Predicting Age from National Health and Nutrition Health Survey 2013-2014

Ismail (Husain) Bhinderwala, Rashid Mammadov, Sienko Ikhabi, Dongchun Chen

2024-12-07

Table of contents

1 Summary

In this project, we developed a logistic regression model to classify individuals into two age groups: Senior (65 years and older) and Adult (under 65 years), using various attributes about the individual. The data is from the National Health and Nutrition Examination Survey (Health Statistics (NCHS) at the Centers for Disease Control and (CDC) 2019). The model we developed uses features such as physical and health-related measurements to make predictions.

We achieved an overall accuracy of 0.73 and an F1 score of 0.61 on the test data. It correctly classified 418 cases; 370 Adults and 48 Seniors, but misclassified 152 cases.

2 Introduction

Age classification is an important aspect of demographic analysis and health resource planning, particularly when distinguishing between seniors (65 years and older) and non-seniors (under 65 years). Accurate age group identification allows for better-targeted healthcare strategies and more efficient resource allocation. Traditional methods often rely on broad assumptions, which can miss important individual differences. In this project, we investigate whether a machine learning model can classify individuals into these two age groups using physical and health-related measurements.

We used data from the National Health and Nutrition Examination Survey (NHANES) 2013-2014 (Health Statistics (NCHS) at the Centers for Disease Control and (CDC) (2019)) to develop a logistic regression model for this purpose. Accurately identifying seniors is especially important, as they are more likely to require regular medical care and management of chronic conditions (Löckenhoff et al. (2016)). By enhancing the precision of age classification, this model has the potential to improve healthcare planning and ensure that interventions are better aligned with the needs of different age groups. This study demonstrates how machine learning can be used to tackle real-world challenges in public health and demographic research.

3 Methods and Analysis

3.1 Data Preprocessing and EDA

The dataset used for this analysis is the National Health and Nutrition Health Survey 2013-2014 (Health Statistics (NCHS) at the Centers for Disease Control and (CDC) (2019)) Age Prediction Subset. It was obtained from the UCI Machine Learning Repository in ZIP format and extracted for preprocessing. The dataset contains 2,278 entries and 8 features described in Table 1 below.

Table 1: Description of the columns in the Age Prediction dataset

Feature	Data Type	Description
SEQN	Numeric	Respondent Sequence Number.
age_group	Categorical	Respondent's Age Group. Encoding [`Senior`, `Adult`]. `Adult` is also referred to as non- senior. This is the target variable.
RIDAGEYR	Numeric	Respondent's age, in years.

Table 1: Description of the columns in the Age Prediction dataset

Feature	Data Type	Description
RIAGENDR	Numeric	Respondent's gender. Encoding [1: `Male`, 2: `Female`]
PAQ605	Numeric	If the respondent engages in moderate or vigorous-intensity sports, fitness, or recreational activities in the typical week. Encoding [1: `Yes`, 2: `No`]
BMXBMI	Numeric	Respondent's Body Mass Index (BMI)
LBXGLU	Numeric	Respondent's blood glucose measurement after fasting
DIQ010	Numeric	If the Respondent is diabetic. Encoding [1: `Yes`, 2: `No`, 3: `Borderline`]
LBXGLT	Numeric	Respondent's oral glucose tolerance (OGTT) result
LBXIN	Numeric	Respondent's blood insulin levels

We noted that target variable exhibited an imbalance with two classes: Adult (84%) and Senior (16%).

The data preprocessing involved several steps. Unnecessary columns such as id and age were removed, focusing on features directly relevant to the prediction task. Categorical variables were converted into more interpretable forms: gender was mapped from numeric values (1 for Male, 2 for Female), weekly_physical_activity was mapped to "Yes" and "No," and diabetic was mapped to "Yes," "No," and "Borderline". Additionally, a single erroneous row with the value 7.0 in the weekly_physical_activity column was removed as it did not align with the dataset's binary format. Descriptive statistics for numerical variables (bmi, blood_glucose_fasting, oral, insulin_level) were calculated to summarize their distributions. The preprocessing was implemented using Python libraries such as NumPy and Pandas (Harris et al. (2020); McKinney et al. (2010)).

Exploratory data analysis (EDA) was performed to better understand the dataset. For numerical features, a correlation analysis was conducted, and the relationships between features were visualized using a heatmap created with Matplotlib (Barrett et al. (2005)). The distribution of categorical features was examined, highlighting imbalances in age_group, gender, diabetic, and weekly_physical_activity. For instance, the diabetic column revealed that most entries were labeled as "No," with only 21 labeled "Yes" and 58 labeled "Borderline."

To address the class imbalance in the target variable, stratified sampling was employed during the train-test split. This ensured that the proportions of the age_group classes were preserved in both training (75%) and testing (25%) datasets. A random seed of 522 was used to make the results reproducible. The processed datasets were exported as CSV files to a structured directory for further analysis and modeling (Pedregosa et al. (2011)).

Once we explored the features in the dataset, we started considering how we would build a classification model using the data. To give us a sense on how each feature is related to the

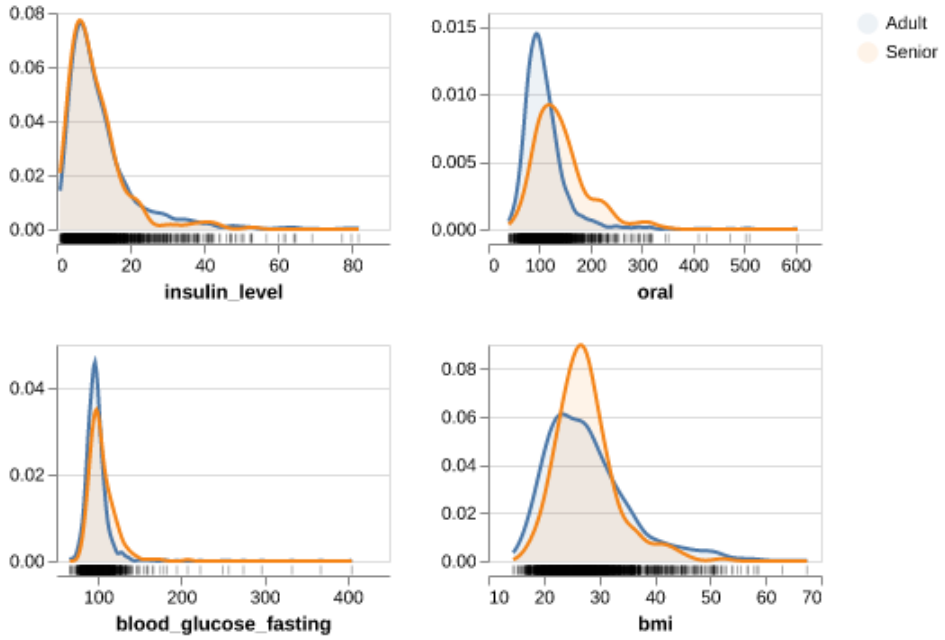

Figure 1: Distribution of Numeric Features by Age Group

target class, we prepared a correlation matrix below that shows the correlation of the input features as a heatmap:

3.2 Model Definition

We chose the logistic regression modeling approach because of two principal reasons:-

- In addition to a hard prediction, we would get a probability value which will be useful for additional interpretation
- Using the learned coefficients we will be able to easily interpret the model and determine feature importance

3.3 Pre-processing

We chose to do pre-processing on the input data as follows:-

- We did not have to do any imputation of values because the data set does not contain any missing values
- We used a `StandardScaler` for the numeric columns `bmi`, `blood_glucose_fasting`, `oral` and `insulin_level`

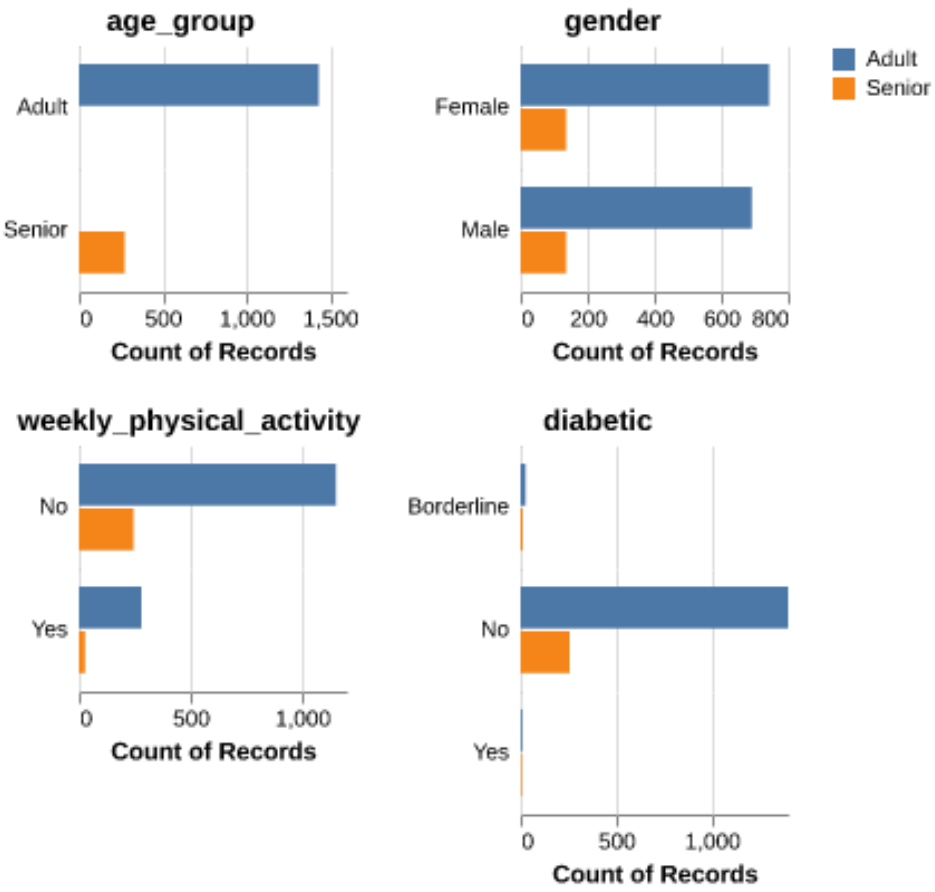

Figure 2: Distribution of Categorical Features by Age Group

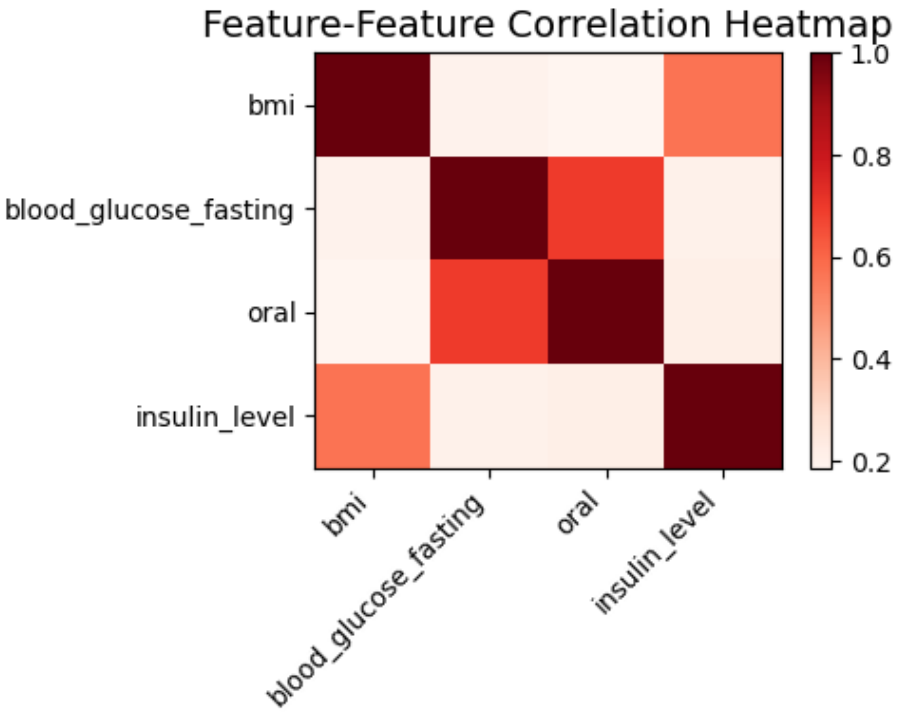

Figure 3: Feature-to-Feature Correlation Heatmap

- The column `diabetic` is categorical, but contained ordered levels of a subject being diabetic (the spectrum being `No` to `Borderline` and finally `Yes` for diabetic subjects). We therefore used a `OrdinalEncoder` for this column
- The other categorical columns, `weekly_physical_activity` and `gender`, contain nominal values and so we applied a `OneHotEncoder`

3.4 Hyperparameter Optimization

We had only one hyperparameter to optimize for our logistic regression model. To obtain the best value for the hyperparameter `C`, we used `GridSearchCV`. Our best value was $C = 0.0001$. Based on our exploratory data analysis, we knew that we had a class imbalance and chose to set `class_weight = 'balanced'`.

In future iterations, we will explore manually adjusting the class weights to see if it would improve the model's performance.

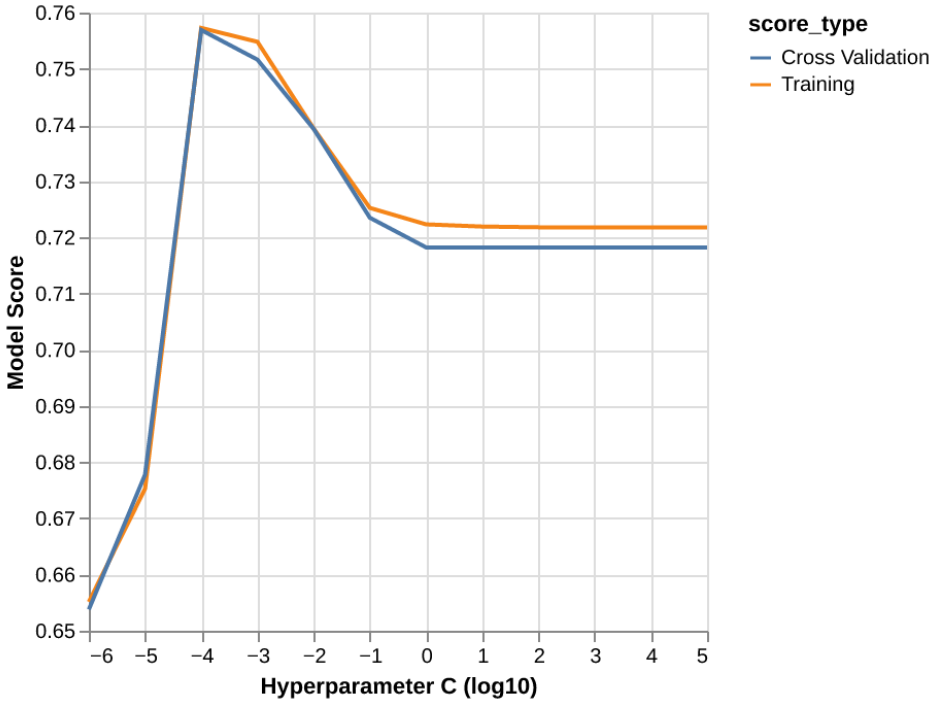

Figure 4: Optimization of model hyperparameter C

3.5 Model Fitting

Once we obtained the best values for the hyperparameter `C`, we defined a logistic regression model pipeline with the column transformers described above, and a logistic regression object initialized with the best hyperparameter values. We then fitted this estimator on the entire training dataset.

4 Results

4.1 Model Evaluation

Our prediction model performed averagely on test data. The classification metrics show a Macro Average F1 score of 0.61 and a decent accuracy of 0.73. The detailed results are presented in Table 2 below.

Table 2: Performance metrics on test data

	precision	recall	f1-score	support
Adult	0.90	0.77	0.83	479
Senior	0.31	0.53	0.39	91
accuracy	0.73	0.73	0.73	1
macro avg	0.60	0.65	0.61	570
weighted avg	0.80	0.73	0.76	570

Looking at the results of the classification for individual classes as shown in Table 3 below, more than half of Senior and Adult categories were correctly classified.

Table 3: Confusion matrix of model performance on test data

Actual label:	Predicted: Adult	Predicted: Senior
Adult	370	109
Senior	43	48

The results in Table 3 can certainly do with improvement, which we aim to do in the next iteration of building this model. We equally plan to engineer additional features, with the new knowledge we acquired from our recent training in Feature Engineering.

5 Discussion and conclusion

This was our very first attempt at fitting a classification model to this data set and so the performance is in line with our expectations. There is certainly more that we will explore to improve this model. As a start, we aim to include some new engineered features to improve the performance of this Logistic Regression model. In addition, we will review the distribution of the probability values for each class to explore whether we can adjust the default threshold of 50% to capture more cases close to the boundary.

Early on we decided to go straight to use a Logistic Regression model. We are aware that we have additional machine learning model types that we can explore and then compare those results with the current results as a baseline. We plan to evaluate K-Nearest Neighbor, and SVC RBF Classifier and Naive Bayes in the next few weeks.

This model is a good baseline that allows us to continue researching these additional questions.

References

Barrett, Paul, John Hunter, J Todd Miller, J-C Hsu, and Perry Greenfield. 2005. "Matplotlib–a Portable Python Plotting Package." In *Astronomical Data Analysis Software and Systems XIV*, 347:91.

Harris, Charles R, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62.

Health Statistics (NCHS) at the Centers for Disease Control, National Center for, and Prevention (CDC). 2019. "National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset." UCI Machine Learning Repository.

Löckenhoff, Corinna E, Chu Hsiao, Julia Kim, and Katya Swarts. 2016. "Adult Age Differences in Health-Related Decision-Making: A Primer." *Handbook of Health Decision Science*, 145–55.

McKinney, Wes et al. 2010. "Data Structures for Statistical Computing in Python." In *SciPy*, 445:51–56. 1.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python Journal of Machine Learning Research." *Journal of Machine Learning Research* 12: 2825–30.