

# Case Study for Omniscient Neurotechnology

Upul Senanayake

December 12, 2021

## Abstract

The case study provided by Omniscient (o8t) as part of the recruitment process for Lead Data Scientist role (LDS) consists of a dataset that has demographic and neuropsychological test (NT) scores of patients originating from Clinical Center for Dementia (CCD). They want to know how and where they can use this data which makes this problem a unique and open-ended one. I have attempted to tackle this in two ways; (i) conducting exploratory analysis to identify obvious data quality issues and redundant tests, and (ii) building a machine learning model to automate the diagnosis process. The former was done in anticipation of identifying any control gaps in CCD's process and to reduce their operational costs as well the costs to the patient as running neuropsychological tests can be very expensive. The latter was implemented to automate the diagnosis process reliably for new incoming patients using the data they have collected and shared herewith. An obvious data quality issue was found in patient weight which did not fit into the expected range of an average adult. Some highly correlated and hence redundant tests were identified and these were further filtered by examining the tests that are most important for the automated diagnosis system to make a recommendation on reducing the number of tests conducted. A baseline machine learning model and an alternate model was implemented along with a continuous integration and continuous development pipeline in Github to iterate and improve on the models. The current best model, which is a random forest model, has a cross-validated accuracy of 87.1%. Lastly, a presentation was created to present these findings to the CCD and uploaded to Github.

## 1 Introduction

The purpose of this short white paper is to act as supplemental information for the study conducted as part of the recruitment process for LDS role. Rather than delving into the exploratory analysis or modelling, this will be used to elaborate the thought process behind formulated research questions and the machine learning operational (MLOps) pipeline developed. Therefore, the rest of this white paper is structured as follows. The next section describes why the research questions were formulated in the current format followed by Section 3 explaining the MLOps pipeline. Section 4 concludes this white paper.

## 2 Research Questions

The case study that was emailed did not set any concrete expectations. As such, it was considered as an open ended study. The provided dataset and context was taken into consideration in designing the research questions. As CCD receives patient referrals and conducts detailed screening, I considered that as their fundamental revenue stream. The dataset consisted of 210 patients and as

it is a year’s worth of data, it appears that they are screening over 200 patients per year. A cursory search on costs associated with running a battery of neuropsychological tests proved that it is an expensive exercise for the patients. It was then assumed that conducting these tests will also be the significant part of operational costs of CCD. Therefore, two connected research questions were formulated having the operational costs in mind.

1. Are there any obvious Data Quality (DQ) issues?
2. Are there any highly correlated neuropsychological tests that CCD may be able to refrain from conducting without compromising the diagnosis accuracy?

Lastly, given this was a labelled dataset, it presented an opportunity to build a supervised machine learning model to automate the diagnosis of Dementia and MCI. In the context of CCD, this is important as it can be used as a standalone diagnosis model or a part of computer assisted diagnosis pipeline for clinicians. Each of these research questions will be explored below.

## 2.1 Data Quality Issues

A fundamental question of any data generating process is whether there are data quality issues in the generated data. This is especially important if the data quality issues are systemic and hence can be avoided. To this extent, the provided dataset was explored to find any anomalous values, outliers and missing values. It did not appear there were any outliers but there were some missing values. The proportion of missing values were minor, and therefore did not hinder the analysis. However, CCD can focus on minimizing this as much as possible.

The measurement for weight was found to have data quality issues as it did not fall into the weight range of an average adult. As there were no measurement unit, standard weight measurement units of KiloGram and Pounds were considered. The weight range in the dataset was from 0 – 20 with a mean of 10.44 and therefore concluded as having data quality issues.

By addressing these data quality issues, CCD will be able to maintain the integrity of their data and mitigate any control gaps identified. This will ultimately translate into cost savings as well as increased revenue from the improved reputation.

## 2.2 Highly Correlated Neuropsychological Tests

A battery of neuropsychological tests can be expensive to conduct and will be equally expensive for the patient. Therefore, it stands to reason that if CCD can minimize the amount of NT carried out as part of the screening and diagnosis process, that will translate into significant operational cost deductions for CCD as well as reduced charges for patients improving the good will towards CCD. An additional benefit of this would be that the time taken to complete the diagnosis will be reduced which can be a relief to elderly patients.

After the exploratory analysis, a correlation matrix was drawn up to identify highly correlated tests as depicted in Figure 1. It can be seen that tests such as MMSE and NPI are highly correlated. This is self-explanatory as some of these tests evaluate the same domain. This can be further examined by using the importance a machine learning model would place on each of these feature when making a patient diagnosis as depicted in Figure 2. It is clear that model considers tests such as MoCAB and MMSE much more important than tests such as HA14 Interview performance. Therefore, this can be used to propose an ideal or reduced set of neuropsychological tests to conduct for diagnosis.

I have however not proposed a set of tests as part of this case study as it needs to be done in close collaboration with subject matter experts such as Neuropsychologists as well as clinicians and Neuroscientists.

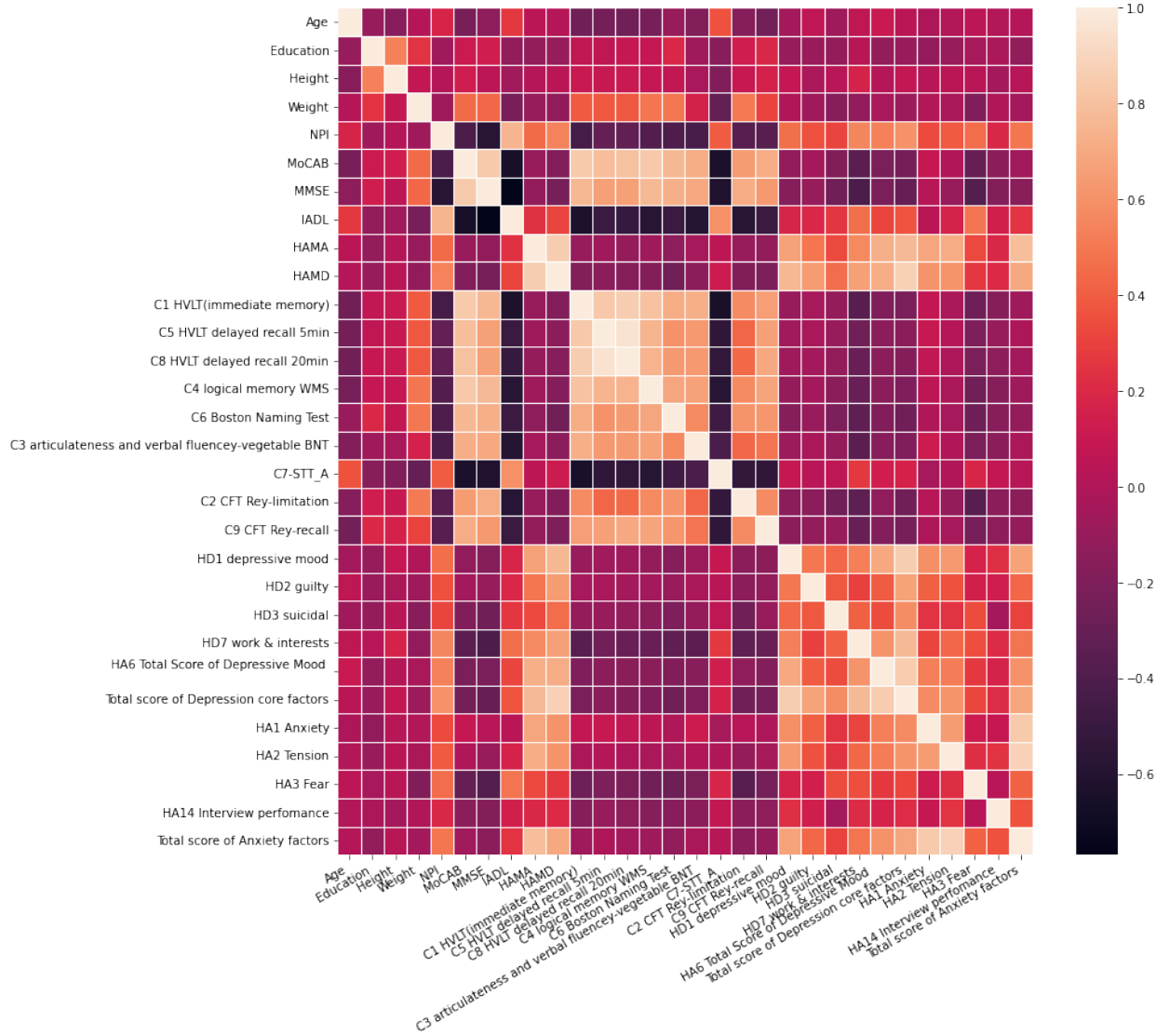


Figure 1: Highly Correlated Features in the Dataset

## 2.3 Machine Learning Model

As part of this case study, I have developed a ML model to automate diagnosis of future patients using currently available data. The objective behind developing this ML model was not to come up with the best in-class model but facilitate a starting point for a conversation with CCD. Therefore, this was modelled as a multi-class classification problem and the baseline model developed was a random forest model without any hyperparameter optimization. An alternate Support Vector Machine (SVM) model was also developed and hyperparamters were optimized using a grid search. These were tested using cross validation as well as train/test split. The random forest model had

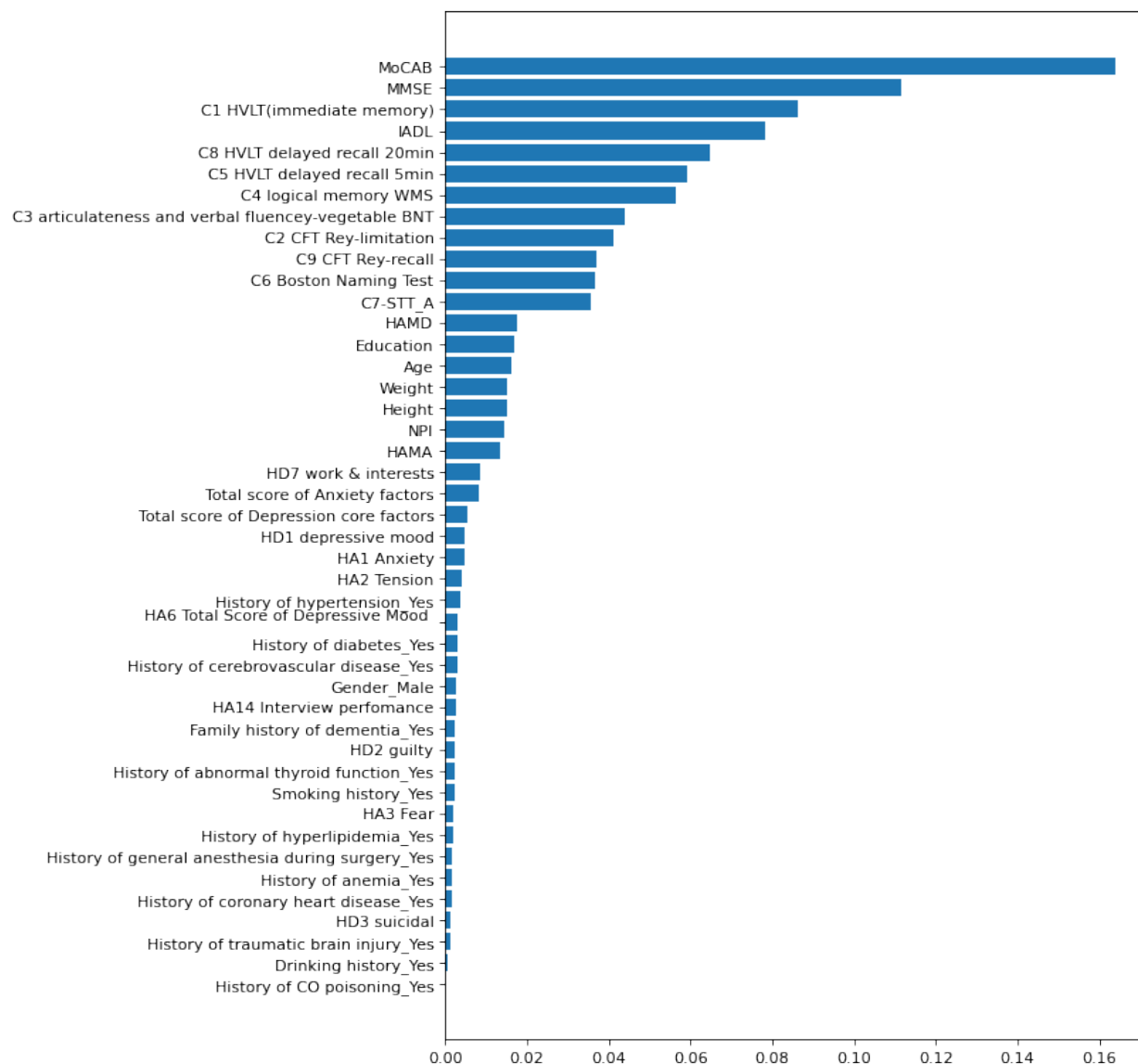


Figure 2: Feature Importance

the best performance when considering the balanced accuracy which is the mean of recall. It was important to consider balanced accuracy as this was an unbalanced dataset in which, accuracy may overstate the true performance of the model. Other metrics such as specificity and precision were considered as well.

The confusion matrix for the random forest model is depicted in Figure 3. As it can be seen, only 2 instances from the test set were wrongly classified. These wrongly classified instances were inspected and a rationale for the wrong classification was provided in the Jupyter Notebook.

### 3 MLOps Pipeline

One of the challenges a data scientist faces when completing a case study or working on a business problem is continuous integration and continuous development (CI/CD) of the models they are

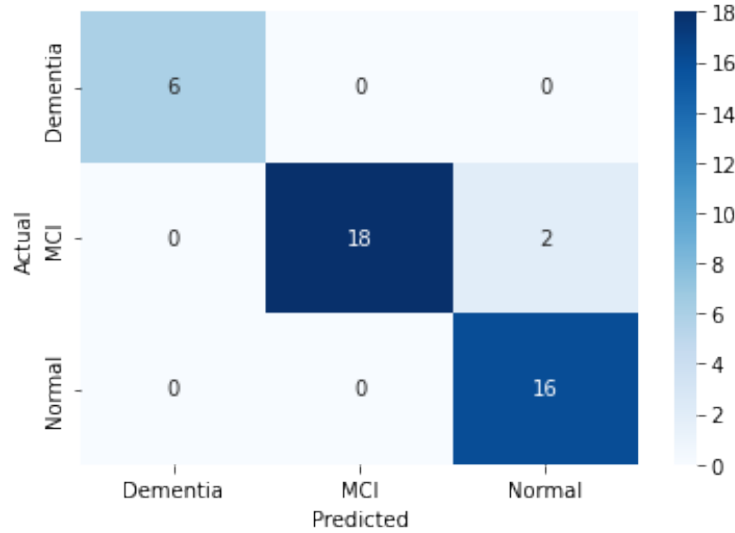


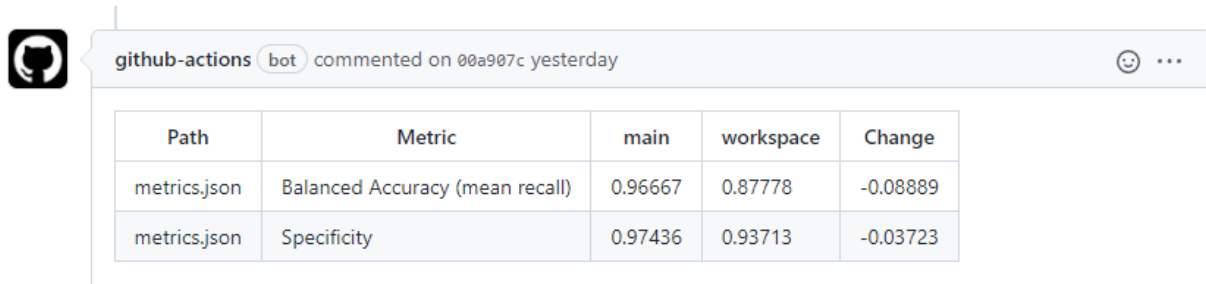
Figure 3: Feature Importance

building. This is different to a traditional DevOps pipeline as data scientists have to build, test and deploy using real data and there are more artefacts than just code to manage. To this extent, I wanted to demonstrate a CI/CD pipeline as part of this case study that allows one to seamlessly iterate and develop on models without the nuances of having to version control their data, models as well as results. I have used DVC (Data Version Control) and CML (Continuous Machine Learning) libraries for this purpose along with GitHub actions. It is using a Github repository and has the following flow.

1. The main branch of this repository will always carry the best model available.
2. When we want to improve upon the current best model or experiment with other models, we can branch the main and start working on a better model (ex: [alternate\\_model branch](#)).
3. Once we are satisfied with the new model, we can commit it to the new branch and create a pull request to merge it to the main branch if the model is better.
4. In the backend, every time a commit is made to the repository, we will setup a docker container with the required packages, execute the model as [follows](#), report its performance and artefacts as a markdown report and also compare the performance of the new model to the current best model and include that in the markdown report as depicted in Figure 4.
5. This comparison can then be used to either approve the pull request if the new model is better or reject and rework on another model.
6. Ultimately, this pipeline provides a data scientist with the opportunity to iterate and build new models that will automatically get tested, evaluated and compared before getting deployed.

The MLOps pipeline will be particularly important for the next steps of this case study as the discussions with CCD ensues and more refinements to the models need to be done. Not only that but it will also keep track of the changes to your data, the changes to your code, changes to

1 comment on commit 00a907c



The image shows a GitHub comment interface. At the top left is the GitHub logo. To its right, the text 'github-actions bot commented on 00a907c yesterday' is displayed. Below this is a table with five columns: Path, Metric, main, workspace, and Change. The table contains two rows of data for 'metrics.json'.

Path	Metric	main	workspace	Change
metrics.json	Balanced Accuracy (mean recall)	0.96667	0.87778	-0.08889
metrics.json	Specificity	0.97436	0.93713	-0.03723

Figure 4: Automated Performance Comparison. For full report, please visit the [commit](#).

your hyperparameters as well as a snapshot of everything over each run which can help with any regulatory reporting CCD will have to undertake as part of automated diagnosis of Dementia/MCI.

## 4 Conclusion

In this short white paper, I have presented my thought process behind the analysis that was carried out as part of this case study. The exploratory analysis as well as the machine learning model development is covered in the Jupyter Notebook and as such, was not covered here. The focus here has been the formulation of the research questions and the rationale behind developing a MLOps pipeline. These will be instrumental to initiate a conversation with CCD and decide on next steps.