Data Quality and Data Wrangling

PROJECT

Task 3: Tidy Up Your Data

https://github.com/CemOguz/Data-Quality-and-Data-Wrangling

Tutor: Sahar Qaadan

Student: Oguz, Cem

Matriculation Number: 32008124 **Course: DLBDSDQDW01**

Table of Contents

List of Figures and Tables

1. Introduction

Stroke, as the fifth leading cause of death in the United States, poses a significant health challenge with severe consequences for survivors. Predicting the underlying risk factors for stroke is crucial for early screening and prevention. The Health and Nutrition Examination Survey (NHANES) data from the National Center for Health Statistics offers a diverse range of variables, encompassing demographics, medical history, physical examinations, biochemistry, dietary, and lifestyle information. The project focuses on data cleaning, addressing the imbalance in the dataset, tidying it up and feature selection from over 1,000 variables of this dataset.

Key steps include the preparation of datasets from various sources, exploration of variables, data cleaning, and feature selection based on correlation. The project then employs models such as XGBoost, and SMOTE and concludes with the productionization of models, saving them for potential deployment in real-world scenarios to contribute to upcoming scientific researches too.

Through this endeavor, the project aims to contribute to the understanding of stroke risk factors, showcasing the potential of machine learning in healthcare for proactive intervention and improved patient outcomes.

In the realm of healthcare, the predictive power of Machine Learning takes center stage, offering a promising avenue for addressing critical challenges. The National Health and Nutrition Examination Survey (NHANES) emerges as a exceptional source of data, encompassing a vast number of variables ranging from demographics to medical history. Much like urban planners utilize diverse data sources to assess air quality, our objective is to find out the patterns and predictors of stroke within this extensive dataset.

As we embark on this journey, the project unfolds through critical stages, including meticulous data cleaning, addressing imbalances in the dataset, and strategically selecting features relevant to stroke risk.

1.1 Research Questions:

Feature Selection**:** What features, including but not limited to blood pressure, cholesterol levels, and lifestyle factors, are most influential in predicting stroke risk?

Imbalanced Dataset Handling**:** How can we effectively address the challenge of imbalanced datasets to enhance the accuracy of stroke prediction models?

Through this project, we aim to contribute valuable knowledge to the field, fostering a deeper understanding of the intricate factors influencing stroke risk. In doing so, Machine Learning becomes a

beacon, illuminating the path toward improved preventive healthcare measures and enhanced patient outcomes, but this can only be done by ensuring data quality which is the primary focus of the project.

1.2 Background and Literature Review

The National Health and Nutrition Examination Survey (NHANES) is a comprehensive program aimed at evaluating the health and nutritional status of both adults and children in the United States. This initiative, overseen by the National Center for Health Statistics (NCHS), a division of the Centers for Disease Control and Prevention (CDC), employs a distinctive approach by integrating interviews and physical examinations.

Commencing in the early 1960s, the NHANES program has evolved into a continuous series of surveys, with a shifting focus on diverse population groups and health-related subjects. The survey, now a pivotal component of NCHS's mission, addresses emerging health and nutrition needs through a representative sample of approximately 5,000 individuals annually. This sample is dispersed across 15 counties visited each year.

The NHANES interview encompasses a wide array of inquiries, spanning demographics, socioeconomic factors, dietary habits, and health-related issues. The examination component involves a meticulous assessment conducted by highly trained medical personnel, encompassing medical, dental, and physiological measurements, alongside laboratory tests.

Over the years, the NHANES program has facilitated the publication of numerous research findings, utilizing its extensive data to contribute valuable insights into various aspects of health and nutrition.

2. Data Retrieval and Cleaning

The original dataset presented a challenge due to disorganized and unclear variable names. Recognizing the importance of clarity and precision in the analysis, a comprehensive scraping process was initiated on the NHANES website. The objective was to carefully gather accurate and detailed descriptions for each variable included in the dataset. This approach was crucial for enhancing the clarity of variable meanings and promoting a deeper understanding of the dataset.

By obtaining precise variable descriptions through web scraping, we aimed to eliminate ambiguity and ensure a clear understanding of each variable's significance in the context of stroke prediction.

This endeavor not only showcases the commitment to data quality, but also emphasizes the significance of using well-documented and comprehensible variable names in healthcare research. The resulting dataset with enriched variable descriptions serves as a solid foundation for further processes, contributing to the overall reliability and interpretability of the study findings.

2.1 Variable Descriptions

An intricate process was employed to streamline the dataset, reflecting a reasonable approach to enhance its quality and suitability. The initial steps involved excluding non-numeric columns and those with over 50% missing values, resulting in a refined dataset comprising 153 columns. This cleaning process aimed to eliminate noise and focus on relevant, high-quality data, laying the groundwork for the entire project.

Furthermore, specific attention was given to the target variable MCQ160F. Its coding was strategically adjusted to create a binary representation (0 for negative and 1 for positive), aligning with the conventions of binary classification tasks. This adjustment not only simplified the interpretation of model outcomes but also optimized the dataset for compatibility with machine learning algorithms.

The NHANES dataset, designed to capture the intricacies of health and nutrition, underwent a transformation through these cleaning steps. The decision to omit non-numeric columns and address missing values underscored a commitment to data quality and precision. The resulting dataset, with its refined structure and precisely coded target variable, stands as a testament to the efforts given in preparing the data for meaningful analysis.

Moreover, the discussion acknowledges the inherent complexity of the NHANES dataset, structured to accommodate the multifaceted aspects of health and nutrition.

2.2 Target Variable

The crux of this analysis centers around the MCQ160F variable, a pivotal indicator of whether an individual has been diagnosed with a stroke by a doctor or health professional. To ensure the precision and reliability of the analysis, a process was undertaken to manage the entries associated with this variable. Entries containing null values or responses indicative of uncertainty or refusal were systematically excluded, culminating in a dataset that encompasses 5583 valid entries.

By rigorously excluding entries with incomplete or ambiguous information related to stroke diagnoses, the dataset's integrity and relevance were preserved. This strategic data-cleaning step not

only contributed to the precision of the subsequent analysis but also underscored the commitment to working with a high-quality dataset. The resulting dataset, focused on individuals with clear and complete information regarding stroke diagnoses, forms the foundation of the project.

2.3 Data Collection Methods

The NHANES interview component includes questions related to demographics, socioeconomic status, dietary habits, and various health-related factors. The examination component involves precise measurements conducted by highly trained medical personnel, covering medical, dental, and physiological aspects, along with laboratory tests.

The NHANES dataset has been widely utilized in research, leading to the publication of numerous findings that contribute significantly to our understanding of diverse health and nutrition phenomena. The accurate documentation of variable descriptions ensures the robustness and reliability of the data for the present study.

	variable	label
0	SEQN	Respondent sequence number
1	SDDSRVYR	Data release cycle
2	RIDSTATR	Interview/Examination status
3	RIAGENDR	Gender
4	RIDAGEYR	Age in years at screening
...	...	...
3846	WHD140	Self-reported greatest weight (pounds)
3847	WHQ150	Age when heaviest weight
3848	WHQ030M	How do you consider your weight
3849	WHQ500	Trying to do about weight
3850	WHQ520	How often tried to lose weight

Table 1 - Section of scraped variable names

3. Methods

3.1 Data Preprocessing and Feature Selection

The initial phase of the study involved comprehensive data preprocessing to ensure the dataset's readiness for the model. Diverse NHANES datasets, encompassing demographic, examination, questionnaire, laboratory, and dietary data were systematically merged. Duplicate entries were also excluded based on the respondent sequence number (SEQN), and the target variable (MCQ160F), representing stroke diagnosis, was meticulously defined. (as can be observed in notebook file in the github project)

In order to enhance variable understanding, codebook featuring variable descriptions and SAS labels were integrated into the analysis. This auxiliary resource provided valuable context for interpreting each variable in the dataset.

	variable	score	label
0	OHX13TC	0.435509	Tooth Count: #13
1	OHX20TC	0.046441	Tooth Count: #20
2	OHX22TC	0.042052	Tooth Count: #22
3	OHDEXSTS	0.035350	Overall Oral Health Exam Status
4	OHDDESTS	0.034990	Dentition Status Code
5	DMDHHSZE	0.028796	# of adults 60 years or older in HH
6	OHX23TC	0.023711	Tooth Count: #23
7	OHX31TC	0.021407	Tooth Count: #31
8	MGQ100	0.016507	Recent pain/aching/stiffness-left hand
9	OHX12TC	0.014648	Tooth Count: #12
10	DMDBORN4	0.014265	Country of birth
11	RIDRETH3	0.010694	Race/Hispanic origin w/ NH Asian
12	CSXQUIPT	0.009009	Tongue Tip 1mM Quinine: What Was Taste?
13	SIAPROXY	0.008897	Proxy used in SP Interview?
14	OHX24TC	0.008300	Tooth Count: #24
15	CSAEFFRT	0.008277	Participant's Understanding of Test
16	OHX25TC	0.008208	Tooth Count: #25
17	OHX14TC	0.007667	Tooth Count: #14
18	OHXIMP	0.007340	Dental Implant: yes / no?
19	CSXONOD	0.007114	Smell Test: Onion Scent
20	BMDSTATS	0.006779	Body Measures Component Status Code
21	OHX06TC	0.006364	Tooth Count: #6
22	RIDEXMON	0.006359	Six month time period
23	OHX10TC	0.006196	Tooth Count: #10

Table 2 – Relevance of Features

To streamline the dataset for model training, non-numeric columns and those with over 50% missing values were excluded. The resultant dataset was further refined by balancing the representation of the target variable as the data is highly imbalanced as we can see in below figure 1.

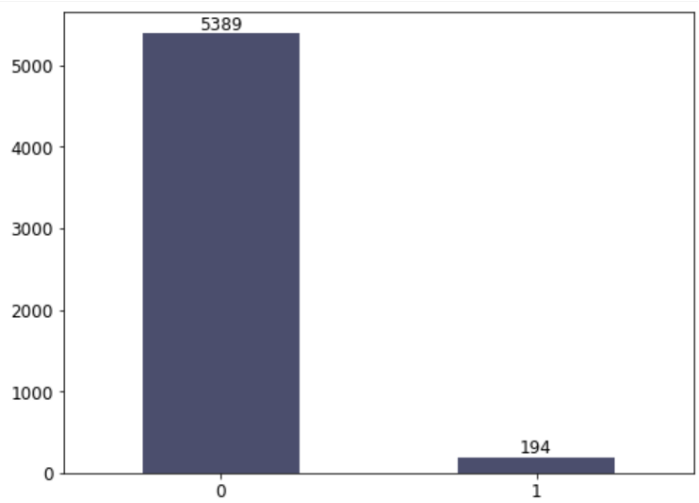

Figure 1 – Target Variable Distribution

And feature selection was conducted using an XGBoost classifier. This classifier identified the top 24 features based on their importance scores, providing valuable insights into the variables influencing stroke prediction.

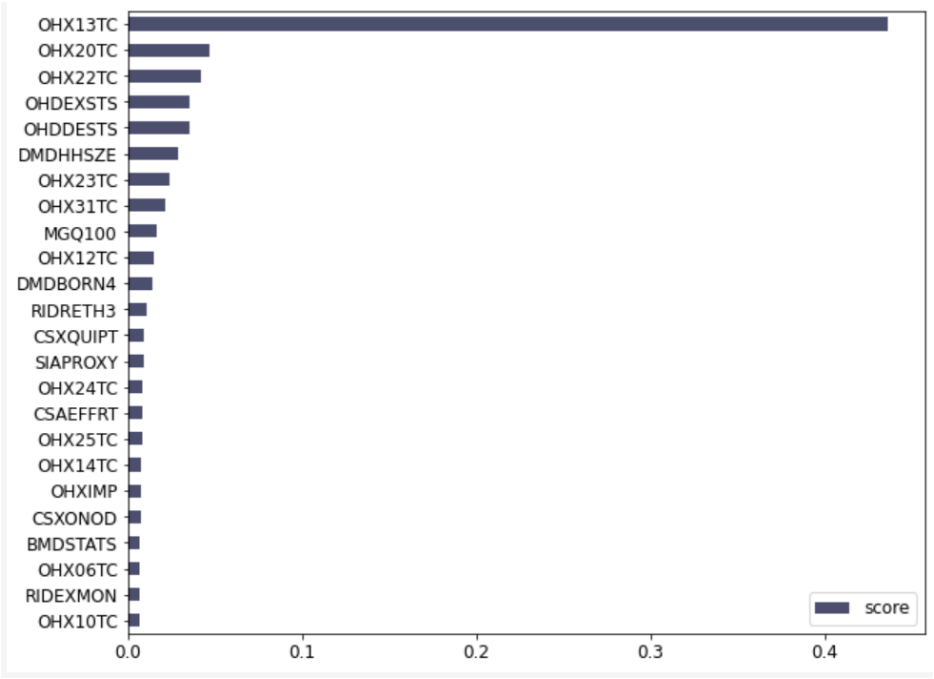

Figure 2 - Importance of features using Xgboost Classifier

3.2 Data Normalization and Upsampling

Normalization plays a pivotal role in ensuring that variables maintain a consistent scale, a crucial aspect for effective machine learning model training. In this analysis, the MinMax scaling technique was applied to normalize the dataset.

Class imbalance, a prevalent challenge in healthcare datasets, was addressed through the implementation of the Synthetic Minority Over-sampling Technique (SMOTE). Significantly, the application of SMOTE was carefully timed. The strategic use of normalization techniques and the cautious application of SMOTE contribute to the robustness and reliability of the subsequent machine learning analysis. These preprocessing steps underscore a commitment to methodological rigor, ensuring the dataset's suitability for predictive modeling.

3.3 Smote

The Synthetic Minority Over-Sampling Technique (SMOTE) is a resampling technique designed to tackle the issue of class imbalance in machine learning datasets. Class imbalance occurs when one class significantly outnumbers the other(s), leading to biased model training and potentially suboptimal performance, particularly for the minority class. This technique is normally used for complete machine learning tasks. However, we used it as a preliminary / transitionary phase throughout the project. The sub steps and working principle is described as below:

Identifying Minority Instances: SMOTE focuses on the minority class, which is the class with fewer instances. In the context of this analysis, the minority class corresponds to individuals diagnosed with a stroke (positive class).

Generating Synthetic Instances: For each minority instance, SMOTE selects its k-nearest neighbors from the same class. Synthetic instances are then created along the line segments connecting the selected instance to its neighbors.

Balancing Class Distribution: The synthetic instances are added to the dataset, effectively increasing the representation of the minority class. This process balances the class distribution, mitigating the impact of class imbalance on model training.

4. DISCUSSION

4.1 Advantages and Disadvantages of Data Pre-processing Steps:

4.1.1 Variable Description Enhancement:

Advantage 1: Improved Interpretability: Accurate and detailed variable descriptions enhance the interpretability of the dataset, reducing ambiguity in feature understanding. This facilitated a clearer understanding of the dataset, aiding researchers and analysts in making informed decisions during analysis.

Advantage 2: Facilitates Feature Selection: Clear variable descriptions aid in the identification of relevant features, contributing to effective feature selection processes, enhancing the efficiency of model training, as irrelevant or redundant features are easily identified and excluded.

Disadvantage 1: Time-Consuming: Scraping variable descriptions from external sources can be time-consuming, particularly for large datasets. The time invested in obtaining detailed variable descriptions may impact the overall speed of the data pre-processing phase.

Disadvantage 2: Dependence on Data Source: The accuracy of variable descriptions depends on the reliability and currency of the external data source. Inaccurate or outdated descriptions may lead to misinterpretation, emphasizing the importance of validating and updating external data sources.

4.1.2 Streamlining and Refinement

Advantage 1: Computational Efficiency: Omitting non-numeric columns and those with significant missing values improves computational efficiency during analysis, accelerating data processing, reducing the time required for subsequent analyses and modeling.

Advantage 2: Focused Feature Set: A reduced set of features (153 columns) enables a more focused and manageable dataset for modeling, simplifying model development and interpretation, contributing to the efficiency of the overall analysis.

Disadvantage 1: Information Loss: Removing columns with missing values may result in the loss of potentially valuable information, especially if the missing values are not entirely at random. Careful consideration is needed to balance data cleanliness with information preservation, minimizing the risk of losing critical insights.

Disadvantage 2: Subjectivity in Thresholds: The choice of the 50% missing value threshold is somewhat subjective and may impact the inclusion or exclusion of certain features. This introduces

variability, and alternative threshold choices might yield different results, emphasizing the need for transparency in decision-making.

4.1.3 Exclusion of Null Values for Target Variable

Advantage 1: Enhanced Target Definition: Rigorous handling of null values in the target variable ensures a well-defined and clear target for predictive modeling, enhancing the precision of model training and evaluation, contributing to more accurate predictions.

Advantage 2: Clean Dataset: The exclusion of entries with null values contributes to a cleaner dataset, minimizing noise in the target variable, simplifying subsequent analyses, reducing the likelihood of errors associated with missing or undefined target values.

Disadvantage 1: Potential Bias: Excluding entries with null values may introduce bias, especially if the null values are not missing completely at random. Researchers must carefully assess the potential bias introduced by excluding entries, considering the implications for the representativeness of the dataset.

Disadvantage 2: Reduced Dataset Size: The exclusion of entries reduces the overall dataset size, potentially impacting the representativeness of the sample. This may affect the robustness of statistical analyses and the generalizability of model predictions.

4.2 Overall Considerations:

4.2.1 Trade-off Between Cleaning and Information Loss:

Balancing the need for data cleaning with the preservation of valuable information is crucial. Careful consideration of the impact on subsequent analyses is necessary to strike the right balance.

4.2.2 Impact on Model Performance:

Evaluating the effectiveness of these pre-processing steps should consider their impact on model performance, aligning with the specific goals and objectives of the analysis.

4.2.3 Iterative Process:

Data pre-processing is often an iterative process. Continuous adjustments based on initial analyses and model performance are essential for refining and improving the overall data quality.

5. CONCLUSION

The data cleaning process plays a pivotal role in the success of machine learning models, influencing their accuracy, interpretability, and generalizability. In the context of predicting strokes using the NHANES dataset, several key data cleaning steps were undertaken to refine and prepare the data for analysis.

The initial dataset presented challenges due to disorganized and unclear variable names. Recognizing the importance of precise variable identification, an elaborate scraping process was implemented on the NHANES website. This process aimed to retrieve accurate and detailed variable descriptions, enhancing clarity and comprehension. As a result, the dataset became more interpretable, reducing ambiguity and contributing to the robustness of subsequent analyses.

A judicious approach was taken to streamline the dataset and eliminate non-numeric columns, as well as those with over 50% missing values. This refinement process resulted in a dataset consisting of 153 columns, ensuring a more focused and manageable set of features for analysis. Furthermore, the adjustment of the target variable, MCQ160F, to binary coding (0 for negative, 1 for positive) facilitated the application of machine learning models, simplifying the interpretation of model outputs.

The focus on the MCQ160F variable, representing the diagnosis of a stroke by a healthcare professional, required meticulous handling of null values. Entries with null values or responses indicating uncertainty or refusal were systematically excluded. This rigorous approach resulted in a dataset comprising 5583 valid entries, ensuring that the target variable was well-defined and suitable for predictive modeling.

In conclusion, the data cleaning process was crucial for transforming the NHANES dataset into a more structured, interpretable, and suitable form for predictive modeling. The combination of enhanced variable descriptions, streamlined features, and meticulous handling of the target variable contributes to the dataset's reliability and sets a solid foundation for subsequent machine learning analyses. The cleanliness and precision achieved through these steps are essential for obtaining meaningful insights and developing accurate stroke prediction models.