# Multi-model Estimators and Ensemble-based Regressors for Predicting Video Memorability

R Gokul Prakash, Jayaraman Bhuvana, Eeswara Anvesh Chodisetty, Arjun Mukesh and T T Mirnalinee

*Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India*

### Abstract

The need to organize prevalent multimedia in our day-to-day activities, presents us various aspects of video importance to be taken into consideration, like video aesthetics and interestingness. Memorability based organization of media is highly effective in cases where media is required to have a long positive retainability in one's memory, like in the field of advertising or educational content creation. Keeping our target group in mind, we have decided to pursue an ensemble-based approach for our model for the task of predicting media memorability in aide to the Benchmarking Initiative for Multimedia Evaluation's list of tasks for the MediaEval 2022 Workshop. Pre-defined ensemble models are versatile enough to incorporate other pre-defined regressors and ensemble models as base estimators and build on the knowledge accumulated by them. Video-level features like the C3D were chosen, and the regressors and ensemblers were trained on these features with parameter optimization.

## 1. Introduction

This paper presents our work on Subtask 1 of the Predicting Video Memorability task [1] in MediaEval 2022. The purpose of this task is to design a system to predict memorability scores for a given video.

Over the previous editions of this task, different approaches have been used. This work concentrates on using Multi-model estimators and ensemble-based regressors to achieve the goal of this task. Proposed methods are based on Ensemble techniques that use multiple regressor models as base estimators and combine their results to achieve a model with higher accuracy. Reason for choosing this method is for it's robustness in reducing the spread of the predictions and for it's ability to attain better performance than any single contributing model.

## 2. Related Work

The task of assigning a score for video memorability can be formulated as a regression problem. In recent years, there have been numerous advances in the study of regression and classification on video in academic literature. The value of combining features from different modalities has been thoroughly demonstrated in numerous earlier works [2][3]. The top-performing models for the Predicting Media Memorability challenge's 2019 and 2020 rounds used ensemble models.

By building on top of the previous editions, we approached this task as an optimisation challenge. Seminal papers on Ensemble learning [4][5] helped us adapt the method to the given dataset without overfitting while producing good scores.

The modalities and traits that are best predictive of memorability remain unknown. In light of this, we decided to use Ensembler-based methods in our quest to achieve a breakthrough.

## 3. Proposed Approach

In this proposed work to measure the video memorability, we opted to use 3D Convolution (C3D) features [6], features that were extracted from the video entirely, rather than features extracted from few particular frames in the video. The pre-extracted C3D features provided by the organizers are fed to our models. C3D, was one of the first ways to learning generic representation from videos. Its homogeneous architecture is built of tiny 3x3x3 convolution kernels. It produces a 4096-dimension feature vector extracted from a video clip after being trained on a generic action recognition dataset. Seven models that are chosen here are, Linear Regression [7], Logistic regression, Decision Tree, Random Forest [8], Bayesian Regression [9], Support Vector Regression, K-Neighbours regression [10] were used. Six ensemble methods employed are Gradient Boosting Regressor [11], AdaBoost Regressor, Voting Regressor [12], XGBoost Regressor [13] and Stacking Regressor Ensemblers [14]. The workflow of the proposed video memorability is shown in figure 1.
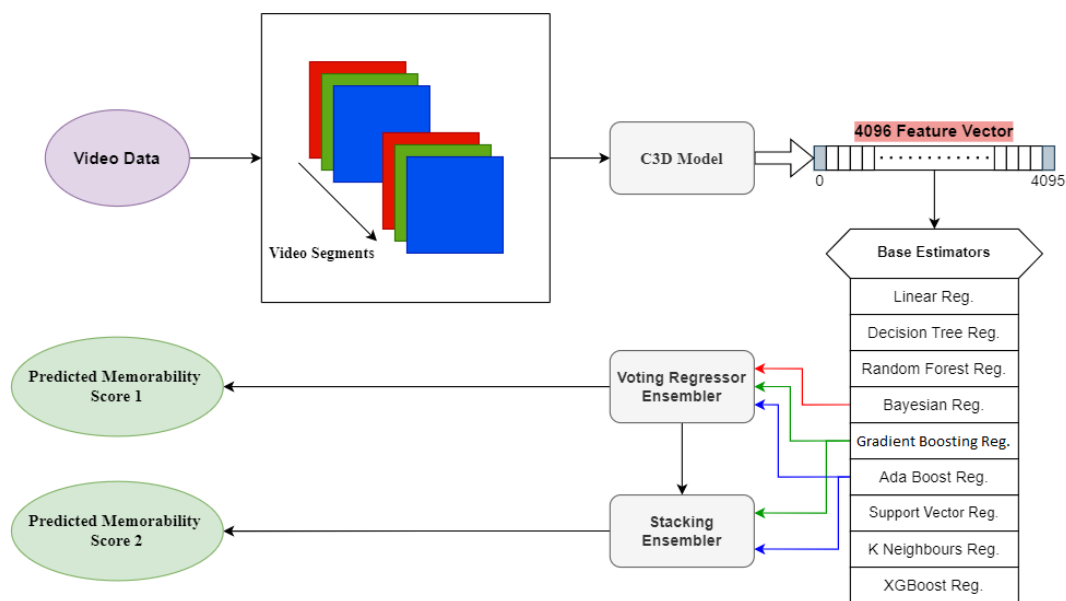


**Figure 1:** Prediction workflow used for the predicting the memorability score of the given video data using Voting and Stacking Ensemblers.

The models were trained using an 80-20 training-validation split based on Video ID. To test the effectiveness of the various models, we ran each model individually on the dataset and carefully handpicked those that gave the best metrics. The metrics were based on Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Spearman rank correlation coefficient. Ensemble learning tends to perform better when there is a diversity among the models used and this was an additional contribution to our choice of models.

Previous work[15] has shown that voting method and stacking method produce high Spearman correlation coefficient for similar data, hence they were among our initial consideration. Further down the task, we were able to test and implement four other models which also

performed on-par with the Stacking Ensembler and Voting Ensembler.

## 4. Implementation and Experiments

RMSE and Spearman's $\rho$ were the deciding factors based on which we chose the methods after rigorous experimentation. Internal tests were done on the 80/20 training/validation split.

### 4.1. Choice of Models

Models that are effective in predicting continuous output variables, memorability scores in this case, were chosen.

The models that achieved low RMSE scores and high Spearman's rank correlation coefficient ($\rho$) on the training set were ascertained and are added to a regressor candidate pool which was initially kept empty. Ensemble models, such as the Gradient Boosting Regressor, AdaBoost Regressor, and XGBoost Regressor, do not require additional base estimators and can be trained independently. Therefore, they were also added to the regressor candidate pool. On the other hand, Voting Regressor and Stacking Regressor permit the use of pre-trained base estimators. The regressors in the candidate pool with the best performance, which can be observed from Table 1, were chosen as the base estimators for Voting and Stacking Regressors. Since Voting Regressor allows 3 base estimators, 4 best models based on performance from the pool were chosen and 4 instances of Voting Regressor were made with different permutations of the 3 base estimators. The best performance was observed in the model with AdaBoost, Gradient Boosting, and Bayes Regression as best estimators. Similarly, 4 instances of Stacking Regressor models were made, whose 3 base estimators were permuted with 4 of the best performing models in the pool which now includes Voting Regressor. The best performance was observed in the model with Voting Regressor, AdaBoost Regressor, and Gradient Boosting Regressor as base estimators.

**Table 1**
Performance of Ensemblers on various metrics over training set

| Ensembler | RMSE | MSE | MAE | Spearman's $\rho$ |
|---|---|---|---|---|
| AdaBoost Regressor | 0.096 | 0.009 | 0.077 | 0.446 |
| XGBoost Regressor | 0.096 | 0.009 | 0.077 | 0.447 |
| Gradient Boosting Regressor | 0.094 | 0.009 | 0.075 | 0.480 |
| **Voting Regressor** | 0.093 | 0.008 | 0.074 | 0.507 |
| **Stacking Regressor** | 0.091 | 0.008 | 0.073 | 0.520 |

### 4.2. Choice of Methods

Six ensemble methods were chosen initially of which two made the final cut. One of the base estimators in the proposed architecture for the final Ensemblers (Voting and Stacking) is the Gradient Boosting Regressor, which serves as an Ensembler itself. *max_depth* is a parameter that was identified as an interesting value to tweak [16], that can improve the accuracy of prediction. It is used to set the maximum depth of the individual regression estimators that limits the number of nodes in the tree. The best value depends on the interaction of the input variables. It was observed that a lower value of *max_depth* = 1 achieved a better performance for the given training and validation set, rather than the default value of 3, which resulted in

our model overfitting the data. The other parameters were set to their default values. Both of the final Ensemblers use AdaBoost Regressor as a base estimator, in which the *n_estimators* parameter, used to set the maximum number of estimators at which boosting is terminated, was set to 100 instead of the default value of 50 which was observed to be insufficient due to the dataset being large. The other parameters were set to their default values.

## 5. Results and Analysis

The performance of the proposed architecture was evaluated using the metrics namely Pearson correlation coefficient, Spearman's rank correlation coefficient, and MSE.

From the results observed on the training set reported in table 1, have given a Spearman coefficient of around 0.52. This indicates a positive and above average correlation between the two metrics. Out of all the regressors in the candidate pool, Voting Regressor and Stacking Regressor performed the best with Spearman's coefficient of 0.507 and 0.520 respectively. The MSE values of both ensemblers were 0.008. The other regressors gave around 0.45 Spearman's coefficient but the best two were picked namely, Stacking Ensembler and Voting Ensembler, and submitted for evaluation. The training set's result of the final ensemblers are highlighted in table 1.

The submitted runs of the Stacking and Voting ensemblers were evaluated and the results are tabulated in table 2. The results using the training data and testing data show a low variance in the performance implying the model is not over-fitted on the training data. The Spearman's coefficient for the Stacking Ensembler and Voting Ensembler are 0.525 and 0.513 respectively which means an above average correlation between our predicted scores and the ground truth values. The Pearson coefficient also indicates the same. The most notable metric from submitted runs is the MSE of 0.008 for both Ensemblers.

In summary, the results indicate that Stacking regressor was able to perform better than the others. Although it was found not to be very different from Voting Regressor, Stacking Regressor achieved the best correlation and MSE scores.

**Table 2**
Submission performances

| Ensembler | Spearman $\rho$ | Pearsons | MSE |
|---|---|---|---|
| Stacking Ensembler | 0.525 | 0.523 | 0.008 |
| Voting Ensembler | 0.513 | 0.513 | 0.008 |

## 6. Discussion and Outlook

In this competition, we've built a strong upper layer based on the foundation and precedents established by previous work. Our key takeaway lies in aiding to establish ensemble learning as a great way to approach the challenge of predicting media memorability.

We emphasise that ensemble models from multiple methods produces the best results. Future work would ideally experiment further with the parameters, different base estimators, and different ensemble techniques other than the ones mentioned in our approach.

# References

[1] L. Sweeney, M. G. Constantin, C.-H. Demarty, C. Fosco, A. García Seco de Herrera, S. Halder, G. Healy, B. Ionescu, A. Matran-Fernandez, A. F. Smeaton, M. Sultana, Overview of the MediaEval 2022 predicting video memorability task, in: MediaEval Multimedia Benchmark Workshop Working Notes, 2023.

[2] L. Sweeney, G. Healy, A. F. Smeaton, Predicting media memorability: comparing visual, textual and auditory features, arXiv preprint arXiv:2112.07969 (2021).

[3] D. Xu, X. Wu, G. Sun, Media memorability prediction based on machine learning., in: MediaEval, volume 2882, 2020.

[4] T. G. Dietterich, et al., Ensemble learning, The handbook of brain theory and neural networks 2 (2002) 110–125.

[5] X. Dong, Z. Yu, W. Cao, Y. Shi, Q. Ma, A survey on ensemble learning, Frontiers of Computer Science 14 (2020) 241–258.

[6] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[7] S. Weisberg, Applied linear regression, volume 528, John Wiley & Sons, 2005.

[8] X. Li, et al., Using" random forest" for classification and regression., Chinese Journal of Applied Entomology 50 (2013) 1190–1197.

[9] C. M. Bishop, M. E. Tipping, Bayesian regression and classification, Nato Science Series sub Series III Computer And Systems Sciences 190 (2003) 267–288.

[10] H. Papadopoulos, V. Vovk, A. Gammerman, Regression conformal prediction with nearest neighbours, Journal of Artificial Intelligence Research 40 (2011) 815–840.

[11] P. Prettenhofer, G. Louppe, Gradient boosted regression trees in scikit-learn, in: PyData 2014, 2014.

[12] S. Chen, N. M. Luc, Rrmse voting regressor: A weighting function based improvement to ensemble regression, 2022. URL: https://arxiv.org/abs/2207.04837. doi:10.48550/ARXIV.2207.04837.

[13] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al., Xgboost: extreme gradient boosting, R package version 0.4-2 1 (2015) 1–4.

[14] S. Acharya, D. Swaminathan, S. Das, K. Kansara, S. Chakraborty, D. Kumar, T. Francis, K. R. Aatre, Non-invasive estimation of hemoglobin using a multi-model stacking regressor, IEEE journal of biomedical and health informatics 24 (2019) 1717–1726.

[15] D. Azcona, E. Moreu, F. Hu, T. E. Ward, A. F. Smeaton, Predicting media memorability using ensemble models, CEUR Workshop Proceedings, 2020.

[16] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: Theory and practice, Neurocomputing 415 (2020) 295–316.