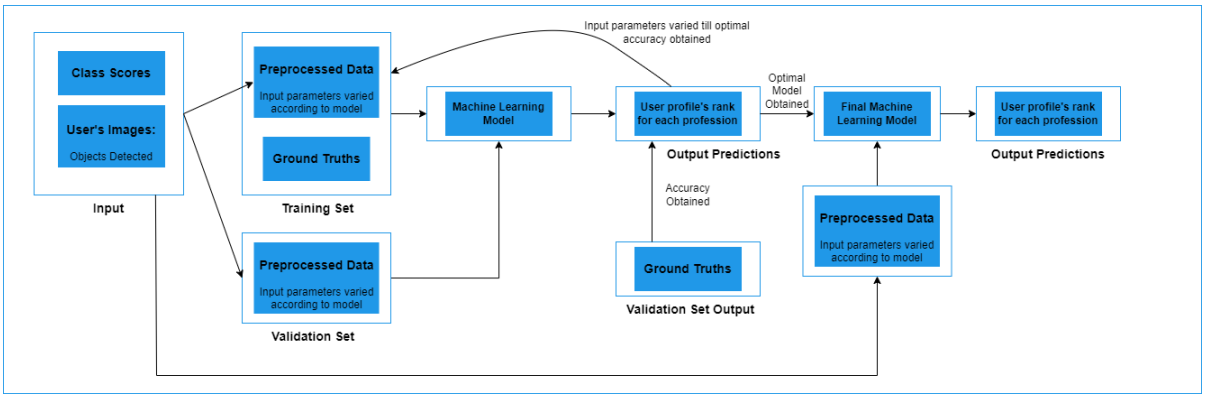


Figure 1: System Design

Figure 1 depicts a high-level abstraction of the proposed system design. The system takes the dataset as input, pre-processes it and then trains a model on it. The model was then evaluated based on the ground truth values and a set of performance metrics such as the Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). On obtaining these values, the model's performance was improved by varying its input and the model's parameters.

The input data consists of the class scores which contains the score that depict the influence of an object on each profession that may be of either positive or negative value. Secondly, the user profiles along with their images and the objects within them are fed as input. However, the data was pre- processed to fit the various machine learning models. The JSON files are read into data frames for easy processing. The final inputs to the machine learning model are varied and their performance as inputs are observed and tabulated as explained in the following paragraphs.

The output JSON file contains the model's predictions of the scores that should be associated to each user's photographic profile on considering the user's images and the objects within them. The expected output given in the ground truth files are used to calculate the accuracy of the trained model using which changes to the model are performed and analyzed.

3.1 Random Forest Regression

The bagging model is used by the Random Forest Algorithm [4] as visualized in Figure 2. This means that subsets of the dataset are used to train various decision trees and the final output is taken through the idea of majority voting. The concept of majority voting is also known as aggregation. Thus, through the methods of replacement and bootstrap aggregation, random forest was observed to be the best algorithm amongst its competitors XGBoost [5] and ANN.

Mentioned below are the different versions of the Random Forest model which are unique due to the input parameters that are fed to the models. The models' accuracies are measured using the error values obtained.

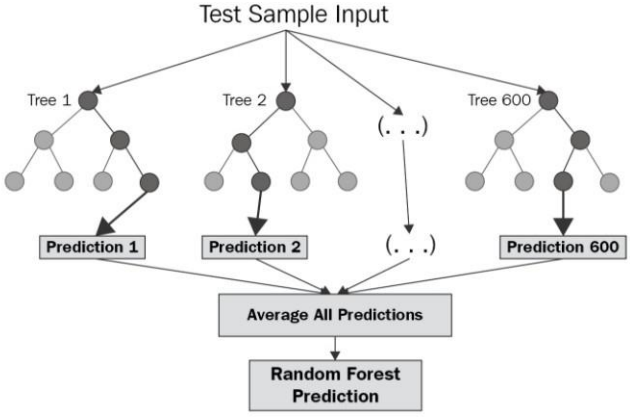


Figure 2: Random Forest Regression

Figure 2: Random Forest Regression explains the working of Random Forest Regression which utilizes the concept of bagging.As depicted in the figure, subsets of the input data are used to train different decision trees, whose predictions are averaged to obtain the final prediction.

3.1.1. Model 1

In this approach, the input was defined as the 'average confidence score', along with 'average impact scores for each of the classes' for a given user. A random forest regressor was defined with the 'number of estimators' parameter ranging from 10 to 1000. This regressor was fit on 80% of the training data with the remaining being reserved for testing. A loss function of 'Mean Squared Error'was used to grade the performance of the result. The same plan of action was adopted for validation data. All the models were measured parametrically using Pearson Correlation Coefficient [7]. Figure 3, illustrates the model's performance on the training data.

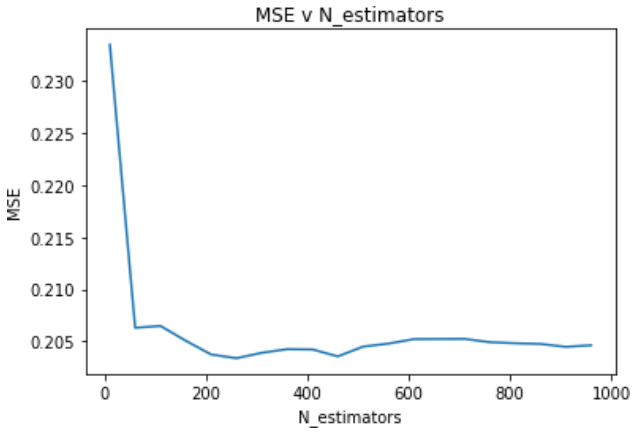


Figure 3: Training Data

The number of estimators for the regressor was decided to be 650. This regressor was then applied to the testing data which gave the following results as given in Table 2. The Pearson correlation coefficient value was calculated to be 0.288.

Table 2
Model 1 Metrics

Metrics	Training Dataset	Validation Dataset
Mean Absolute Error (MAE)	0.36241	0.37192
Mean Squared Error (MSE)	0.20515	0.23345
Root Mean Squared Error (RMSE)	0.45294	0.48317

3.1.2. Model 2

In addition to the confidence score and the average impact scores, the objects detected were represented using a matrix whose index represents the object and the value at the index represents the count of the object was also provided as the input.

To find the optimal number of estimators, a similar approach as the previous method was adopted. Figure 4 illustrates the model's performance over the training and validation phases.

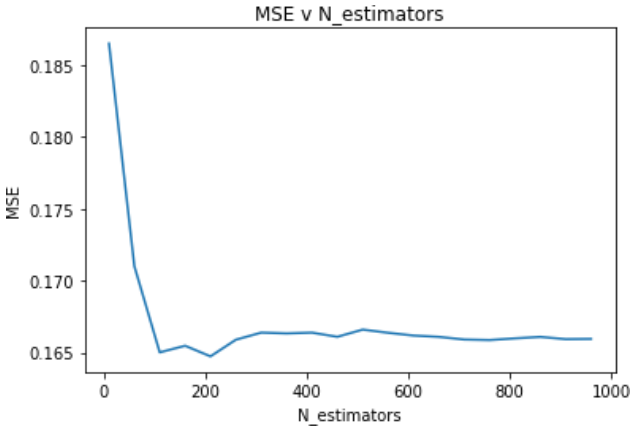


Figure 4: Training Data

The number of estimators for the regressor was decided to be 650. This regressor was then applied to the testing data which gave the following results as given in Table 3. The Pearson correlation coefficient value was calculated to be 0.544.

Table 3
Model 2 Metrics

Metrics	Training Dataset	Validation Dataset
Mean Absolute Error (MAE)	0.371921	0.328819
Mean Squared Error (MSE)	0.176763	0.166127
Root Mean Squared Error (RMSE)	0.407587	0.420433

3.1.3. Model 3

Having ascertained that Random Forest Regression is the best fit model for the problem, parameter tuning was performed to optimize the model. The following parameters were considered and set with different options using grid search [8] as follows:

1. 'bootstrap': [True, False],
2. 'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
3. 'max_features': ['auto', 'sqrt'],
4. 'min_samples_leaf': [1, 2, 4],
5. 'min_samples_split': [2, 5, 10],
6. 'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]

The options were then validated using 3-fold cross validation approach to determine the leading model. The optimal parameters retrieved were: 'bootstrap': True, 'max_depth': 50, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 2000. The Pearson correlation coefficient value was calculated to be 0.542 and Table 4 tabulates the metrics obtained for the same.

Table 4
Model 3 Metrics

Metrics	Training Dataset	Validation Dataset
Mean Absolute Error (MAE)	0.333413	0.339431
Mean Squared Error (MSE)	0.185843	0.193637
Root Mean Squared Error (RMSE)	0.431096	0.440042

3.1.4. Model 4

Like the above versions, an additional feature was added in order to improve the model's accuracy. In the dataset used for training, the objects in various images are identified along with their confidence score and the coordinates of their bounding box. With the knowledge of the coordinates, the area of the bounding box was calculated using simple geometry as they always form a rectangular shape. The area of the bounding was used as it would account for the importance weight of the object in the image along with its confidence score.

Figure 5 illustrates the model's performance of training data. The Pearson correlation coefficient values was resulted as 0.519 and Table 5 tabulates the metrics obtained for the same.

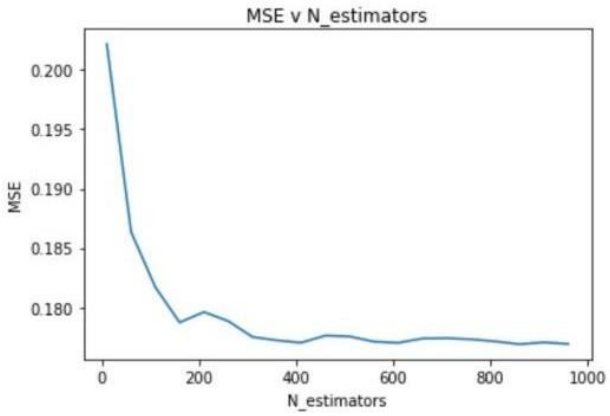


Figure 5: Training Data

Table 5
Model 4 Metrics

Metrics	Training Dataset	Validation Dataset
Mean Absolute Error (MAE)	0.343690	0.337131
Mean Squared Error (MSE)	0.198958	0.189130
Root Mean Squared Error (RMSE)	0.446047	0.420433

4. Implementation and Results

In this section, the aforementioned machine learning models are trained, and the corresponding results are compared and analyzed using performance metrics, namely the Mean Squared Error, Mean Absolute Error and Root Mean Squared Error.

4.1 System Specification

The hardware and software specification required for the implementation the machine learning models includes, Intel i7 processor with NVIDIA MX100 2GB graphics card, 8GB RAM, 1TB disk space, Windows 11 OS, Jupyter Notebook, Python 3.7 packages with required libraries like Sklearn, Tensorflow, Numpy, Pandas, etc.

4.2 Results of Machine Learning Models

Since the problem warrants a multivariate regression approach with multiple target variables, XGBoost, Artificial Neural Network, Random Forest Regression models were considered. Thorough experiments were conducted, and the results are depicted below as follows in Table 6.

Table 6
Model accuracy comparisons

ML Models	Training Dataset (MSE)	Validation Dataset (MSE)
Random Forest Regression	0.1661	0.1767
XGBoost	0.1870	0.1966
Artificial Neural Network	0.1761	0.2160

From the above data it can be inferred that the Random Forest Regression model is well suited for the problem statement and hence was chosen as the baseline model for the given task.

The accuracies of the regression models were evaluated on the test set based on the Mean Squared Error (MSE) metric. In the first run, the 'Model 1' was trained by setting the number of estimator parameter as 650 after iterating over values in the range of [10,1000]. This model took as input the 'average confidence score', an 'average impact scores for each of the classes.' In Run 2, 'Model 2' took in similar inputs but additionally accounts for all the images detected in a user's profile. In Run 3, 'Model 2's' hyperparameters are altered to achieve better performance. Besides the inputs which were given to 'Model 2', area of the bounding boxes of the objects detected was calculated and was used as an input.

Table 7
Brief description about each run

Run Number	Approach	MSE-Training	MSE-Validation	Pearson's Correlation Coefficient
1	Model 1	0.1661	0.1767	0.288
2	Model 2	0.1870	0.1966	0.544
3	Model 3	0.1761	0.2160	0.542
4	Model 4	0.1989	0.1891	0.519

Inferring from the results tabulated in Table 7Table , it is evident that the inclusion of the objects detectedper user was a key factor in improving the prediction performance metrics. Furthermore, it was observedthat hyperparameter tuning did not affect the performance of the model substantially.

5. Conclusion and Future Works

The paper aims to devise a solution for the ImageCLEFaware 2022 Task. The task aims to provide a solution to generate contextual feedback scores for a user's social profile and its influence in their job prospects. The paper describes the various models that were implemented, and their performances have

been compared. It was observed that the Random Forest Regression model performed far better than the other models such as XGBoost and ANN. On inferring the same, the inputs given to the Random Forest Regression Model were tweaked and different Random Forest models were developed. These models were compared based on Mean Absolute Error, Mean Squared Error and Root Mean Squared Error. Model 2 has fared better than the other models as per the Pearson Correlation Coefficient value. This could be a direct consequence of the consideration of the objects detected being fed as an input parameter. Hence, this can be worked on further to enhance its correlation with the required output.

In the future, the dataset can be made more diverse to cover all edge cases and thus aid in developing a more robust algorithm. Other ensemble learning algorithms can be experimented to arrive at meticulous conclusions that will help improve the model's performance. Thus, fine tuning the hyper parameters of the algorithms such as epochs, learning parameters, cross-validation etc. can increase the efficiency and improve the results that are currently obtained.

6. Acknowledgements

We express our deep gratitude towards CLEF Initiative labs for coming up with the problem statement for us to work on and giving us timely assistance. It was due to ImageCLEF 2022 Aware [9,10] that we learnt a lot during the contest, so we're forever indebted to them. We appreciate AI4Media to support this task. We are grateful to the YDSYO Team for sharing with us the anonymized dataset. We would also like to take this opportunity to thank our college, Sri Sivasubramaniya Nadar College of Engineering, Department of Computer Science and Engineering for motivating us with the opportunityto work on this task.

7. References

[1] Van-Khoa Nguyen, Adrian Popescu, and Jérôme Deshayes-Chossart. "Unveiling Real-Life Effects of Online Photo Sharing." IEEE WACV 2022.

[2] Bogdan Ionescu, Henning Müller, Renaud Péteri, Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, Serge Kozlovski, Vitali Liauchuk, Yashin Dicente, Vassili Kovalev, Obioma Pelka, Alba García Seco de Herrera, Janadhip Jacutprakart, Christoph M. Friedrich, Raul Berari, Andrei Tauteanu, Dimitri Fichou, Paul Brie, Mihai Dogariu, Liviu Daniel Ştefan, Mihai Gabriel Constantin, Jon Chamberlain, Antonio Campello, Adrian Clark, Thomas A. Oliver, Hassan Moustahfid, Adrian Popescu, Jérôme Deshayes-Chossart, Overview of the ImageCLEF 2021: Multimedia Retrieval in Medical, Nature, Internet and Social Media Applications, in Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), Bucharest, Romania, Springer Lecture Notes in Computer Science LNCS, September 21-24, 2021.

[3] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, Li-Jia Li. "YFCC100M: The New Data in Multimedia Research" (2016) URL: https://doi.org/10.48550/arXiv.1503.01817

[4] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). URL: https://doi.org/10.1023/A:1010933404324

[5] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, (2016): 785-794.

[6] Afroz Chakure, Random Forest Regression, Medium Article. URL: https://miro.medium.com/max/1400/0*f_qQPFpdofWGLQqc.png

[7] Kirch W, Pearson's Correlation Coefficient. Encyclopedia of Public Health. Springer, Dordrecht (2008). URL: https://doi.org/10.1007/978-1-4020-5614-7_2569

[8] Petro Liashchynskyi, Pavlo Liashchynskyi. "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS" (2019) URL: https://doi.org/10.48550/arXiv.1912.06059

[9] Bogdan Ionescu, Henning Müller, Renaud Péteri, Johannes Rückert, Asma Ben Abacha, Alba García Seco de Herrera, Christoph M. Friedrich, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Serge Kozlovski, Yashin Dicente Cid, Vassili Kovalev, Liviu-Daniel Ștefan, Mihai Gabriel Constantin, Mihai Dogariu, Adrian Popescu, Jérôme Deshayes-Chossart, Hugo Schindler, Jon Chamberlain, Antonio Campello, Adrian 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), Springer Lecture Notes in Computer Science LNCS, Bologna, Italy, September 5-8, 2022.

[10] Adrian Popescu and Jérôme Deshayes-Chossart and Hugo Schindler and Bogdan 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), Lecture Notes in Computer Science LNCS, Springer, Bologna, Italy, September 5-8, 2022.