MODEL X

Optimization Sprint Report

FalconCode

Name	University	NIC
Vibodha Lakshan	University of Moratuwa	200404011049
Imandi Perera	University of Moratuwa	200476001229
Chamod Madubashana	University of Moratuwa	200428904643
Dinuka Dasanayaka	University of Moratuwa	200409210216

01.Dataset Details

The available dataset represents a longitudinal study clinic dataset, considering a total of 195,196 visits made by 52,537 individuals. The biggest challenge associated with this particular dataset includes the fact that it has a significant number of individuals, which are 35,671, contributing more than one visit.

The variable of interest is DEMENTED, a binary flag that indicates whether the participant has been diagnosed as having dementia when the visit occurs.

Process Flow

The project involved the following structured machine learning workflow:

1. **Data Loading & Exploration:** The raw data in the form of a CSV file was imported, and the distribution of the target variable as well as participant IDs was explored.
2. **Feature Selection:** Non-medical variables were chosen from Form A1 (Demographics), Form A5 (Self-reported Health), and Form A3 (Family History), as defined in the Data Dictionary, based upon the hackathon conditions.
3. **Feature Engineering:** The raw features were cleaned (where special values like 99, 888 were taken care of), along with the addition of new, more predictive features like IS_MULTIRACIAL, BMI, and one-hot encoded variables.
4. **Training & Test Data Splits**: A GroupShuffleSplit was applied, using NACCID as the grouping variable. This was a crucial step in the process, as it allowed the data from each individual to remain in the same splits. This prevents potential data leakage.
5. **Model Development & Tuning**: Four powerful models based on decision trees were trained: RandomForest, XGBoost, CatBoost, and

LightGBM. The hyperparameters were tuned in each model using GridSearchCV or RandomizedSearchCV.

6. **Ensemble Modeling**: A final Stacking Ensemble was constructed, where a LogisticRegression meta-model was used to integrate the predictions made by the other four base models.
7. **Evaluation & Explainability**: The models were evaluated on the test set using the ROC-AUC metric. The model was then used to explain the results using the SHAP technique.

02.Feature Engineering

1. Selected Non-Medical Features

In light of the clarification from the hackathon, the features considered were those that a person can "know about themselves" – demographics, lifestyles, as well as straightforward known diagnoses.

- **Form A1 (Demographics):**
  - Features: NACCAGE, SEX, EDUC, RACE, MARISTAT, RESIDENC, NACCLIVS, INDEPEND, etc.
  - Justification: These are common demographic variables and social context factors, which are also specifically allowed as non-medical variables.
- **Form A5 (Self-Reported Health):**
  - Features: CVHATT (Heart Attack), CBSTROKE (Stroke), DIABETES, HYPERTEN, TOBAC100 (Smoking), THYROID, B12DEF, etc.
  - Justification: These are "simple diagnoses" that a person can understand and self-report, as has been clarified in the hackathon. These are very effective non-medical predictors.

- **Form A3 (Family History):**
    - Features: NACCFAM, NACCMOM, NACCDAD.
    - Justification: These were listed as the "borderline" criteria. These reflect information that the individual would know (it’s not a genetic finding) but is important in terms of the non-medical risk profile.

2. Feature Creation

New elements were designed in order to integrate information so it could be more useful for the models.

- **IS_NON_ENGLISH_HOME:** This binary flag is derived from PRIMLANG.
- **IS_MULTIRACIAL:** This binary flag was obtained through the process of checking if there were any entries in RACESEC or RACETER.
- **RACE_OTHER:** Binary indicator for the 'Other' race variable.
- **BMI:** This was calculated from the variables HEIGHT and WEIGHT, where NACCBMI was lacking, giving it a much cleaner and more complete feature.
- **One-Hot Encoding:** The pd.get_dummies function was applied to the variables MARISTAT, RESIDENC, and RACE to encode the variables in a numerical form that could be processed by the models.
- **A5/A3 Flags:** The A5/A3 flags were translated to numerical flags with values 0, 1, or 2, as well as NaN using the "a5_to_flag” helper, flagging “No” as 0, “Recent/Active” as 1, and “Remote/Inactive” as 2.
-

3. Finalized Features

The final set of features X included the cleaned set of A1 features, the newly engineered features, as well as all the one-hot columns and the flags in A5/A3. This created a sparse data set that was ideal for the use of tree-based models.

03.Data Preprocessing

Preprocessing Steps

1. **Handling of Special Codes:** The helper functions, clean_educ, were applied to treat non-informative codes (for example, code '99'-unknown education) as np.nan.
2. **Handling of Missing Values**: The NaN values were designed to be passed without any handling. This was because the models chosen (namely XGBoost, CatBoost, LightGBM) were designed to handle this issue, which was more effective than the mean/median approach. The RandomForest model was specified to handle the class_weight="balanced" parameter.
3. **Categorical Conversion:** This was achieved, as seen in the Feature Engineering process, by using the get_dummies function for one-hot encoding.

Train-Test Split

- **Method:** GroupShuffleSplit (70% for train, 15% for validation, 15% for test).
- **Justification:** This was the most important piece of the preprocessing work. In the dataset, there are several rows that correspond to the same individual. This means that if a standard train_test_split were performed, it would result in data leakage, where the model would train on a person's Visit 1 and test on their Visit 5, which wouldn't make the result meaningful.
- Use of the GroupShuffleSplit splitter with the groups parameter set to NACCID ensures that all data records for the same participant are contained in one split. This allows us to ensure that our model is

evaluated based upon the prediction of risk in individuals it has not seen before, which is the whole point of this project.

04.Model Building

Models Trained

Various powerful models based on the concept of a tree were used for training. In this regard, the models chosen are as follows:

1. **RandomForestClassifier:** A strong, parallelizable ensemble model that serves as a high-quality baseline.
2. **XGBClassifier (XGBoost):** This is the industry-standard GBT algorithm, which was chosen for its accuracy and speed.
3. **CatBoostClassifier:** This GBT algorithm is quite modern, performs well when the variables are categorical (though in this scenario, they were one-hot encoded).
4. **LGBMClassifier (LightGBM):** This algorithm is a GBT model that's fast, efficient, and suitable for use on large datasets.
5. **Stacking Ensemble:** The final meta-model utilizes a LogisticRegression to combine the predictions of the four base models.

Hyperparameter Tuning

- **RandomForest:** The parameters were tuned using GridSearchCV on a small grid defined by parameters like n_estimators, max_depth, min_samples_split, etc.
- **XGBoost, CatBoost, LightGBM:** The strategy RandomizedSearchCV was applied with the parameter n_iter=40 in order to search the parameter space in a more intensive manner, which could be applied when there were many parameters.

05.Model Evaluation

Evaluation Metrics

- **Primary Metric: ROC-AUC (Area Under the Receiver Operating Characteristic Curve)**
  - Justification: The variable DEMENTED, the target variable, has a class imbalance problem. The use of accuracy as the evaluation metric in this case can be very misleading, as a model predicting “Not Demented” every time would result in a model having a high accuracy but little use. ROC-AUC evaluation should be used in this case. A perfect classifier would attain a value of 1.0, while a random classifier would attain a value of 0.5.
- **Secondary Metric: Accuracy**
  - Justification: Added to provide completeness, but not used in model selection.

06.Model Comparison

Comparison between the models was based on the ROC-AUC score obtained from the held-out test set.

Model	Validation AUC	Test AUC
Random Forest	0.9344	0.9334
XGBoost	0.9369	0.9353
CatBoost	0.9377	0.9367
LightGBM	0.9382	0.9362
Stacking Ensemble	0.9383	0.937

07.Final Model

The Stacking Ensemble was chosen as the final model.

- **Justification:** From the table, the stacking model performed the best based on the ROC-AUC scores in the validation set as well as the test set. The meta-model, which utilizes the prediction of the base models, can leverage the strength of the base models, which are diverse, to generate a better prediction result.

08.Explainability & Model Interpretability

Techniques Used

In analyzing the learnings of the model, the tool used was called SHAP (SHapley Additive exPlanations). This tool gives each feature an "importance value" for every single prediction made, enabling us to understand the model outputs. The tool, named SHAP TreeExplainer, was applied to the best single model, which was XGBoost.

Insights Gained

The summary plots obtained from the SHAP tool revealed the core variables responsible for the prediction of the risk of dementia based on the model:

1. **AGE :** This was, rather convincingly, the most important piece of information in the model. The clear use of the SHAP values in the visualization illustrates the tremendous positive impact of the factor 'age' upon the model output (as indicated by the red dots).
2. **EDUC_YEARS (Years of Education):** This was the second most important variable. From the graph, it can be seen that there was a very strong negative relationship, as higher education (red dots) had a negative SHAP value, meaning that this variable contributed to reducing the predicted risk.

3. **INDEPEND (Level of Independence):** This was also a very important factor. A lower level of independence (as in “requires assistance”) greatly increased the probability.
4. **Medical History:** Self-reported health variables such as CBSTROKE_FLAG (Stroke), DIABETES_FLAG, and CVHATT_FLAG (Heart Attack) were all important predictors, validating the relevance of these well-known non-medical variables.
5. **Social Factors:** MARISTAT (Marital Status) as well as NACCLIVS (Living Situation) were also used in the prediction equations, showing the model learned from the participant's social context.

09.Tools Used

- Language: Python
- Core Libraries: Pandas, NumPy
- ML & Splitting: Scikit-learn (GroupShuffleSplit, GridSearchCV, RandomizedSearchCV, LogisticRegression, RandomForestClassifier)
- Boosting Models: XGBoost, CatBoost, LightGBM
- Explainability: SHAP
- Plotting: Matplotlib
- Environment: Jupyter Notebook

10.GitHub Repo Link -

https://github.com/vibodhalakshan2004/Dementia-Prediction-Model-Notebook