Classification of the Alzheimer's Disease Using a SVM Model*

Kaggle Competition Report

Rahma Binth Mohammad

This report is based on project that used supervised machine learning techniques to build a model that classifies those diagnosed with Alzheimers. After preprocessing and feature selection, we evaluated several algorithms and selected the best-performing model based on predictive accuracy. Our final model achieved a Kaggle score of _____, ranking _____ among the other models.

Table of contents

*Code and data are available at: https://github.com/rahmabinth/alzheimers.

1 Introduction

There were over 55 million people worldwide living with dementia in 2020 and by 2050, it is projected to affect 139 million people (cite Alzheimer's International). Due to the fact that this is a huge amount of the population, it is important to identify the characteristics and lifestyle choices of the people it affects in order to find a pattern or a prevention method. Finding a pattern can help with early detection, which would allow for better prevention. Thus, this paper focuses on predicting Alzheimer's Disease in patients using a dataset from Kaggle that contains data collected from patients. We treated the diagnosis as a binary outcome (1 = Yes, 0 = No) and aimed to use machined learning algorithms to find the best model for performance.

Our final and best performing model achieved a score of _____ and ranked _____ on the leaderboard. The model performed as follows. Evaluation metrics —–> Strongly in AUC or F1 score (find out).

Our model matters because it allows for early detection of Alzheimers and demonstrates the potential of machine learning in highlighting key predictive factors in patient data.

The remainder of this paper is structured as follows. Section 2....

2 Data

Kaggle Username: rahmabm

Final Kaggle Ranking:

Final Private Score:

2.1 Overview

We conducted our analysis using the statistical programming language R (R Core Team 2023). Our data was pulled from the Kaggle Competition called Classification of the Alzheimer's Disease dataset [cite competition]. The dataset provides patient data for _____ patients. Each row contains a different individuals patient data with _____ different variables found at [cite]]. It contains demographic details, lifestyle factors, medical history, cognitive assessments, and symptoms for each patient.

2.2 Measurement

While the dataset is based on real-world patient assessments and health records, certain variables are expressed as measurable features with the use of standardizes clinic tests. This includes several of the cognitive and function assessments in the reported data such as Mini-Mental State Examination Score, Functional Assessment Score, and Behavioural Problems. Others have been self-reported as binary variables, such as the symptoms. Variables such as Confusion, measured the presence of confusion, along with variables like Disorientation and Forgetfulness.

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

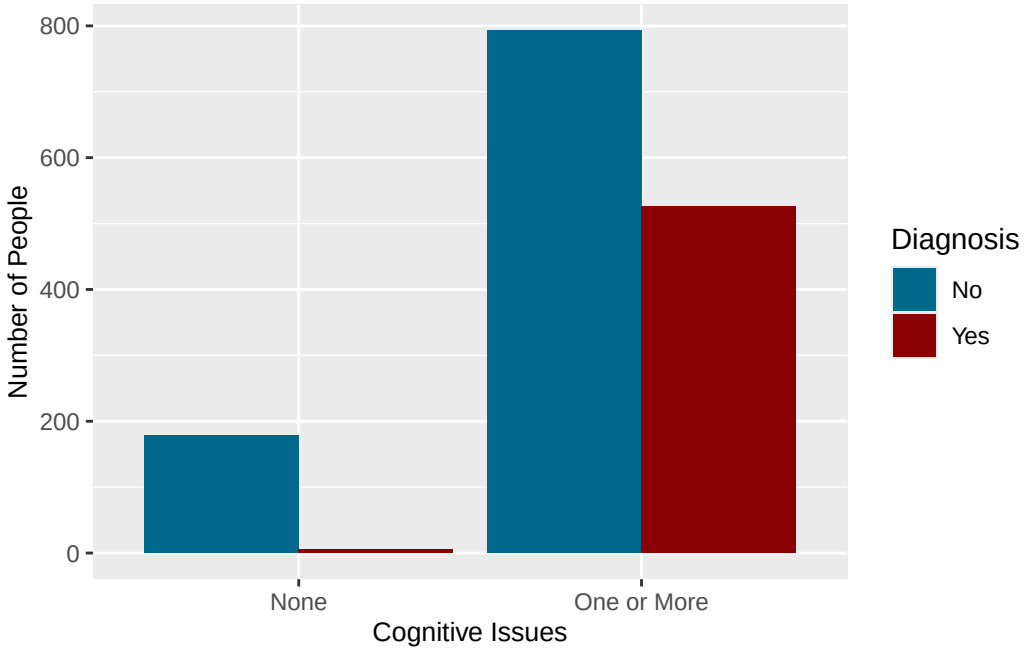

Figure 1: Diagnosis vs. Cognitive Issues

2.3 Data Preprocessing and Feature Engineering

We started by inspecting the data for missing values and irrelevant variables. We found variables such as the "DoctorInCharge", and "PatientID" to be irrelevant to our problem and removed them. This also allowed for randomness in our data. We looked for patterns within the data but segregating the data into separate categories. We identified those with anytime type of cognitive issues and looked for a pattern within those diagnosed and undiagnosed with Alzeheimers. These issues were found through cognitive and functional assessments as mentioned earlier. They include _____. We found that 88% of the patients had one or more cognitive issue and approximately 40% of them were diagnosed with the diseases. The most vital/significant find was that only 1% of those undiagnosed with the disease had one or more cognitive issues and approximately 40% of those with cognitive issues were diagnosed with the disease. This indicated a very high chance/correlation that if a patient had a cognitive issue, they were more likely to have the disease.

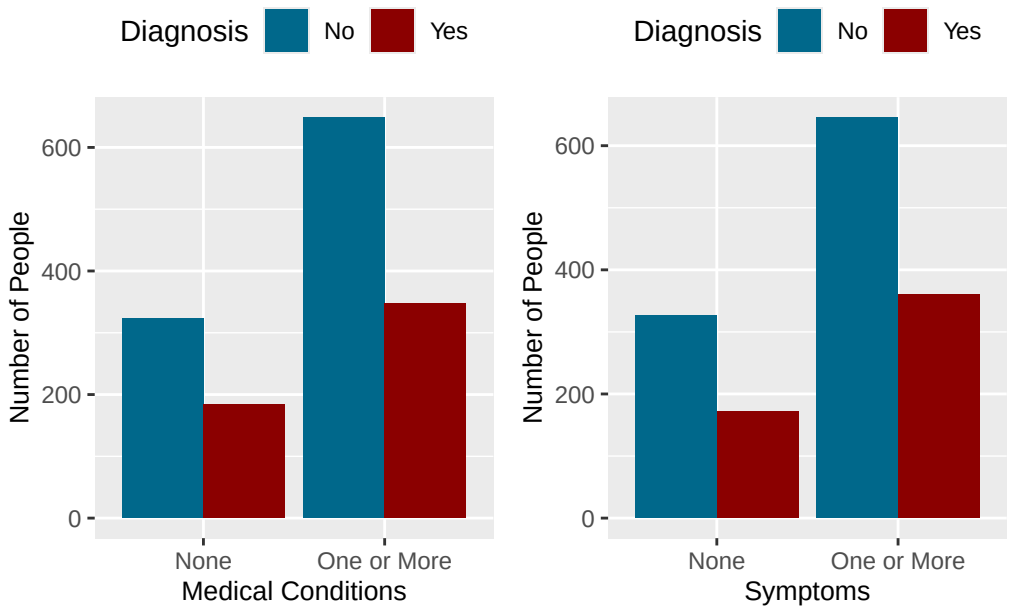

Figure 2: Diagnosis vs. Medical History and Diagnosis vs. Symptoms

We did the same with medical history. If the patient had one or more of the the following, we grouped them, a family history of the disease, cardiovascular disease, diabetes, depression, history of head injury, hypertension. We found approximately 66% of the patients had one or more medical condition and 35% of them were diagnosed. It was also shown that 35% of those diagnosed didn't have one or more of these medical condition. It was less conclusive then those with cognitive issues but we found it relevant to test out models with this feature later on, since a significant number of those with the condition were diagnosed. We saw a similar

pattern with patient symptoms, where we grouped together those with one or more of the symptoms, "Confusion", "Disorientation, "PersonalityChanges", "DifficultCompletingTasks", and "Forgetful". See Figure 2.

2.4 Feature Selection

In addition, we looked at how each variable/factor played into the diagnosis. **?@sec-appendix.2** provides the code to replicate this by changing the name for the variable. Our analysis revealed that certain features had the most impact. For example, by examining all the variables, we saw 61% of those diagnosed were Caucasians, and 61% of those diagnosed either didn't attend high school or only attended high school and didn't hold a Bachelor's Degree or Higher Education. However, gender didn't seem to play a role as the number of men and women diagnosed were approximately the same.

3 Model

Our initial goal for our modelling strategy was twofold, balancing interpretability and predictive accuracy. With this in mind, we used the three factors mentioned in Section 2.3, cognitive issues, medical conditions, and symptoms, we tried to create a logistic regression model and used cross validation method to train and test data but failed to get higher than 62% predictability rate. Hence, we prioritized predictive performance by choosing to do a Support Vector Machine model, which offered limited interpretability. The SVM performed significantly better than our previous tested models. However, interpretability decreased. **?@sec-discussion** identifies this tradeoff.

3.1 Model set-up

We define the outcome variable $y_i \in \{0, 1\}$ as the diagnosis of Alzheimer's Disease, where $y_i = 1$ indicates a positive diagnosis and $y_i = 0$ indicates a negative diagnosis. To predict this outcome, we trained a Support Vector Machine (SVM) classifier using the following selected predictor variables whether the feature for individual i is denoted by $x_i = (x_1, x_2, ..., x_p)$.

x_1: Ethnicity, x_2: Gender, x_3: MMSE score (Mini-Mental State Examination score), x_4: MemoryComplaints, x_5: ADL, x_6: FunctionalAssessment, x_7: BehavioralProblems, x_8: HeadInjury, x_9: Forgetfulness

The model estimates a function $f(\mathbf{x}_i)$ such that:

$$\hat{y}_i = \begin{cases} 1 & \text{if } f(\mathbf{x}_i) \geq 0 \\ 0 & \text{if } f(\mathbf{x}_i) < 0 \end{cases}$$

We run the model in R (R Core Team 2023) using the `e1071` package (**e1071?**). In our SVM model, $f(\mathbf{x})$ is the decision function — a number that tells how far a point is from the decision boundary (i.e., the separating surface between classes).

If $f(\mathbf{x}_i) \geq 0$, the model predicts class 1 (positive diagnosis).
If $f(\mathbf{x}_i) < 0$, it predicts class 0 (negative diagnosis).

3.1.1 Model justification

3.2 Evaluation Metrics and Model Performance

Before submitting on Kaggle, we worked with the "train_csv" file and split it into two, 70% of the data was used for training and 30% used for validation. Our validation set proved useful when testing the models. Hyperparameter tuning via 10-fold cross validation helped identify the optimal parameters, cost = 10 and gamma = 0.1. These parameters provided a cross-validated error rate of 0.0997, which shows that the model misclassifies about 10.26% of observations and will do so on new and similar data.

Our Support Vector Machine (SVM) classifier with these parameters achieved an accuracy of 89% and sensitivity of 90.1% (accurately identifying the undiagnosed patients) and specificity of 87% (accurately identifying the diagnosed patients) on the validation set. The ROC curve in Figure 3 shows this relationship and plots how well our model separates the two classes. The area under the curve (AUC) of 0.941 shows that our model has excellent ability to distinguish between Alzheimer's-positive and Alzheimer's-negative patients.

Note we chose a radial kernel rather than a linear one as it provided about 9% less accuracy.

4 Results

As mentioned in Section 2.4, all the variables were examined to see a _____. However, when we simply inputted those features into the SVM, it only gave us an accuracy of 86%. To improve our model and better understand each feature's importance, we conducted a leave-one-feature-out sensitivity check, essentially removing one feature at a time, retraining the model, and checking how the accuracy changes. Using this, we found certain features to be extremely important to the model's predictive ability. For example, MMSE and FunctionalAssessment were very important. This check also allowed us identify which features to retain and exclude. The final set of features used is listen in **?@sec-model-set** and reflects the most significant contributors to the Alzheimer's disease. **?@sec-discussion** will examine the relevance of these features.

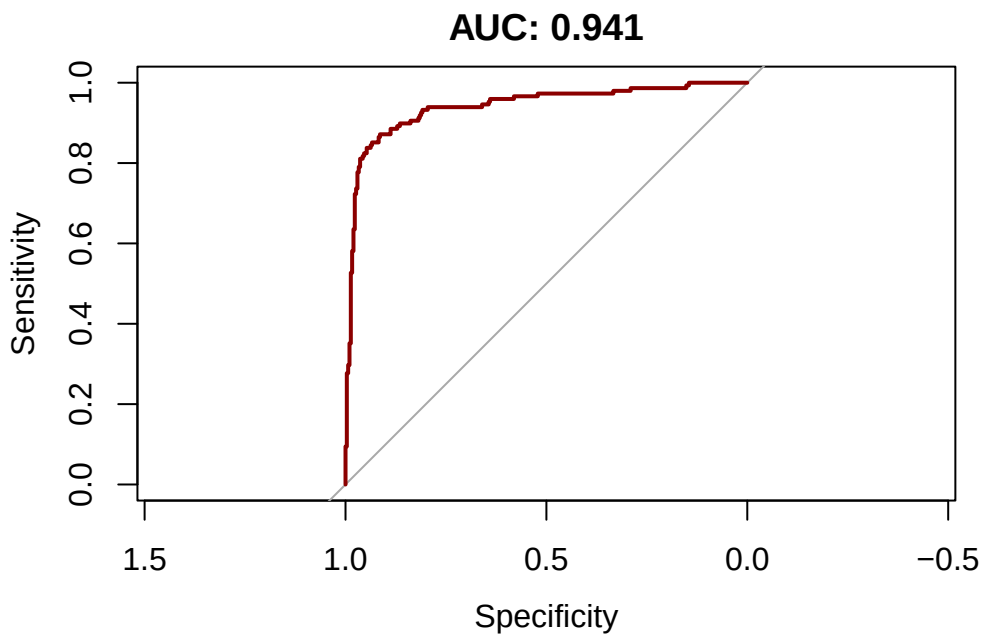

Figure 3: ROC curve

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In we implement a posterior predictive check. This shows...

In we compare the posterior with the prior. This shows...

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...

Checking the convergence of the MCMC algo-
rithm

References

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.