Interpretable models for dementia diagnosis

Tejas Sujit Bharambe
tbharamb@usc.edu

Jon Luo
jzluo@alumni.cmu.edu

Jhony Mejia
ja.mejia12@uniandes.edu.co

1 Project overview

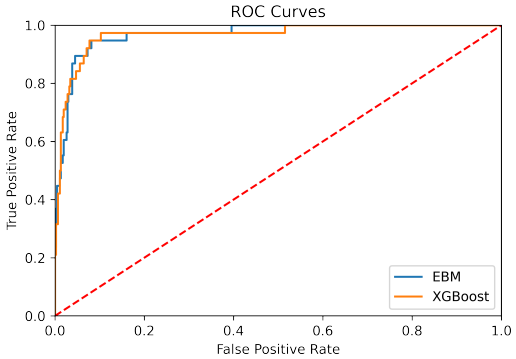
Our project is two-fold: first, we propose the use of interpretable models for the diagnosis of dementia. Second, we attempt to use these models to investigate the most important features and potentially uncover underlying patterns in the features. To this end, we used two approaches: XGBoost with calculation of SHAP values for model explanation, and explainable boosting machine, an implementation of GA2M (generalized additive models plus interactions).

Both approaches showed high Accuracy and AUC, specifically for cognitive tests and daily activities reported by an informant. Interestingly, both approaches seem to indicate that cognitive tests and daily activities might be equally important for a diagnosis of dementia. The novelty of our job is that we were able to interpret which features play a role in clinical diagnosis including at the individual prediction level, which can be relevant for clinical implementation.

1.1 Top-down approach

Using most of the features collected in the LASI-DAD study including cognitive tests, activity scores, informant questionnaires, demographics, and so on, both models performed well.

	AUC	Accuracy	Recall	Precision	Kappa
EBM	0.969	0.951	0.579	0.71	0.615
XGB	0.968	0.96	0.684	0.765	0.702



Data augmentation techniques mentioned later to combat the class imbalance and hyperparameter tuning could be explored in the future to assess further improvement.

More importantly, we are able to look at the features that contribute the most to the predictions. As expected, the model identifies the cognitive tests and activity scores are the most important in diagnosing dementia (Figure 1).

The model also allows for deeper inspection of each feature's contribution. We can observe from the shaping plot of the informant questionnaire feature that the model has learned that the questionnaire indicates an increased risk of dementia beginning around score 4 (Figure 2).

This kind of interpretable model allows for looking into the reasoning behind the model's individual predictions, which is critical for healthcare diagnoses. For example, Lee et al. suggested implementing their SVM as virtual rater in the consensus diagnosis

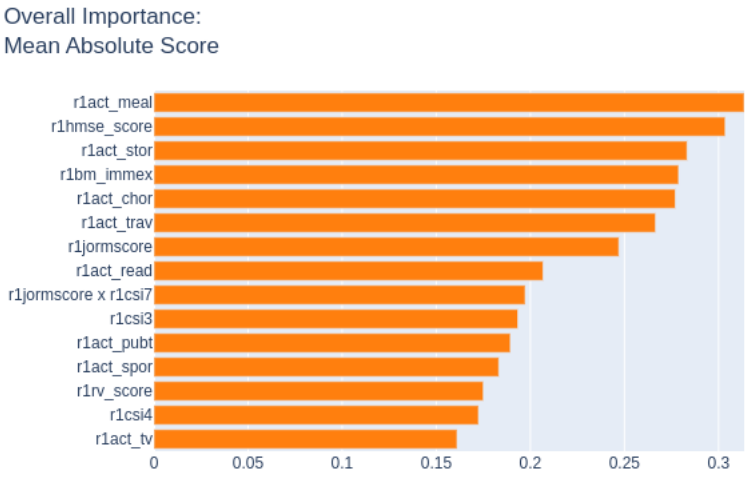


Figure 1: Global feature importances determined by EBM. Score is in log-odds.

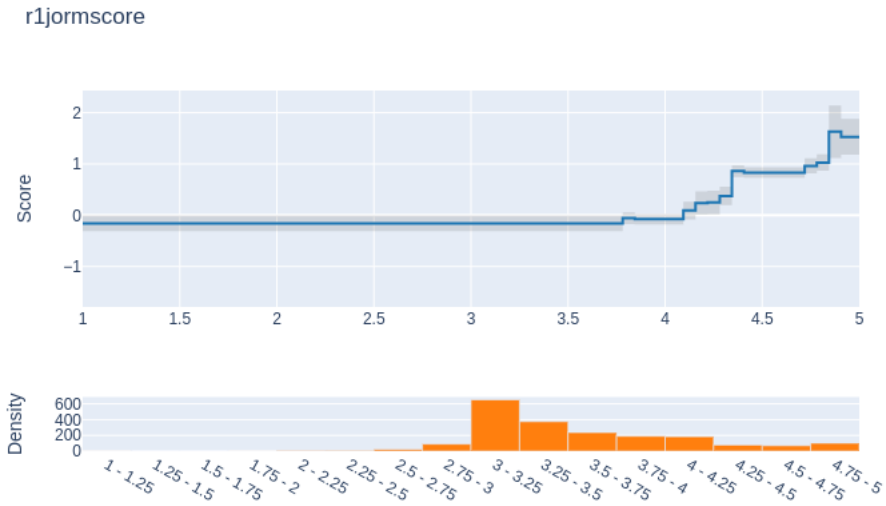
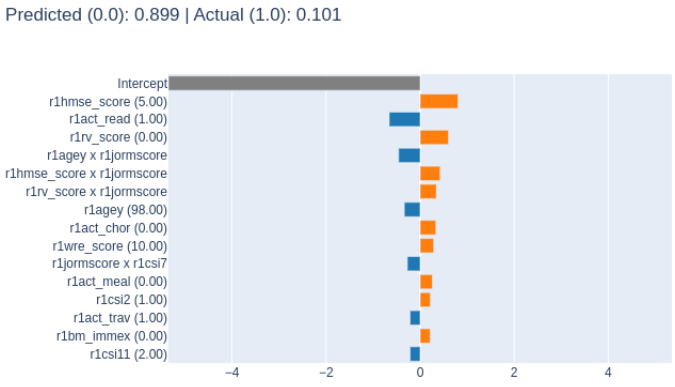


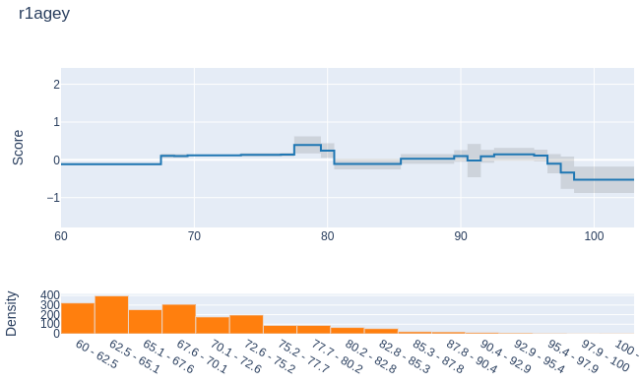
Figure 2: Shaping plot of informant questionnaire. Score is in log-odds.

process for LASI-DAD. Consider a situation where the model disagrees with the clinicians; with an interpretable model, the clinicians would be able to look at the individual prediction and see why the model made its choice. Figure 3 shows an example where the model disagreed with the other raters, predicting no-dementia where the consensus was dementia. It is immediately apparent that the participant is 98 years old, which the model unexpectedly says lowers dementia risk. In fact, the second largest factor in the negative diagnosis is the interaction of the informant questionnaire and age (Figure 3a).

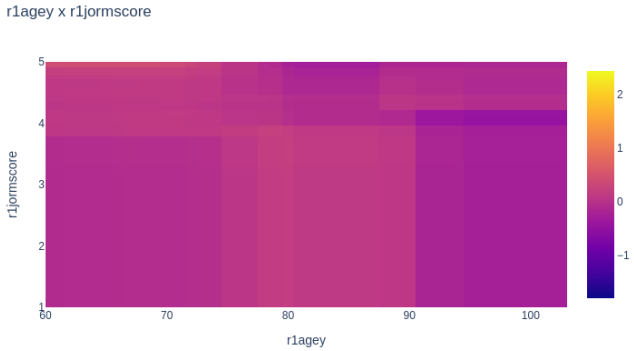
That the model learned that being in the upper 90s in age actually decreases risk of dementia (Figure 3b). We know this isn't true - the density plot for age suggests that there is simply a lack of data on people that old. Indeed, we find in the data a lot of outliers in that upper age range with low CDR. Figure 3c illustrates how the model views the interaction of age and informant questionnaire score, and we see the effect the lack of older participants in the cohort affects the data. So in this manner, we can also find issues with the data using this interpretable model, and can be more confident in dismissing the model's prediction for this participant.



(a) Local interpretation



(b) Importance plot for age



(c) Interaction of age and informant questionnaire (jormscore)

Figure 3: Local explanation of misclassified participant.

1.2 Bottom-up approach

We further investigated the features category-wise to determine whether a sufficiently performant model could be built using easily obtainable data to approach the problem in a similar manner as a clinician would. We looked at the most commonly available features first before assessing whether or not to subject a patient to further testing which can be costly or intensive. We use XGBoost with stratified K-Fold CV for hyperparameter tuning, and address class imbalance using upsampling techniques like SMOTE or SMOTEENN, or simply downsampling down to the category with fewer subjects.

We focused on four categories of data: Demographics (age, gender, education, literacy and rurality); Daily activities reported by a respondent; Cognitive tests; and Venous Blood Samples. Tests on venous blood assays resulted in poor performance, and so will not be considered. The results for the other three categories of data can be found in Table 1.

		AUC	Accuracy	Recall	Precision
Demographics	SMOTEENN	0.712	0.767	0.553	0.172
	Downsampling	0.761	0.692	0.736	0.040
Daily Activities	SMOTEENN	0.870	0.847	0.605	0.271
	Downsampling	0.876	0.734	0.842	0.052
Cognitive Tests	SMOTEENN	0.807	0.919	0.289	0.440
	Downsampling	0.900	0.778	0.842	0.062

Table 1: Performance metrics for individual categories of features.

By only looking at demographics we obtained an acceptable performance (AUC = 0.76, Acc = 0.77). Additionally, despite the roughly homogeneous low levels of education and young LASI population, we found that Age and Educations still are the greatest predictors of dementia (Fig 4).

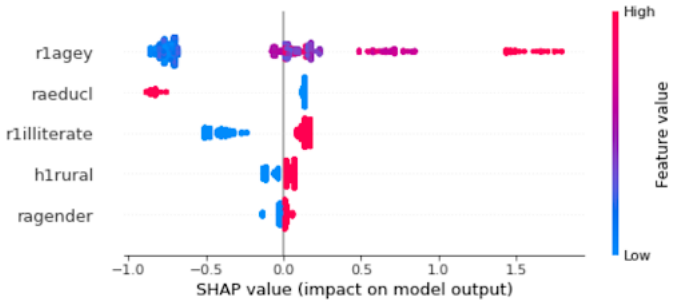


Figure 4: SHAP plot for demographics features.

Similarly to the Top-down approach, we found that cognitive tests and daily activities had a better performance at diagnosing dementia (AUC = 0.7, Acc = 0.7) (Table 1). Our findings of the importance of cognitive tests and daily activities supports previous literature that both factors can reflect an increased risk of dementia [1]. Furthermore, the best scores of Recall (0.842) and Precision (0.271) for our daily activity models are similar to the ones obtained by a Swedish study that aims to diagnose dementia using an informant questionnaire (Precision=[0.28-0.63], NPV=[0.96-1]) [2].

In general, we found that by looking at individual categories we could obtain an acceptable recall, but a bad precision. We also found that metrics like AUC and Accuracy might not be enough for correctly assessing the performance of a model, as we always had AUC greater than 0.71 but low values of precision. In terms of methods to deal with unbalanced data, we found similar AUC and Accuracy results for SMOTEEEN and simple downsampling. However, it seems that downsampling offers an improvement in Recall but an extremely low Precision, while SMOTEEEN can create decreases in Recall at the expense of getting a better Precision. In general, we advise caution and trying different strategies when dealing with unbalanced data. In sum, we think that some individual categories can be useful for screening due to its high recall. Then, a combination of those affordable and easily collected categories can be used to increase the precision as shown in the top-down approach.

References

[1] Potter, G. G., Plassman, B. L., Burke, J. R., Kabeto, M. U., Langa, K. M., Llewellyn, D. J., ... Steffens, D. C. (2009). Cognitive performance and informant reports in the diagnosis of cognitive impairment and dementia in African Americans and whites. Alzheimer's Dementia, 5(6), 445–453. doi:10.1016/j.jalz.2009.04.1234

[2] Svensson, A., Granvik, E., Sjögren Forss, K. (2020). Performance of the Eight-item Informant Interview to Differentiate Aging and Dementia within a context similar to the Swedish primary healthcare sector: a systematic review of diagnostic test accuracy studies. Scandinavian Journal of Primary Health Care, 38(4), 454–463. doi:10.1080/02813432.2020.1844370