Predicting a Diagnosis of Alzheimer's Disease

Shelley Jones

Agenda

- Context
- Business Question
- Data Dictionary
- Exploratory Data Analysis
- Feature Correlation
- Modelling
- Model Evaluation
- Conclusions
- Questions

Context

- Alzheimer's Disease (AD) is a cruel disease that steals away and never gives back
- It can progress gradually or quickly, but patients inevitably lose their memory, their ability to think, to plan and to recognise even those closest to them
- AD affects around 50 million people worldwide, is one of the 21st century's most challenging health problems and is expected to triple by 2050
- Currently, the diagnosis of AD is largely based on clinical symptoms, including cognitive testing (e.g. through questionnaires designed to measure mental processes such as memory, attention, problem-solving and language); it is only accurate in 70-80% of cases
- This means that up to 30% of families have been told their loved one has Alzheimer's disease when actually they don't.

The bad news: There is currently no cure and many experimental treatments have failed clinical trials over the years
The good news: A LOT of research going on in this area including the use of biomarkers for earlier prediction and management (fingers crossed).

Business Question

"Using data that is already publicly available, can we establish a classification model that accurately predicts a diagnosis of Alzheimer's Disease with at least an 85% accuracy using clinical features (e.g. cognitive test scores, clinical data) to enhance early diagnosis and support clinical decision-making within the next 4 weeks?"

- **Specific**: The question targets the use of modelling for predicting Alzheimer's disease based on specific types of data (cognitive test scores and clinical data)
- **Measurable**: The performance of the model will be assessed using specific metrics like accuracy (min 85%), recall, and AUC
- **Achievable**: The goal of achieving at least 85% accuracy is realistic given current machine learning and data analysis capabilities
- **Relevant**: Question addresses the importance of improving early diagnosis and supporting clinical decisions, which is critical for effective management of Alzheimer's disease
- **Time-bound**: Sets a deadline of 4 weeks to achieve the target 85% accuracy score.

Data Dictionary

The dataset contained 35 features in which 'Diagnosis' was the target feature.

- Patient Information: Patient ID (a unique number)
- Demographic Details (4 features)
- Lifestyle Factors (6 features)
- Medical History (6 features)
- Cognitive and Functional Assessments
  - MMSE: Mini-Mental State Examination score, ranging from 0 to 30. Lower scores indicate cognitive impairment
  - FunctionalAssessment: Functional assessment score, ranging from 0 to 10. Lower scores indicate greater impairment.
  - MemoryComplaints: Presence of memory complaints, where 0 indicates No and 1 indicates Yes.
  - BehavioralProblems: Presence of behavioral problems, where 0 indicates No and 1 indicates Yes.
  - ADL: Activities of Daily Living score, ranging from 0 to 10. Lower scores indicate greater impairment.
- Symptoms
  - Confusion: Presence of confusion, where 0 indicates No and 1 indicates Yes.
  - Disorientation: Presence of disorientation, where 0 indicates No and 1 indicates Yes.
  - PersonalityChanges: Presence of personality changes, where 0 indicates No and 1 indicates Yes.
  - DifficultyCompletingTasks: Presence of difficulty completing tasks, where 0 indicates No and 1 indicates Yes.
  - Forgetfulness: Presence of forgetfulness, where 0 indicates No and 1 indicates Yes.
- Diagnosis Information: Diagnosis status for Alzheimer's Disease, where 0 indicates No and 1 indicates Yes.
- Confidential Information: DoctorInCharge

Initial Exploratory Steps

[4]:

	0	1	2	3	4
PatientID	4751	4752	4753	4754	4755
Age	73	89	73	74	89
Gender	0	0	0	1	0
Ethnicity	0	0	3	0	0
EducationLevel	2	0	1	1	0
BMI	22.927749	26.827681	17.795882	33.800817	20.716974
DifficultyCompletingTasks	1	0	1		
Forgetfulness	0	1	0		
Diagnosis	0	0	0		
DoctorInCharge	XXXConfid	XXXConfid	XXXConfid	XX	

- 2149 rows, 35 columns
- No missing data
- Variables were a mixture of integers, objects and floats
- Checked categorical columns for ‘unexpected’ entries: none found
- Checked for duplicate rows: None were identified (this was expected as we have a unique pt ID)
- Removed ‘PatientID’ and ‘DoctorInCharge’ as they added no value

EDA Cont’d

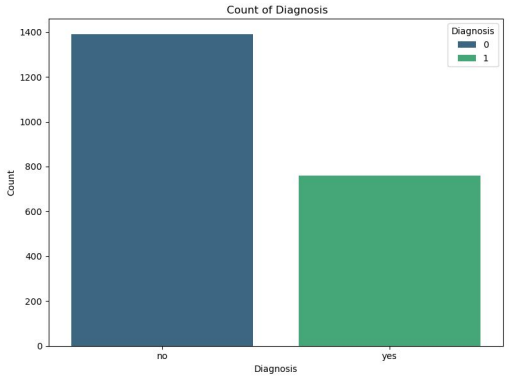
- Ran initial stats (df.describe)
  - All Yes / No questions answered appropriately i.e. no ‘rogue’ data
  - Nothing remarkable about the data although some patients had ‘nearly perfect’ scores for assessments (these max scores are not ‘true’ - they should be a whole number)
- Checked categorical columns for unique entries to ensure ‘responses’ were appropriate

MMSE	2149.0	14.755132	8.613151	0.005312	7.167602	14.441660	22.161028	29.991381
FunctionalAssessment	2149.0	5.080055	2.892743	0.000460	2.566281	5.094439	7.546981	9.996467
MemoryComplaints	2149.0	0.208004	0.405974	0.000000	0.000000	0.000000	0.000000	1.000000
BehavioralProblems	2149.0	0.156817	0.363713	0.000000	0.000000	0.000000	0.000000	1.000000
ADL	2149.0	4.982958	2.949775	0.001288	2.342836	5.038973	7.581490	9.999747

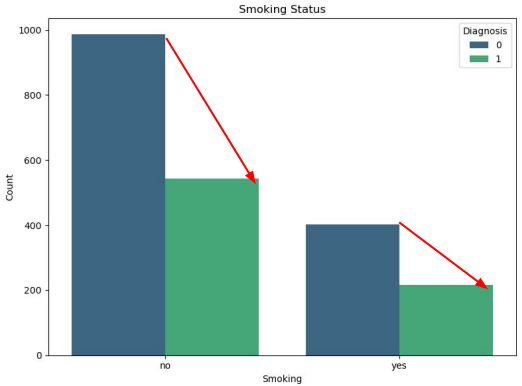
EDA cont’d

Target was moderately imbalanced

- ~35% +ve
- ~65% -ve

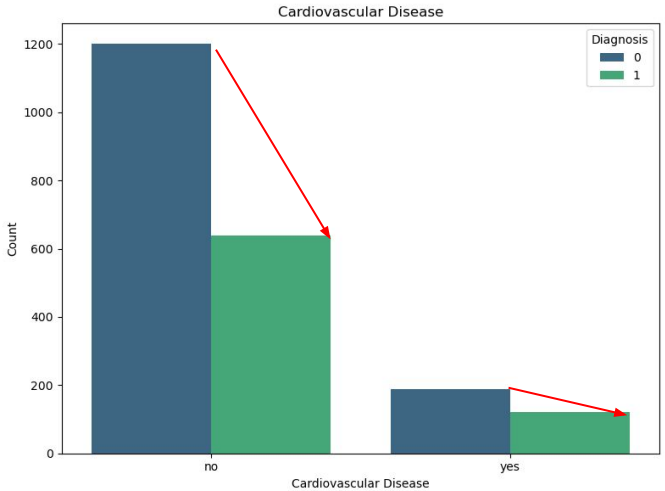


Other categorical data generally balanced across yes / no Diagnosis……

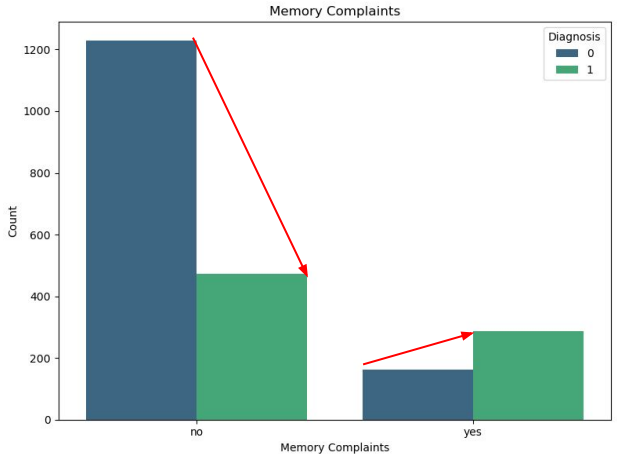


EDA cont’d

…..Except for e.g. CV disease, Hypertension



Other categorical data depicted expected increases with +ve diagnosis

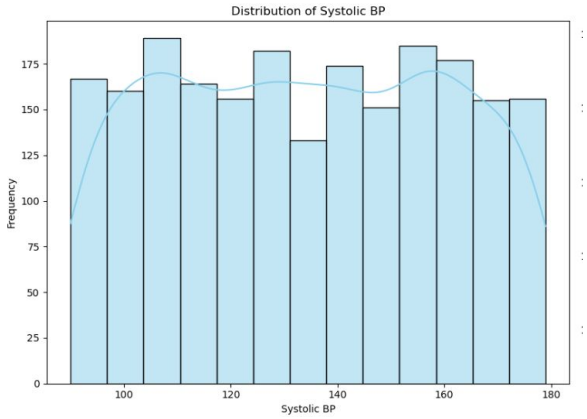


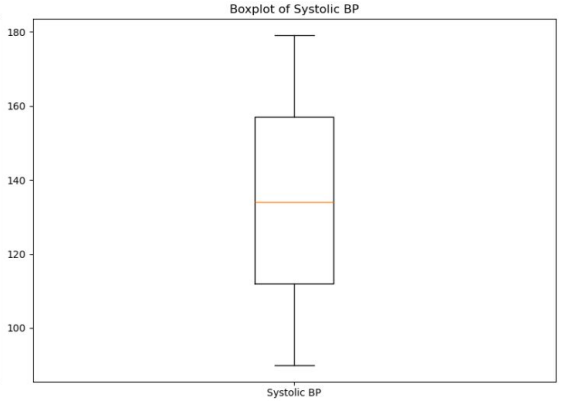
EDA Cont’d

Continuous data

- No normal distributions?
- No outliers?
- No ‘tails’?

? Highly edited data



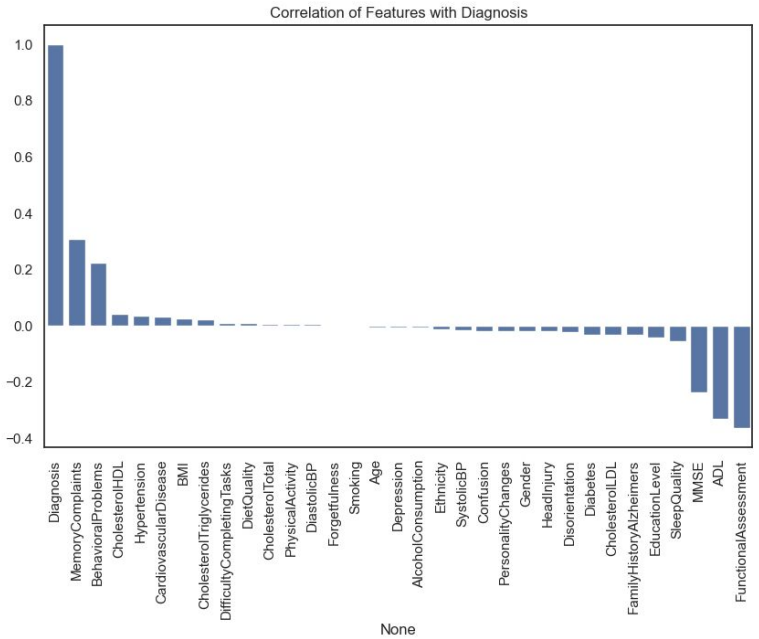


Feature Correlation

- The following are strongly positively correlated:
  - Memory complaints
  - Behavioural problems

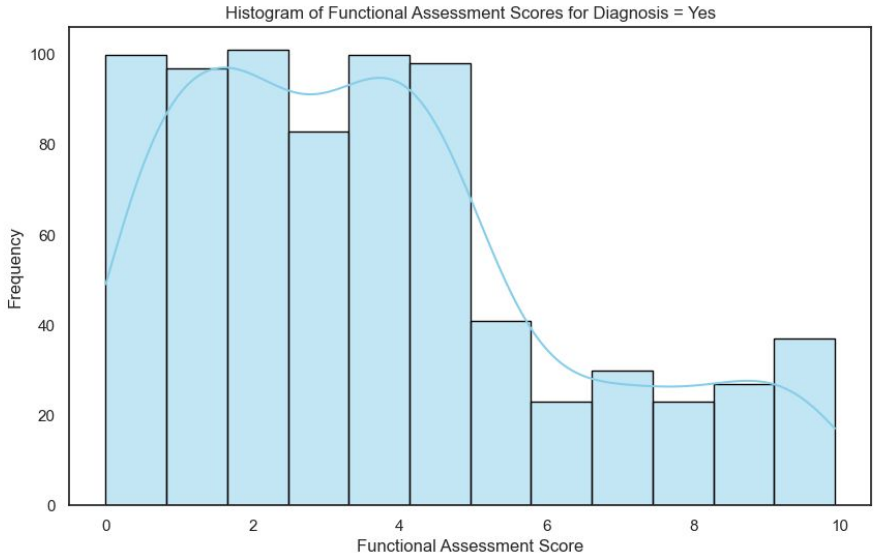
- The following are strongly negatively correlated:
  - Mini-Mental Score Examination (MMSE)
  - Functional Assessment
  - Activities of Daily Living (ADL)

- These results made sense to me as memory complaints and behavioural problems score higher with AD whereas patients with AD will have lower scores for MMSE, functional assessment and ADL



Features

Visualisation of Functional Assessment (+ve diagnosis)



Data Pre-processing

- Data needed to be scaled but the dataset had a lot of binary columns
  - Created ‘binary’ dataset
  - Created a ‘non-binary’ dataset
  - Scaled the ‘non-binary’ using StandardScaler
  - Combined the two datasets

[69]:

	Age	Ethnicity	EducationLevel	BMI	AlcoholConsumption	PhysicalActivity	DietQuality	SleepQuality	SystolicBP	DiastolicBP	...	HeadInjury	Hypertension	MemoryComplaints
0	-0.212368	-0.700408	0.788833	-0.655225	0.565923	0.492525	-1.253593	1.119918	0.298159	-1.014750	...	0	0	0
1	1.567757	-0.700408	-1.422782	-0.114751	-0.954895	0.945093	-1.538442	0.056836	-0.742572	-1.469595	...	0	0	0
2	-0.212368	2.311955	-0.316974	-1.366428	1.653006	1.023896	-1.088855	1.487380	-1.359301	1.486898	...	0	0	0
3	-0.101111	-0.700408	-0.316974	0.851625	0.376930	1.227995	0.839804	0.760833	-0.626935	1.430043	...	0	0	0
4	1.567757	-0.700408	-1.422782	-0.961607	1.461793	0.486696	-1.443293	-0.824566	-1.552029	1.543754	...	0	0	0

Modelling

Set up

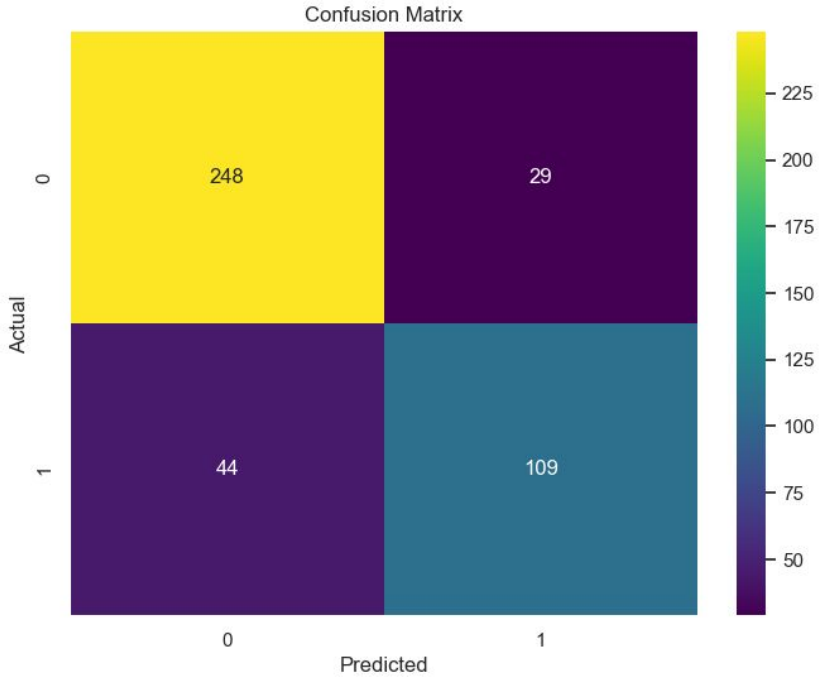
- Used all the features ‘as is’
- Defined the target and predictor features:
  - Target = Diagnosis
  - Predictor features: All other columns
- Performed Train-Test split
  - Test size= 0.20
  - Set Random State for reproducibility
  - Checked shape of train / test arrays
- Created a model for Logistic Regression
- Fit Linear Regression Model

Model Evaluation

Model Testing

- Accuracy with training data = 85%
- Predicted the diagnosis using test data
    - +ve AD: 138 (~32%)
    - -ve AD: 292 (~68%)

Visualised the confusion matrix



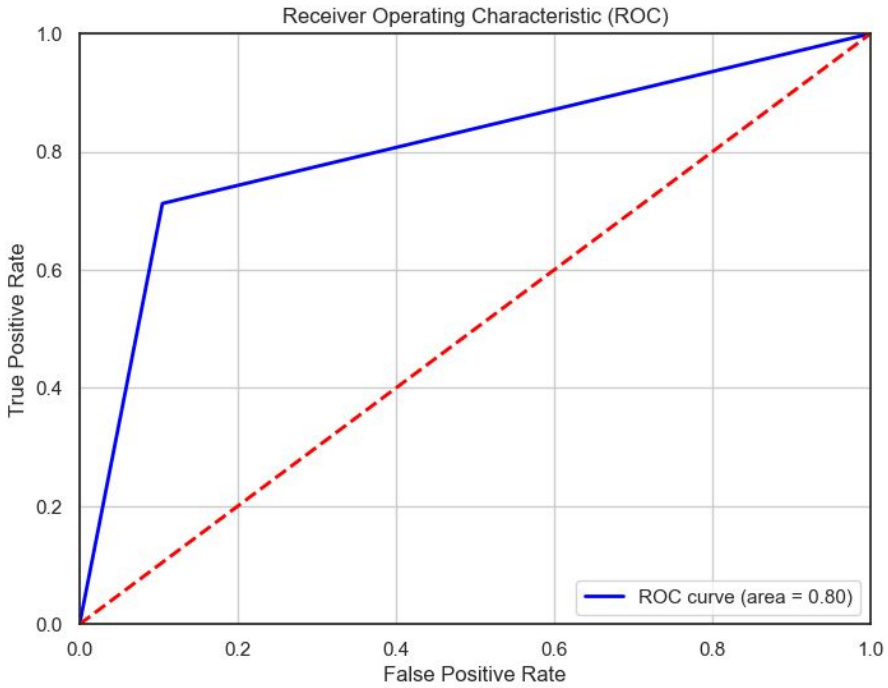
Ran classification report:

- Accuracy = 83%
- Recall: 0.90 (-ve) / 0.71 (+ve)
- F1 score: 0.87 (-ve) / 0.75 (+ve)

Model Evaluation Cont’d

Plotted ROC Curve

- AUC = 0.80



Model Evaluation

Use of Principal Component Analysis

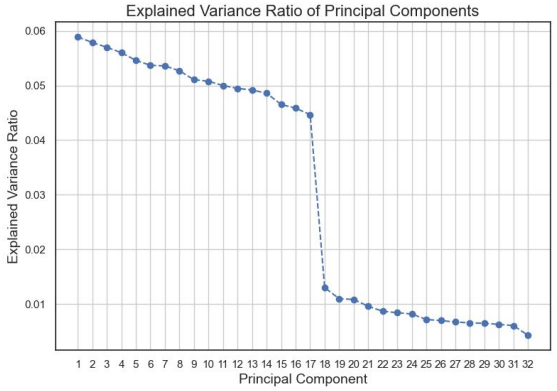
- Do I need all the features?
- Top 5 seemed most important?
- PCA suggested 19 features (90% threshold)
- Ran Logistic Regression with the 19 features
- Evaluated output as before

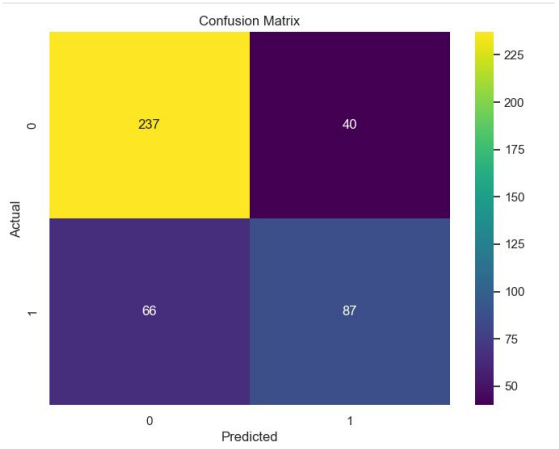
Result

- Accuracy = 75%
- Recall: 0.86 (-ve) / 0.57 (+ve)
- F1 score: 0.82 (-ve) / 0.62 (+ve)
- AUC = 0.71

Conclusion

- Use of PCA did not improve the model
- All features were required to some extent



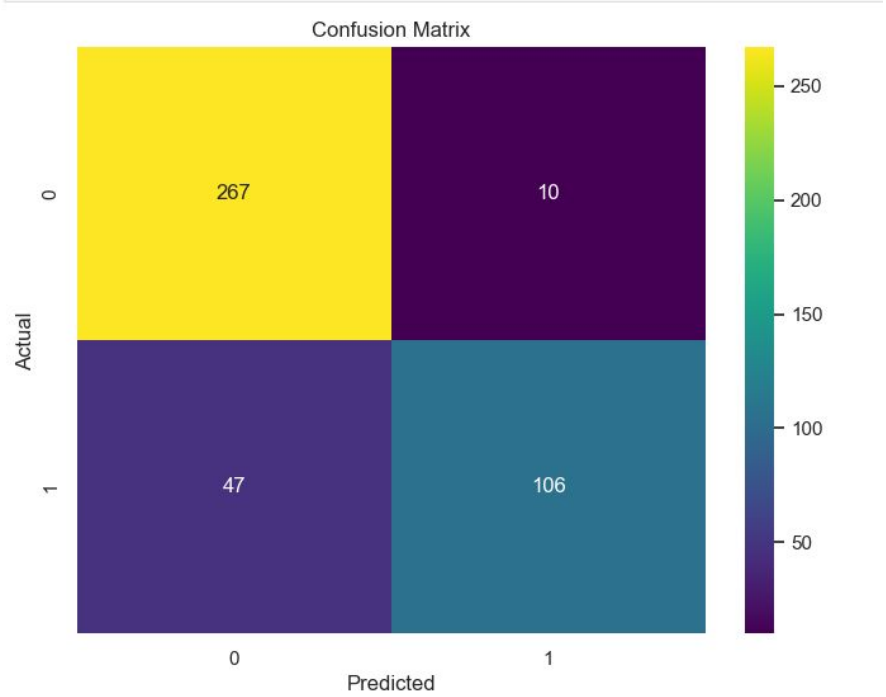


Model Evaluation - Decision Tree

Decision Tree Classifier

- Accuracy with training data = 89%
- Predicted the diagnosis using test data
    - +ve AD: 116 (~27%)
    - -ve AD: 314 (~73%)

Visualised the confusion matrix →



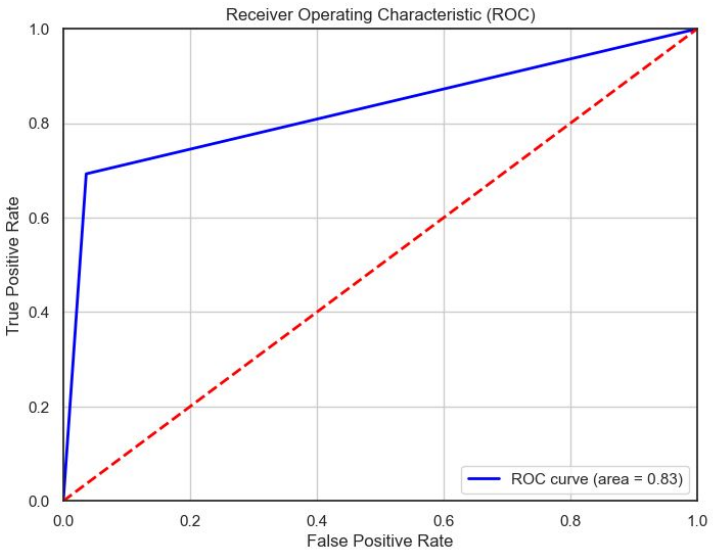
Ran classification report:

- Accuracy = 87%
- Recall: 0.96 (-ve) / 0.69 (+ve)
- F1 score: 0.90 (-ve) / 0.79 (+ve)

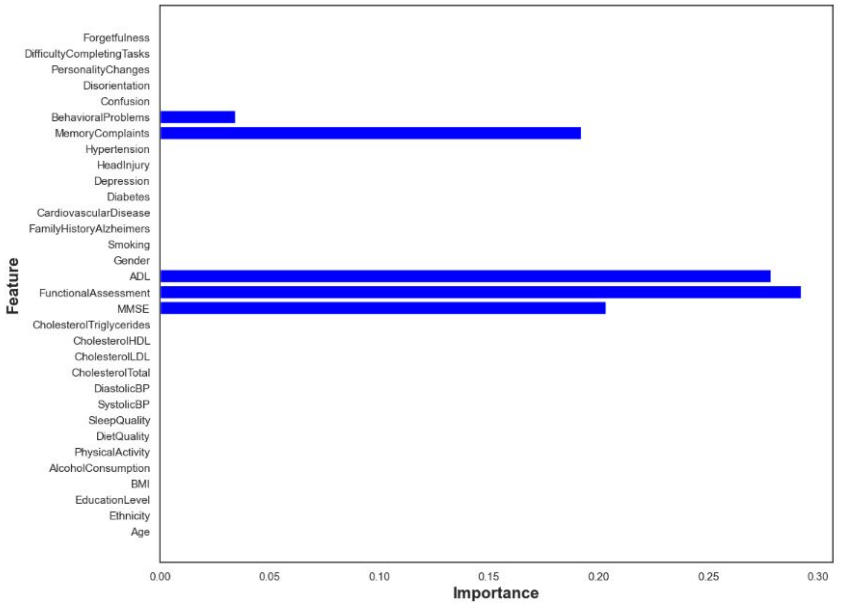
Model Evaluation - Decision Tree Cont'd

Plotted ROC Curve

- AUC = 0.83



Feature Importance

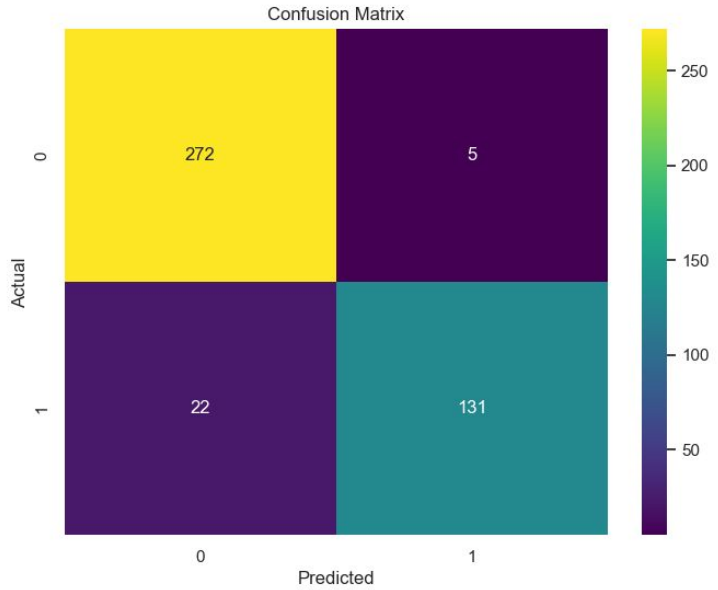


Model Evaluation - Random Forest

Random Forest Classifier

- Accuracy with training data = 100% (!!)
- Predicted the diagnosis using test data
    - +ve AD: (~32%)
    - -ve AD: (~68%)

Visualised the confusion matrix →



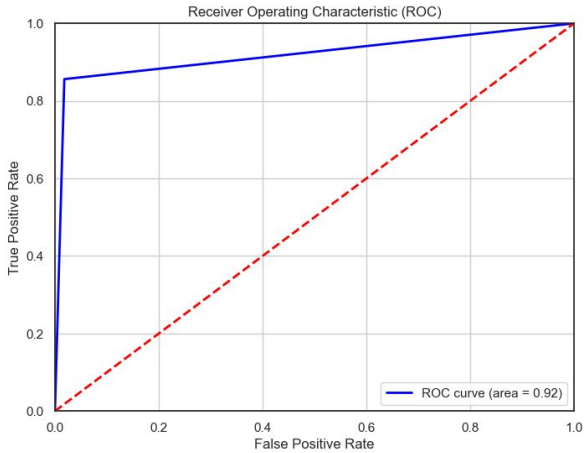
Ran classification report:

- Accuracy = 94%
- Recall: 0.98 (-ve) / 0.86 (+ve)
- F1 score: 0.95 (-ve) / 0.91 (+ve)

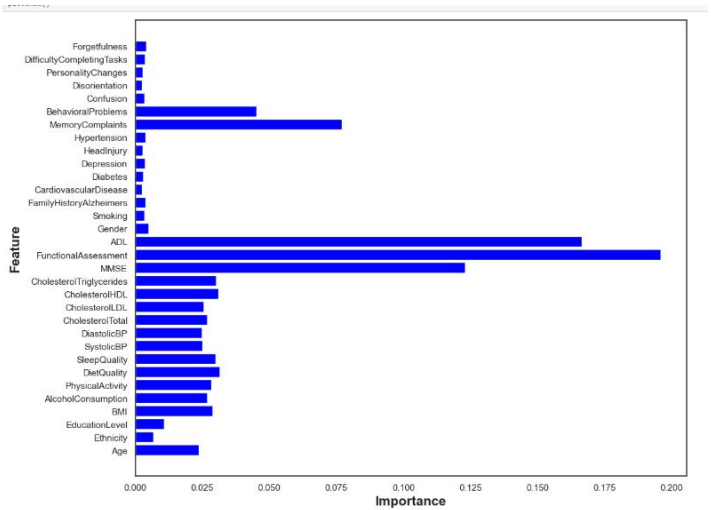
Model Evaluation - Random Forest Cont'd

Plotted ROC Curve

- AUC = 0.92



Feature Importance



Comparisons

Method	Accuracy	Recall (Sensitivity)	F1 Scores*	AUC^
Logistic Regression	Train: 85% Test: 83%	-ve: 0.90 +ve: 0.71	-ve: 0.87 +ve: 0.75	0.80
LR+PCA	Train: 85% Test: 75%	-ve: 0.86 +ve: 0.57	-ve: 0.82 +ve: 0.62	0.71
Decision Tree	Train: 89% Test: 87%	-ve: 0.96 +ve: 0.69	-ve: 0.90 +ve: 0.79	0.83
Random Forest	Train: 100% (!!) Test: 94%	-ve: 0.98 +ve: 0.86	-ve: 0.95 +ve: 0.91	0.92

- *F1 score takes both FPR and FNR into account: more informative than accuracy in this scenario as dataset is imbalanced
- *F1 is also useful when the +ve class is of greater interest e.g medical diagnosis
- ^AUC is less affected by class imbalance (evaluates how well +ve cases are ranked relative to -ve cases, irrespective of class distribution

Conclusions

Headlines:

- Random Forest Classification performed best overall
  - It generalised well to unseen data
  - Good F1 Scores and Recall scores for +ve data
  - Good AUC score
  - Model exceeded the accuracy target of 85%

Other Points to Consider:

- Need a bigger data set that has not been extensively edited
- Other classification methods could be considered e.g. KNN, SVM
- Hyperparameter tuning
- Use GridSearch
- Dataset is somewhat imbalanced, could use weighted classes

Finally:

- Use this model with caution due to dataset size and edited data - however it is likely not based on a true reflection of real world data

Questions?

References

https://www.kaggle.com/datasets/ilysha/alzheimers-disease-dataset