DATA 471 Project

Exploring the relationship between different health risk factors and stroke

Hien Nguyen, 300199540

Due: 26 Oct 2021

Executive Summary

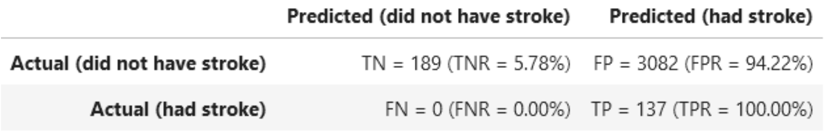
The project's analysis is based on National Health and Nutrition Examination Survey (NHANES) dataset, which aims to collect data on adults' and children's health and nutritional status in the United States. The survey consists of six primary data sets: Demographics, Dietary, Examination, Laboratory, Questionnaire, and Limited Access Data. Using the NHANES dataset, the project aims to identify the important risk factors for stroke and narrow it down to only top important factors, focusing mainly on the Demographic and Questionnaire datasets.

The NHANES datasets are extensive and include lots of health information. Therefore, the Random Forest machine learning algorithm was used to pick out the most important risk factors for stroke. The final chosen dataset contains 17,037 participants with 19 features, including the prediction target (whether a participant has been diagnosed with stroke or not). Other features are related to participants' demographics, physical functioning, and medical conditions such as high blood pressure, age, and moderate recreational activities. To find out the relationship between these features and stroke, Logistic Regression was chosen to get the statistical effects. For the Logistic Regression algorithm, the dataset was split into 80% training set and 20% test set, and both sets include synthetic data created by data imputation methods. They are MICE - multiple imputation method which was used for numerical variables and mode for categorical variables. The synthetic data occupied around 21% of the dataset, which was also the number of missing values.

People with high blood pressure, high cholesterol, high creatinine refrigerated serum level, or who have limited work can do and are 70 years old or older have a higher odds of having a stroke than those who have the opposite stated demographic, physical functioning, and medical conditions.

The model has potential sampling biases since the dataset only contains participants who have been formally diagnosed with specific diseases and not those who may have those disorders but have never been found out. Additionally, the dataset often has more data on particular groups of people. The unbalanced factors in the dataset are ethnicity, marital status, income level, educational level, type of work, type of physical activities, and some diagnosed diseases. We could use the model for inference purposes but should not assume it accurately reflects the actual population where the data is collected, that is, the US. The model correctly predicted people who have had a stroke. However, it did so with a considerable trade-off that it also miss-classified many people who should not have a stroke. In particular, the test set has 3,408 participants, true positive (TP) is 137, and thus

the true positive rate (TPR) is 100%. However, out of 3,271 people who did not have a stroke in the test set, the model only correctly predicts 189 true negative (TN) participants, and the false positive rate (FPR) is 94.22%.



	Predicted (did not have stroke)	Predicted (had stroke)
Actual (did not have stroke)	TN = 189 (TNR = 5.78%)	FP = 3082 (FPR = 94.22%)
Actual (had stroke)	FN = 0 (FNR = 0.00%)	TP = 137 (TPR = 100.00%)

Background

The NHANES data is available to download from the Centers for Disease Control and Prevention (CDC)'s website. Findings from this survey may be used to identify the risk factors for some major diseases.

This project aims to explore the effects of various health factors on stroke, focusing on Demographic and Questionnaire datasets to quickly identify the people at risk of stroke before going on and doing further lab tests.

Three datasets were being used to conduct this report. The initial dataset was massed downloaded from CDC's website. Then it was reduced further based on the number of missing values and became the second dataset. The final dataset, which is also the center of this report, was chosen based on the machine learning Random Forest's feature importance method on the second dataset. In other words, the third dataset was a subset of the second.

The initial dataset was taken from 19 data files from 4 primary datasets: Demographics, Examination, Laboratory, and Questionnaire, which were collected from 2013 to 2018. Every participant is unique. Since the project's focus is stroke, only participants who replied to the question "Ever told you had a stroke?" (MCQ160F) were retained in the dataset. From that, we have a total of 17,037 instances in which 4% of them (684 cases) answered "yes," and 16,353 people answered "no" to the question and 606 features. Out of 606 columns, we had 73 of them with no missing values (22 of them were from the Demographic dataset) and 72 columns with up to 10% missing values. The rest that had NA values of more than 10% was removed from the dataset. That left 145 columns in the dataset (24% of the initial dataset). After examination, we still had most stroke-related features, such as high cholesterol, high blood pressure, diabetes, physical activity, smoking, and triglycerides. Hence, this dataset (17037 instances, 145 features) would move forward for further examination.

There were 14 duplicated pairs from the Laboratory dataset. They represented the same feature in different units, so 14 lab features were removed. Some data in the Demographic that only gave information on the survey itself but not the stroke disease, such as Data release cycle, Interview/Examination status, Language of SP Interview, Household reference person's gender, etc were also excluded. At this point, the dataset has 116 columns left. The feature of Questionnaire Mode Flag (code SMAQUEX2 and SMAQUEX) in the Smoking - Cigarette Use survey had only one unique value. It means values were the same for all participants, so these features were excluded. In this Smoking survey, there were two features about the Cigarette Brand (SMD100BR) and Cigarette Product Code (SMDUPCA), each had more than 300 unique values which was a lot and at this point of the project such details should not add much value to answer the question which factors are important to stroke. Therefore, these two features were also removed. We had 111 columns left (not including the SEQN column), and all of them will be used for a Random Forest model and would be ranked according to the model's feature importance. This second dataset had one target column, "Ever told you had a stroke?”, 77 categorical variables, 33 numerical variables, and a total of 76,452 missing values.

Data Description

Regarding the second dataset of 111 features, numerous data preprocessing methods are applied to the dataset before feeding it to the Random Forest model. They are MICE - multiple imputation method used to replace missing values, Ordinal Encoding to transform categorical data, Label Encoding to transform target column and Standard Scaler to make the numeric data standard normally distributed.

Multiple imputation quantifies the uncertainty in estimating what the missing values might be, avoiding creating false precision (as can happen with single imputation). The imputed numbers are derived from treatment effects in randomized trials, sample means of specific variables, correlations between 2 variables, and the related variances. Hence, it reduces the chance of false-positive or false-negative conclusions (Li, 2015).

The next step is to narrow down the research and determine the effect sizes of Demographic and Questionnaire features on stroke, focusing on questionnaires around a person's physical activity. Thus, out of 111 features of the second dataset, the project will only analyze features mainly from Demographic, Physical Functioning, Occupation, Physical Activity, Blood Pressure & Cholesterol, Diabetes, and Medical Conditions data files. There are 56 features represented in these data files.

Correlation and VIF analysis were used to identify multicollinearity among 56 features. Many features give a very similar correlation with the dependent variable, suggesting high multicollinearity. However, as the variance_inflation_factor function does not accept data with missing values, Simple Imputer with mean strategy was used to fill the NA. The VIF analysis showed that the VIF score is high (up to 252.44). After removing all those features with high VIF scores, the dataset had 18 features left. Though half of them have VIF scores higher than the cut-off value of 10 (see Table 2), they are known risk factors for stroke or interesting features from the Demographic dataset that might be useful for stroke inference.

Details of the chosen features are:

- Limited in amount of work you can do: are you/Is the participant limited in the kind or amount of work {you/s/he} can do because of a physical, mental, or emotional problem?
- Moderate work activity: does {your/ participant’s} work involve moderate-intensity activity that causes small increases in breathing or heart rate, such as brisk walking or carrying light loads for at least 10 minutes continuously?

- Moderate recreational activities: in a typical week {do you/does participant} do any moderate-intensity sports, fitness, or recreational activities that cause a small increase in breathing or heart rate such as brisk walking, bicycling, swimming, or volleyball for at least 10 minutes continuously?
- Walk or bicycle: this question excludes the physical activity at work that participants have already mentioned. What is the usual way {you travel/participant travels} to and from places? For example, to school, for shopping, to work. In a typical week {do you/does participant} walk or use a bicycle for at least 10 minutes continuously to get to and from places?
- Minutes sedentary activity: this question is about sitting at school, at home, getting to and from places, or with friends, including time spent sitting at a desk, traveling in a car or bus, reading, playing cards, watching television, or using a computer. Do not include time spent sleeping. How much time {do you/does participant} usually spend sitting on a typical day?
- Type of work done last week: the question is about {your/ participant's} current job or business. Which type of work {were you/was participant} doing last week.
- Other features include "Close relative had diabetes?", "Doctor told you have diabetes", "Doctor told you - high cholesterol level", "Ever told you had high blood pressure", "Gender", "Country of birth", "Marital status", "Education level", "Race", "Annual family income", "Age at screening", and "Creatinine, refrigerated serum (mg/dL)".

Most of them are categorical variables, except for "Minutes sedentary activity" and "Creatinine, refrigerated serum (mg/dL)." They are numerical variables. Out of these 18 independent variables, only Gender, Race, and Age at screening have no missing values. The lab data - Creatinine refrigerated serum (mg/dL) and the demographic data - Annual family income have the highest number of NA (1660 and 1243 respectively). The total missing values for this dataset were 3580 (see Table 1).

As the project's main purpose is inference instead of prediction, all the categorical variables are transformed into dummies variables. The Annual family income feature also includes the categories "under 20k" and "20k and over". Thus, I combined them with "15-19.99k" and "35-44.99k," respectively (as seeing these two categories are among the most frequent). The get_dummies function provides a parameter called dummy_na, and if this is set to False, NaNs are ignored, and thus they will have the value of the corresponding reference level. We have 1,794 NAs among categorical variables, which means 10% of the dataset has been imputed. However, we still have 126 missing values in Minutes sedentary activity and 1660 in Creatinine refrigerated serum (11% of NAs came from numerical variables). In addition, both numerical variables are skewed and in different scales. As a result, MICE imputer and Standard Scaler were used once again. The data was split into 80% training set and 20% test set with stratifying method to ensure we have a balanced number of people who had strokes between training and test set (each set equally has 4% of people who had strokes).

Ethics, Privacy and Security

Ethics

Ethical considerations are crucial in this project, where the outcome of our machine learning model may be used to predict the likelihood of a stroke. Any bias in our model or its interpretation may lead to adverse health outcomes for users. Furthermore, through poor model design, or improper use of our model, the result of our analysis may lead to or reinforce biases against some groups of people. It is especially true if the model developed is shared more widely (i.e., outside of this course), such as in the form of an online stroke risk calculator.

The survey question, “Ever told you had a stroke?" is a potential source of sample selection bias. Those excluded include those who have had a stroke and died, are now in residential care, or have suffered memory damage that stops them from answering this question accurately. It will not discriminate amongst factors that make someone more likely to die from a stroke, as we have no information on these people. In addition, these types of questions include the ones such as "Doctor told you have diabetes”, “Doctor told you - high cholesterol level," etc. It creates a bias in diagnosis because these questions are about what people have been formally diagnosed with, and so may miss people with these diseases but have never been diagnosed.

Almost all factors in the dataset are imbalanced. We do not have enough data on people diagnosed with high blood pressure, high cholesterol, diabetes, or people with a birth country outside of the US or black and other races rather than white people. The dataset also lacks participants who are active in moderate work/recreational activities or who have limited work can do. This type of negative set bias is due to not having enough samples representative of “the rest of the world” (Srinivasan, 2021).

The ethnic makeup of the NHANES data set is of concern as it is designed to represent the US population but may not be representative of health in New Zealand. The ethnicity categories in the data are lackluster and strongly reflect US sensibilities and priorities. Using US health data in the New Zealand context may not correctly reflect the stroke risk of the ethnically diverse communities in New Zealand. We can assume there are few Māori and Pasifika communities in the dataset. The Asian community is lumped into 'Other Race and Multiracial' in the dataset, despite being a significant ethnic community in New Zealand (15.1% in 2018 (Stats NZ, 2020)).

More data needs to be collected, especially for imbalanced factors to improve the quality of this dataset and apply the model to New Zealand population. In particular, the Race's categories may need to be tweaked to reflect the actual races in New Zealand, such as including a new class of Māori and Pasifika. On top of that, some measurement of "fairness metrics" should be tracked continually to determine whether any errors by the algorithm were biased against certain ethnic groups or other minorities.

Privacy

The health data collected in the NHANES survey is highly sensitive, and a great amount of harm can be caused if identifiable information is released without an individual's consent. In the case of the NHANES data set, participants consent to only de-identified data being released. National Centre for Health Statistics (NCHS) will never release any identifiable information to police, the courts, or any other government agency without a person's consent (NCHS, 2021).

As a starting point, it is worth considering the privacy steps already taken by the NCHS during the data collection. Privacy rules for NHANES are governed by the Confidential Information Protection and Statistical Efficiency Act of 2002 (US). According to the NCHS (2021), all identifiable information is removed, and results for minimal geographic locations are not publicly released if there is a chance that the data can be traced back to a person living in that area.

Given the NCHS claims that the data has already been de-identified. However, Na et al. (2018) showed that third parties who hold physical activity data about an individual might identify participants in the NHANES survey. People with enough time and money could also re-identify patient-level data through this predictive model, though only a few people would be able to acquire enough resources and motivation to do so.

Security

A range of measures could be used to protect our project data, including by:

- Using a strong password for accessing my computer or GitHub account as a group of people was working on the project.
- Using permissions for different user types in our project or the option for changes to be peer-checked by another user.
- Only share the data with authorized people who need to research the data and understand the importance of data ethics and privacy.
- Ensuring there is redundancy in the system, store the data in the cloud software. Google Drive could be an option. Google encrypts our files while they're being transferred and stored. We could leverage their technology to protect our data.

Exploratory Data Analysis

To get to the final dataset, the Random Forest model was applied to the second dataset to find out which health factors are among the most important ones for stroke.

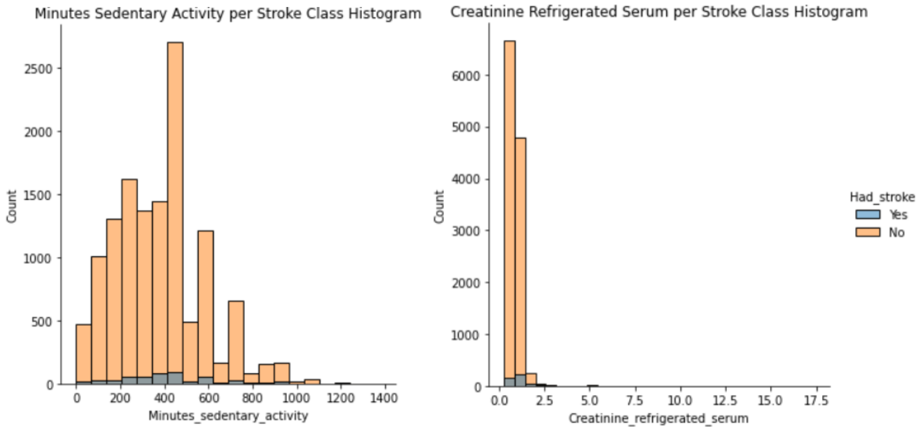
According to the factor's importance rank (see Table 3), one question that was ranked among the top in the importance ladder is "Are you/Is participant limited in the kind or amount of work {you/s/he} can do because of a physical, mental or emotional problem?". This question belongs to the Physical Functioning survey. Therefore, it looks like there is a high correlation between having a stroke and patients' physical conditions, and thus it is worth to be explored more.

Additionally, Age ranked first. Therefore, the project will include Age and take some Demographic features such as Annual family income, Education level, Gender, etc., into the analysis.

The second place was a feature from the Laboratory dataset Creatinine, refrigerated serum. Due to its high rank, this feature will also be included, but it will be the only one from the Laboratory dataset. Laboratory dataset often requires time and proper procedure to obtain. Additionally, some of them are pretty costly, and different hospitals may charge different prices. Moreover, the project aims to study on Demographic and Questionnaire dataset to quickly identify the people at risk of stroke before going on and doing further lab tests. However, as pointed out from the ranking list and stated from various research, high cholesterol (Phillips, 2013), high blood pressure (Phillips, 2013), and diabetes (Peters, 2014) are the known risk factors for stroke. Thus, questions related to these features will also be included. These features form the content of the third and final dataset.

As the name suggests, exploratory data analysis, or in short EDA, is a process of summarizing, visualizing, and becoming intimately familiar with the essential characteristics of the data. Results obtained from EDA could help significantly in data preprocessing and transforming, thus improving model performance. The EDA for the final dataset is divided into two parts, distribution and features interaction.

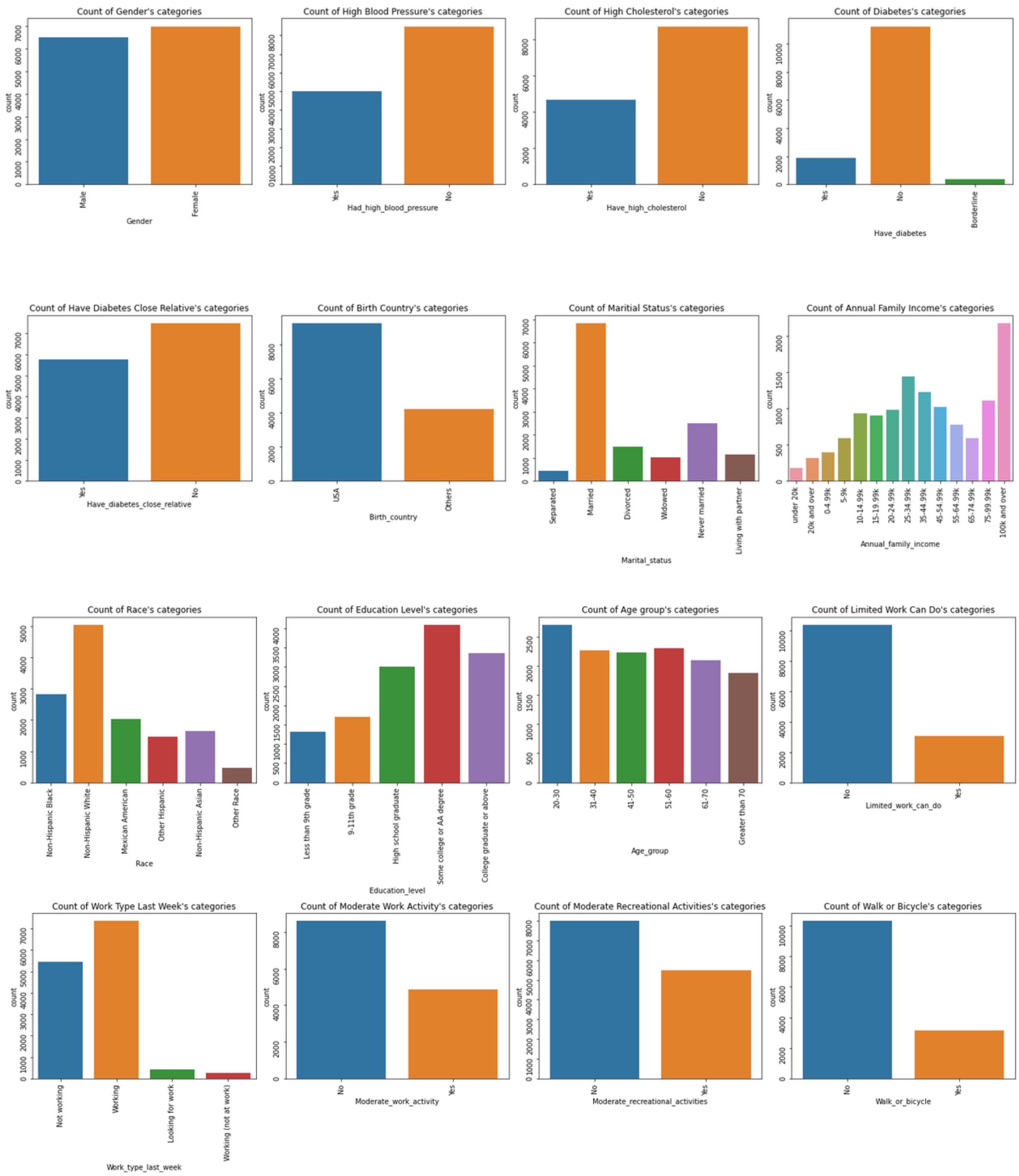
Distribution



Minutes Sedentary Activity per Stroke Class Histogram

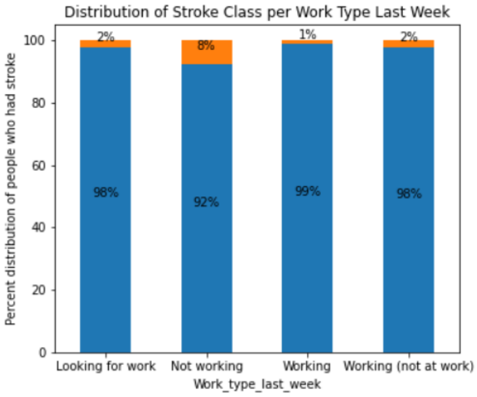
Creatinine Refrigerated Serum per Stroke Class Histogram

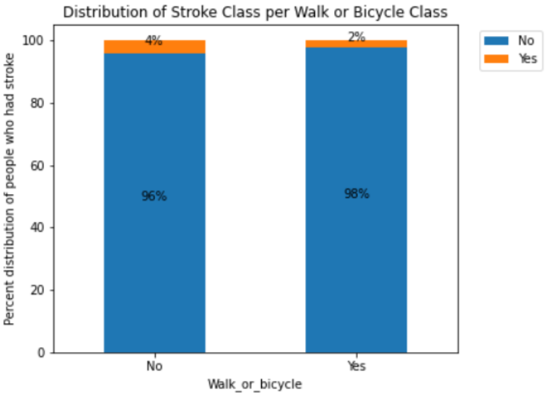
Both numerical variables are right-skewed. Regarding Minutes of sedentary activity, the distribution between people who had a stroke and those who did not have a stroke is quite similar. People spent most of the time sitting or lying down for around 400 to 430 minutes in both categories. On the other hand, most people have Creatinine, refrigerated serum around 0.5 to 1.25 mg/dL; however, the data has a few outliers up to 17.5 mg/dL.

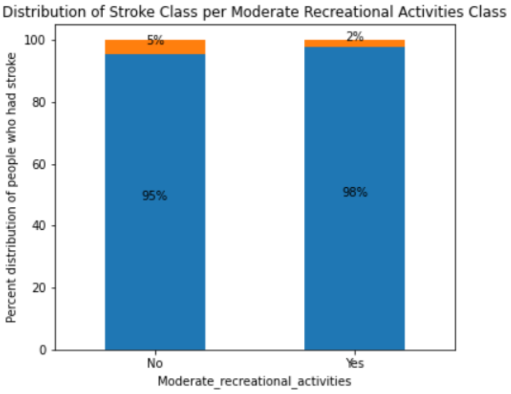


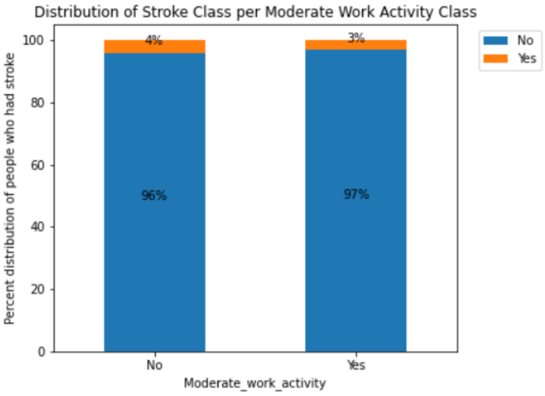
The dataset is imbalanced in almost all categorical features. For example, not counting the imbalanced in the dependent variable Had stroke as mentioned in the Background part, out of 16 categorical variables, only Gender and Age have somewhat the same number of observations between their sub-categories. However, one thing worth noting is that the dataset has no data of people under 20 years of age and that once a person gets over 70 years old, they are in the same group compared to the other age groups, which has the size of 9 years. Overall, we have fewer people diagnosed with high blood pressure, high cholesterol, diabetes, have close relatives with diabetes than people who are not diagnosed with these diseases. We also have more Non-Hispanic White people than the other races, more people with higher education, more working, married people, and more people earning 25k and above. Regarding the physical activity survey, more people replied "no" than "yes" in all questions if they have limited work can do, do moderate work activity, do moderate recreational activities, and walk or ride a bicycle.

When deep-diving into the physical activity questions, there are noticeable patterns between these categorical and dependent variables.



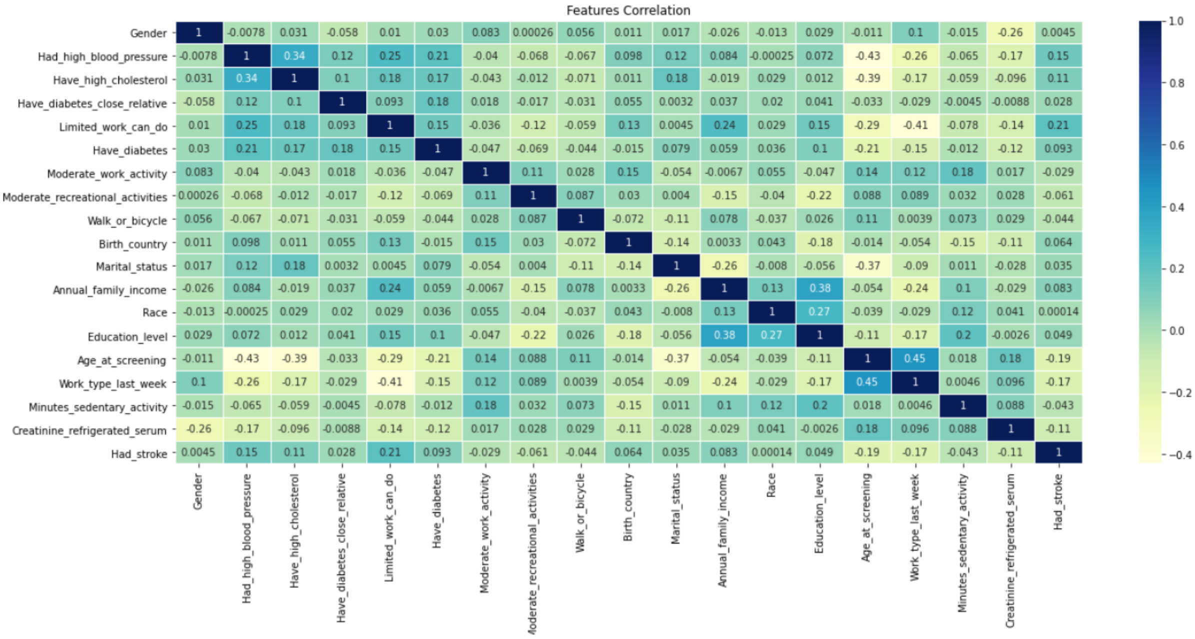




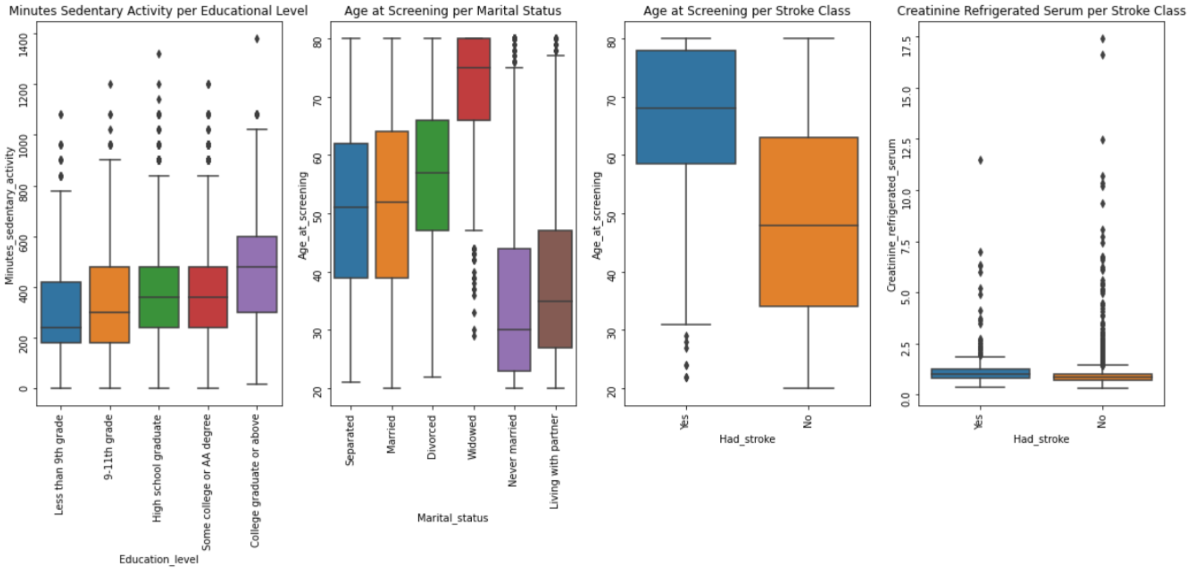


- 8% of not working class have had a stroke while it is maximum only 2% for other types of work.
- 4% of people who don't walk or use bicycles have had a stroke, while only 2% of people had a stroke if they walked or used bicycles.
- More people had a stroke than did not have a stroke if they answered "no" to Moderate recreational activities and moderate work activity compared to answered "yes."

Features interaction



All data values of the original dataset were either recorded in number or coded with number even for categorical variables, such as "yes" is coded 1 and "no" is coded 2. It makes computing correlation possible for categorical variables. Limited work can do had the highest correlation (0.21) with stroke, followed by Age (-0.19), Work type (-0.17), High blood pressure (0.15), High cholesterol (0.11), and Creatinine, refrigerated serum (0.11). However, in general, Age and Work Type have the highest correlation among all (0.45), and the next highest pair is Annual family income and Education level. High cholesterol and High blood pressure also belong in the top 3 highest correlation (0.34).



There is a strong positive relationship between Age and stroke, while stroke and Creatinine, refrigerated serum have a weak positive relationship. Additionally, the higher the Education level, the more significant number of minutes sedentary people have on average. Widowed people have the highest average age, and never married people have the lowest average age.

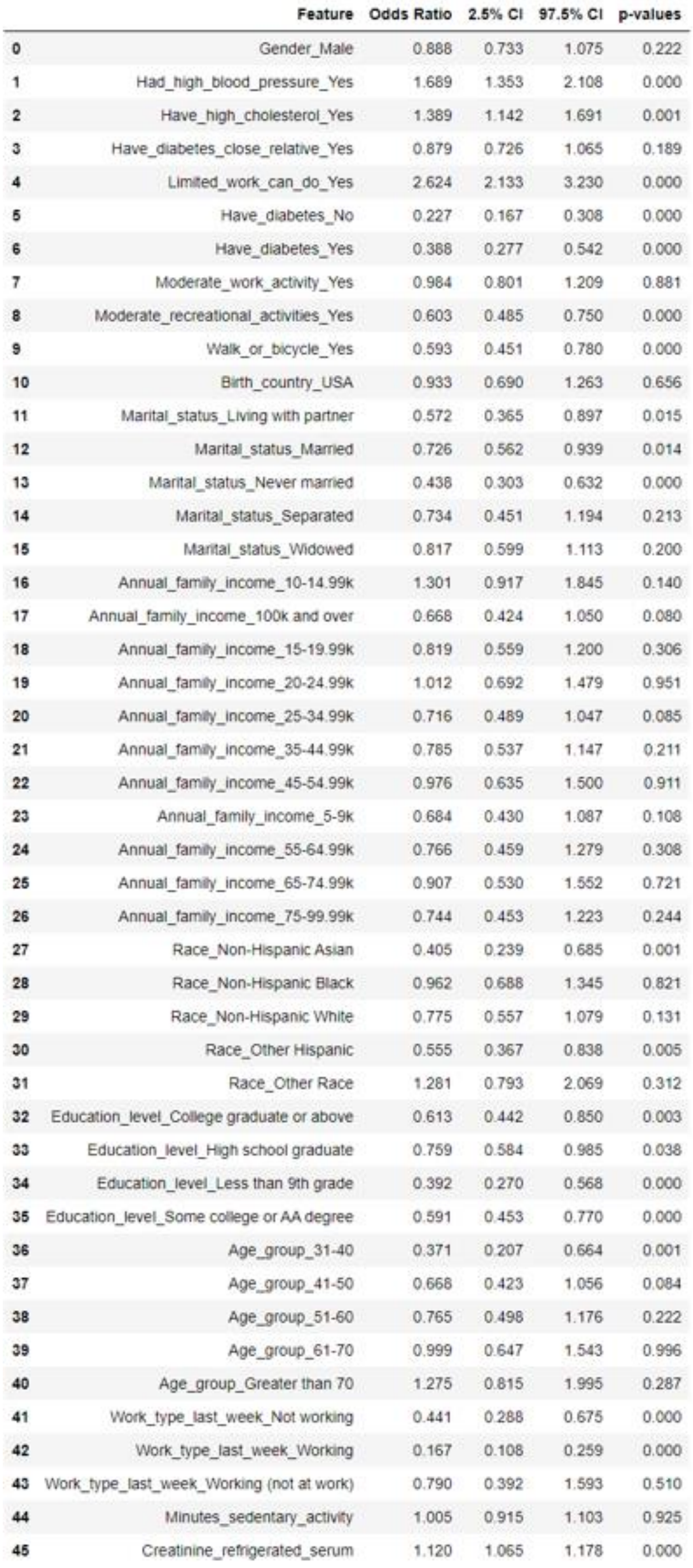
Detailed Analysis Results

As mentioned before, the Random Forest algorithm was used to help find the subset of the second dataset. Random Forest is a method of ensemble learning. It uses the result of a group of Decision Tree to predict the outcome that gets the most votes, hence, achieving better results. Feature importance is an excellent quality of Random Forest. It gives information on the measurement of the relative importance of each feature. The algorithm achieves this by looking at how much the tree nodes that use that feature reduce impurity on average (across all trees in the forest), in other words, a weighted average computation (Géron, 2019).

Logistic Regression is a popular method that gives information about the statistical effect and significance of features. Furthermore, as we aim to explain the relationship between the Demographic and some health factors on stroke, this algorithm is ideal for our final dataset of 18 features.

According to the Logistic Regression summary (Table 4 below), we have strong evidence to suggest that High blood pressure, High cholesterol, Limited work can do, Have diabetes, Moderate recreational activities, Walk or bicycle, Race, Education level, Age, Work type last week and Creatinine refrigerated serum have a statistically significant relationship with stroke if one is numerical variable or statistically significant relationship with its reference level if it is a categorical variable, as their p-values are smaller than significance level alpha equal to 0.05. Those features with p-values larger than 0.05 are Gender, Have diabetes close relative, Moderate work activity, Birth country, Marital status, Annual family income, and Minutes of sedentary activity. It is somewhat similar to the Random Forest's feature importance, especially Gender, Moderate work activity, and the Birth country, as they are all at the bottom of the list of 110 features. Thus, the result's analysis will only explain the "effect sizes" of the significant features.

Table 4: Results of logistic regression model



	Feature	Odds Ratio	2.5% CI	97.5% CI	p-values
0	Gender_Male	0.888	0.733	1.075	0.222
1	Had_high_blood_pressure_Yes	1.689	1.353	2.108	0.000
2	Have_high_cholesterol_Yes	1.389	1.142	1.691	0.001
3	Have_diabetes_close_relative_Yes	0.879	0.726	1.065	0.189
4	Limited_work_can_do_Yes	2.624	2.133	3.230	0.000
5	Have_diabetes_No	0.227	0.167	0.308	0.000
6	Have_diabetes_Yes	0.388	0.277	0.542	0.000
7	Moderate_work_activity_Yes	0.984	0.801	1.209	0.881
8	Moderate_recreational_activities_Yes	0.603	0.485	0.750	0.000
9	Walk_or_bicycle_Yes	0.593	0.451	0.780	0.000
10	Birth_country_USA	0.933	0.690	1.263	0.656
11	Marital_status_Living with partner	0.572	0.365	0.897	0.015
12	Marital_status_Married	0.726	0.562	0.939	0.014
13	Marital_status_Never married	0.438	0.303	0.632	0.000
14	Marital_status_Separated	0.734	0.451	1.194	0.213
15	Marital_status_Widowed	0.817	0.599	1.113	0.200
16	Annual_family_income_10-14.99k	1.301	0.917	1.845	0.140
17	Annual_family_income_100k and over	0.668	0.424	1.050	0.080
18	Annual_family_income_15-19.99k	0.819	0.559	1.200	0.306
19	Annual_family_income_20-24.99k	1.012	0.692	1.479	0.951
20	Annual_family_income_25-34.99k	0.716	0.489	1.047	0.085
21	Annual_family_income_35-44.99k	0.785	0.537	1.147	0.211
22	Annual_family_income_45-54.99k	0.976	0.635	1.500	0.911
23	Annual_family_income_5-9k	0.684	0.430	1.087	0.108
24	Annual_family_income_55-64.99k	0.766	0.459	1.279	0.308
25	Annual_family_income_65-74.99k	0.907	0.530	1.552	0.721
26	Annual_family_income_75-99.99k	0.744	0.453	1.223	0.244
27	Race_Non-Hispanic Asian	0.405	0.239	0.685	0.001
28	Race_Non-Hispanic Black	0.962	0.688	1.345	0.821
29	Race_Non-Hispanic White	0.775	0.557	1.079	0.131
30	Race_Other Hispanic	0.555	0.367	0.838	0.005
31	Race_Other Race	1.281	0.793	2.069	0.312
32	Education_level_College graduate or above	0.613	0.442	0.850	0.003
33	Education_level_High school graduate	0.759	0.584	0.985	0.038
34	Education_level_Less than 9th grade	0.392	0.270	0.568	0.000
35	Education_level_Some college or AA degree	0.591	0.453	0.770	0.000
36	Age_group_31-40	0.371	0.207	0.664	0.001
37	Age_group_41-50	0.668	0.423	1.056	0.084
38	Age_group_51-60	0.765	0.498	1.176	0.222
39	Age_group_61-70	0.999	0.647	1.543	0.996
40	Age_group_Greater than 70	1.275	0.815	1.995	0.287
41	Work_type_last_week_Not working	0.441	0.288	0.675	0.000
42	Work_type_last_week_Working	0.167	0.108	0.259	0.000
43	Work_type_last_week_Working (not at work)	0.790	0.392	1.593	0.510
44	Minutes_sedentary_activity	1.005	0.915	1.103	0.925
45	Creatinine_refrigerated_serum	1.120	1.065	1.178	0.000

- The reference level for Had high blood pressure, high cholesterol, limited work can do, Moderate recreational activities, Walk or bicycle is No. As for Have diabetes, the reference level is Borderline. For

Age, the reference level is 20-30 years old, for Race is Mexican American, for Educational level is 9-11th grade, and for Work type last week is Looking for work.

- Each additional mg/dL of Creatinine refrigerated serum is associated with a multiplicative change of 11.7% or 1.117 (95% CI: (1.061, 1.176)) in the odds of stroke (i.e., an increase in the odds).
- Those who have high blood pressure are more likely to have a stroke than those who do not have. It means that if a person has high blood pressure, the odds of having a stroke is 1.695 (95% CI: (1.356, 2.120)) times the odds of having a stroke for a person who does not have high blood pressure.
- Those who have high cholesterol are more likely to have a stroke than those who do not have. It means that if a person has high cholesterol, the odds of having a stroke is 1.387 (95% CI: (1.136, 1.693)) times the odds of having a stroke for a person who does not have high cholesterol.
- Those who have limited work can do are more likely to have a stroke than those who do not have. It means that if a person has limited work can do, the odds of having a stroke is 2.647 (95% CI: (2.149, 3.261)) times the odds of having a stroke for a person who does not have limited work can do.
- Those without diabetes are less likely to have a stroke than those on the borderline of having diabetes. It means that if a person does not have diabetes, the odds of having a stroke is 0.229 (95% CI: (0.167, 0.316)) times that of one who is on the borderline of having diabetes. In addition, the odds of having a stroke for those who have diabetes is 0.392 (95% CI: (0.278, 0.554)) times that of one who is on the borderline of having diabetes.
- Those who do moderate recreational activities are less likely to have a stroke than those who do not. It means that if a person does moderate recreational activities, the odds of having a stroke is 0.603 (95% CI: (0.485, 0.751)) times the odds of having a stroke for a person who does not do moderate recreational activities.
- Those who walk or ride bicycles are less likely to have a stroke than those who do not. It means that if a person walks or rides a bicycle, the odds of having a stroke is 0.597 (95% CI: (0.454, 0.786)) times the odds of having a stroke for a person who does not walk or ride bicycles.
- There is no significant difference between Non-Hispanic Black people or Non-Hispanic White people, or Other Race with Mexican American people in the odds of having a stroke. However, there is a significant difference between Non-Hispanic Asians or Other Hispanic with Mexican American people. In particular, Non-Hispanic Asians or Other Hispanics are much less likely to have a stroke than Mexican Americans. The odds of having a stroke are 0.407 (95% CI: (0.239, 0.694), for Non-Hispanic Asians) and 0.558 (95% CI: (0.368, 0.846), for Other Hispanic people) times the odds of having a stroke for Mexican Americans.
- Those with an education level that is not 9-11th grade are less likely to have a stroke than those with the highest education level in the 9-11th grade. In general, the odds of having a stroke for those with education levels other than 9-11th grade is ranging from 0.395 (for Less than 9th grade) to 0.758 (for High school graduate) times the odds of having a stroke for a person who has the highest education level in the range 9-11th grade.

- Since the reference level for Age is 20-30 years old, once a person reaches 70 years, it increases the odds of having a stroke to 1.290 (95% CI: (0.816, 2.039)) times the odds of having a stroke for a 20-30-year-old person.
- Those who are working, working but not at work, or not working are less likely to have a stroke than those looking for job. The odds of having a stroke are 0.168 (for working people), 0.792 (for working but not at work people), and 0.449 (for not working people) times the odds of having a stroke for looking for work people.

As the dataset is highly imbalanced, the accuracy score is not suitable for measuring the predictive performance. Additionally, if false positive and false negative have different associated costs, it is preferable to maximize precision-recall AUC. In our case, there are related costs between false positive and false negative, especially in the medical context. For example, false-negative should be more dangerous. The damage from not knowing that a patient might have a stroke is more severe than believing that they have a stroke though it may be a wrong assumption. Therefore, instead of keeping the Logistic Regression model's default threshold (0.5), the new threshold of 0.001669 was chosen to ensure the model would not miss any stroke patient and aim to have a 100% true positive rate. If any individual has a probability of having a stroke larger than 0.1669%, the model will classify that person as positive or having a stroke. As a result, the model is less likely to have false-negative and is more likely to have more false-positive. Though the false positive rate is very high, the recall score is 100%.

Conclusions and Recommendations

Using Logistic Regression, we get some great insights on the statistics effects of various health factors such as Creatinine refrigerated serum, high blood pressure, high cholesterol, diabetes, physical activities, and demographics such as age and ethnicity on stroke. Among these features, there is strong evidence to suggest that High blood pressure, High cholesterol, Limited work can do, Have diabetes, Moderate recreational activities, Walk or bicycle, Race, Education level, Age and Work type last week have a strong relationship with each corresponding reference level and Creatinine refrigerated serum also has statistically significant relationship with stroke. Furthermore, people with high blood pressure, high cholesterol, high creatinine refrigerated serum level, or who have limited work can do and are 70 years old or older have a higher odds of having a stroke than those who have the opposite health conditions. Additionally, if a person often does moderate recreational activities, walks or rides bicycles, it is less likely to have a stroke than who does not.

It is important to stay healthy and thus reduce the risk of stroke by avoiding diseases such as high blood pressure and high cholesterol and increasing moderate recreational physical activities such as walking or bicycling. However, the dataset used in this project could have many potential biases against certain groups of people, such as black people or those who are not married or who are looking for work. In addition, the data was also collected based on the US population only. Therefore, any user of this model needs to be aware of its limitations and be critical while using it.

Some potential interactions between the mentioned diseases with Age have not been included in this project. In addition, there is evidence that tobacco consumption is an independent risk factor for both ischemic and hemorrhagic stroke (Peters, 2013). These features could be the omitted variables and should be considered for future research.

Reference

Géron, Aurélien (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition. 2nd edition. *O'Reilly Media, Inc*., 2019. Print.

Na, L., Yang, C., Lo, C., Zhao, F., Fukuoka, Y. & Aswani, A. (December 21, 2018). Feasibility of Re-identifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning. *Jama Network Open, 1(8)*. https://dx.doi.org/10.1001%2Fjamanetworkopen.2018.6040

National Center for Health Statistics. (2021). How the National Health and Nutrition Examination Survey Keeps Your Information Confidential. https://www.cdc.gov/nchs/data/nhanes/participant/Confidentiality-Brochure-2021.pdf

Li P, Stuart EA, Allison DB. Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA*. 2015;314(18):1966–1967. doi:10.1001/jama.2015.15281

Phillips, L. A. (2014). "Stroke Survivors' Endorsement of a 'Stress Belief Model' of Stroke Prevention Predicts Control of Risk Factors for Recurrent Stroke." *Psychology, Health & Medicine*, 19.5, 519–524.

Peters, S. A. E., Huxley, R. R. & Woodward, M. (2014). "Diabetes as a Risk Factor for Stroke in Women Compared with Men: a Systematic Review and Meta-Analysis of 64 Cohorts, Including 775385 Individuals and 12539 Strokes." *The Lancet (British edition)*, 383.9933, 1973–1980.

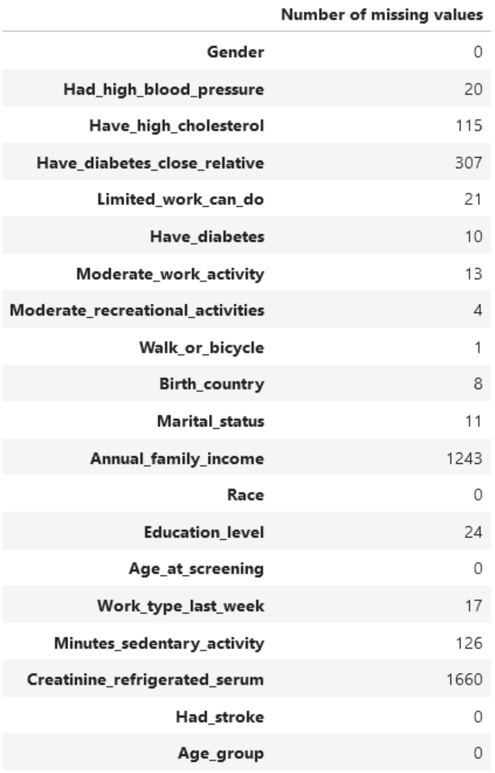
Peters, S. A. E., Huxley, R. R. & Woodward, M. (2013). "Smoking as a Risk Factor for Stroke in Women Compared with Men: a Systematic Review and Meta-Analysis of 81 Cohorts, Including 3,980,359 Individuals and 42,401 Strokes." *Stroke (1970)*, 44.10, 2821–2828.

Srinivasan, R., Chander, A. 2021. Biases in AI Systems: A survey for practitioners. *Queue, Volume 19, Issue 2*. https://dl.acm.org/doi/10.1145/3466132.3466134

Stats NZ. (2020, September 3). Ethnic group summaries reveal New Zealand's multicultural makeup. https://www.stats.govt.nz/news/ethnic-group-summaries-reveal-new-zealands-multicultural-make-up

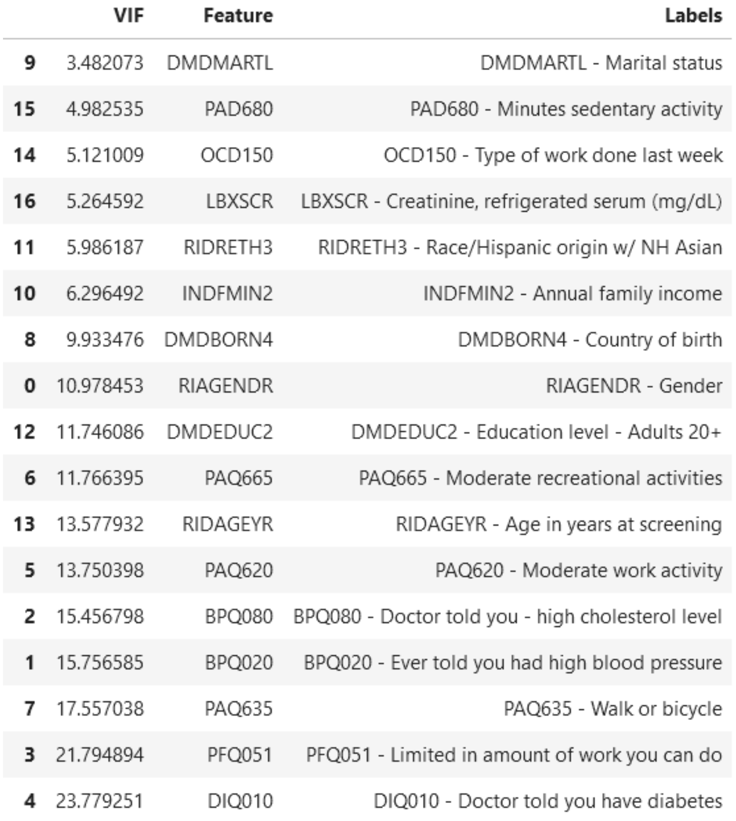
Appendices

Table 1: Number of missing values per feature (final dataset)



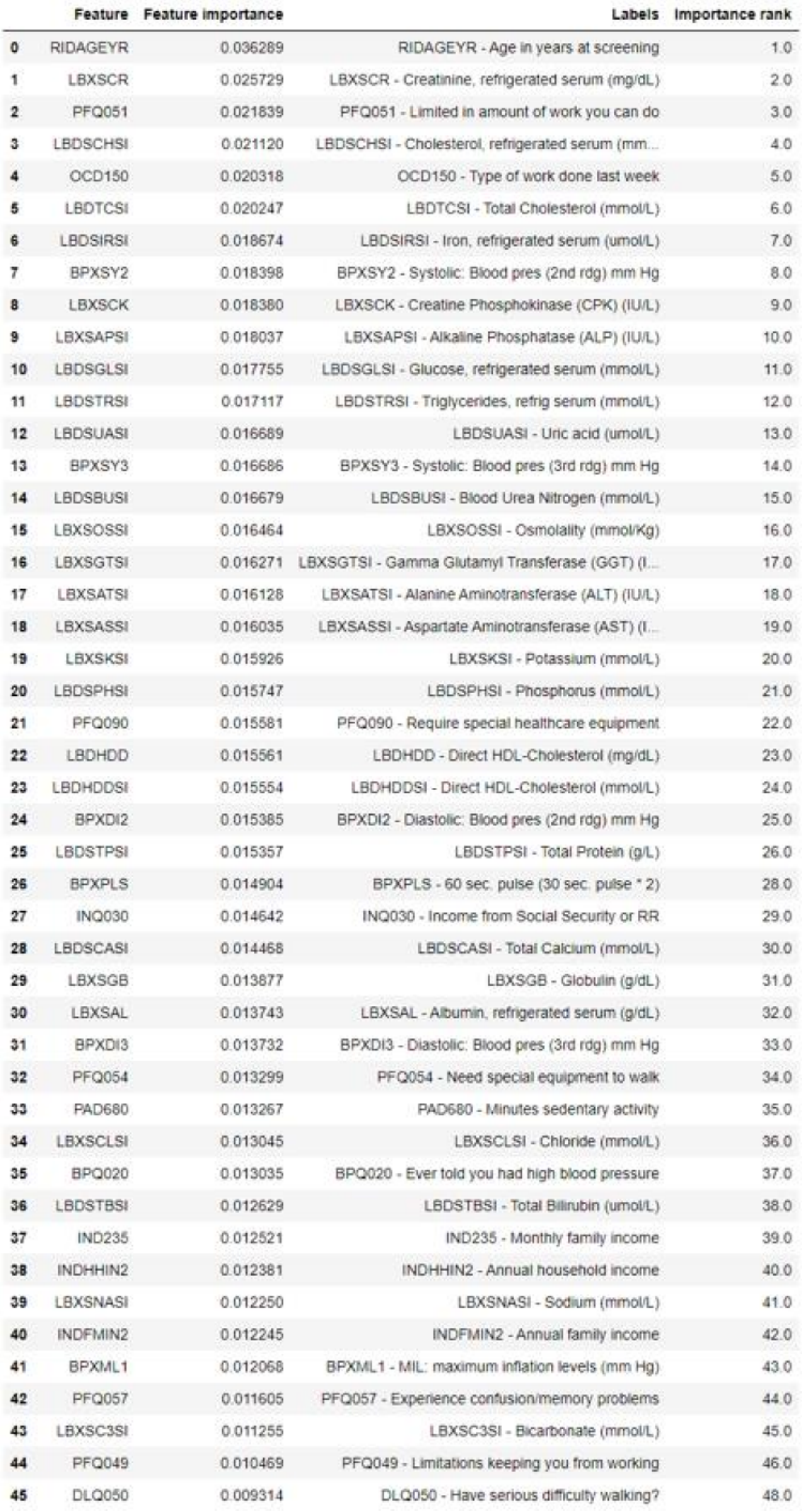
	Number of missing values
Gender	0
Had_high_blood_pressure	20
Have_high_cholesterol	115
Have_diabetes_close_relative	307
Limited_work_can_do	21
Have_diabetes	10
Moderate_work_activity	13
Moderate_recreational_activities	4
Walk_or_bicycle	1
Birth_country	8
Marital_status	11
Annual_family_income	1243
Race	0
Education_level	24
Age_at_screening	0
Work_type_last_week	17
Minutes_sedentary_activity	126
Creatinine_refrigerated_serum	1660
Had_stroke	0
Age_group	0

Table 2: VIF values of chosen features

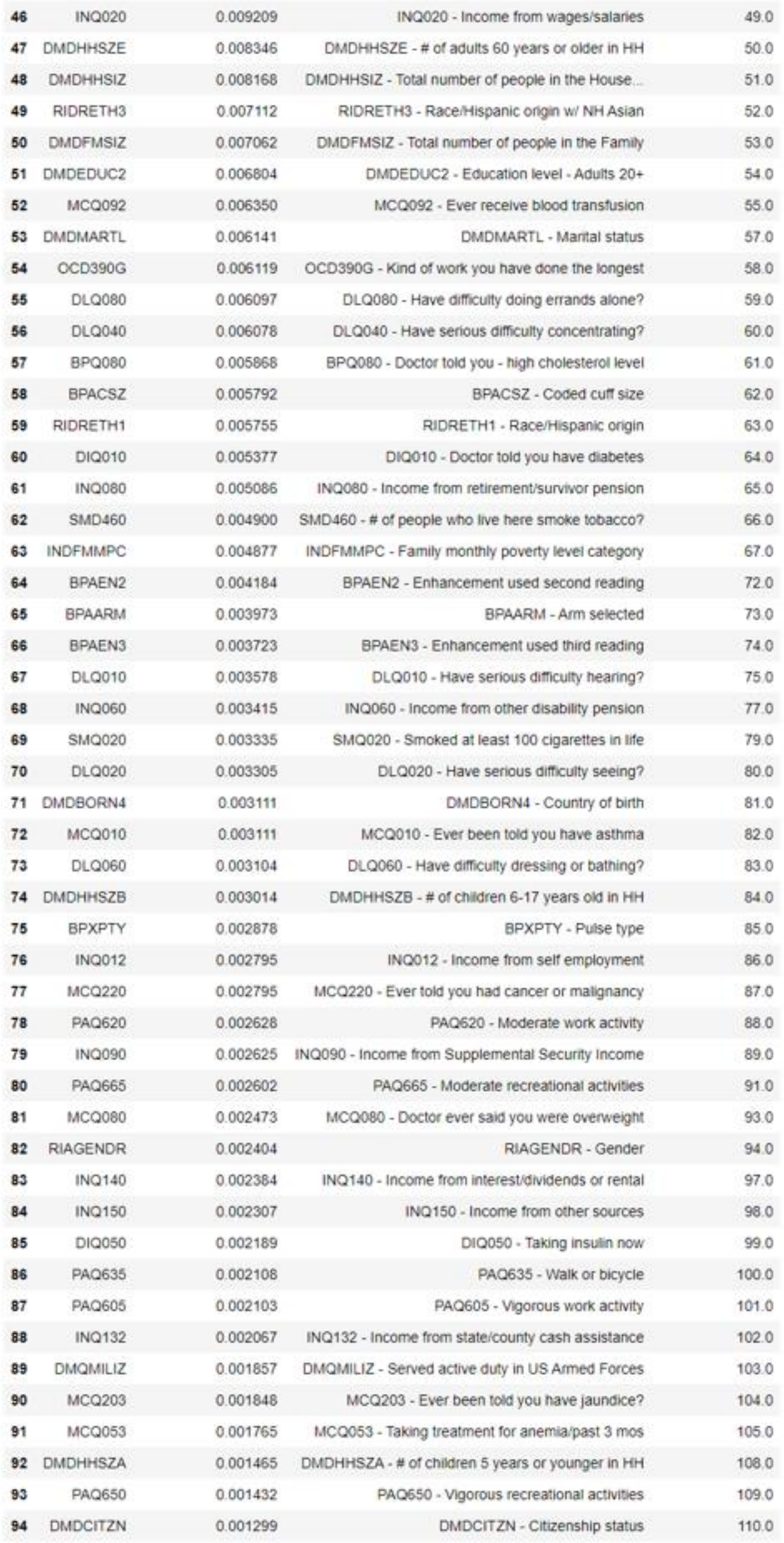


	VIF	Feature	Labels
9	3.482073	DMDMARTL	DMDMARTL - Marital status
15	4.982535	PAD680	PAD680 - Minutes sedentary activity
14	5.121009	OCD150	OCD150 - Type of work done last week
16	5.264592	LBXSCR	LBXSCR - Creatinine, refrigerated serum (mg/dL)
11	5.986187	RIDRETH3	RIDRETH3 - Race/Hispanic origin w/ NH Asian
10	6.296492	INDFMIN2	INDFMIN2 - Annual family income
8	9.933476	DMDBORN4	DMDBORN4 - Country of birth
0	10.978453	RIAGENDR	RIAGENDR - Gender
12	11.746086	DMDEDUC2	DMDEDUC2 - Education level - Adults 20+
6	11.766395	PAQ665	PAQ665 - Moderate recreational activities
13	13.577932	RIDAGEYR	RIDAGEYR - Age in years at screening
5	13.750398	PAQ620	PAQ620 - Moderate work activity
2	15.456798	BPQ080	BPQ080 - Doctor told you - high cholesterol level
1	15.756585	BPQ020	BPQ020 - Ever told you had high blood pressure
7	17.557038	PAQ635	PAQ635 - Walk or bicycle
3	21.794894	PFQ051	PFQ051 - Limited in amount of work you can do
4	23.779251	DIQ010	DIQ010 - Doctor told you have diabetes

Table 3: Random Forest’s feature importance (rank from the most important to the least important)



	Feature	Feature importance	Labels	Importance rank
0	RIDAGEYR	0.036289	RIDAGEYR - Age in years at screening	1.0
1	LBXSCR	0.025729	LBXSCR - Creatinine, refrigerated serum (mg/dL)	2.0
2	PFQ051	0.021839	PFQ051 - Limited in amount of work you can do	3.0
3	LBDSCHSI	0.021120	LBDSCHSI - Cholesterol, refrigerated serum (mm...	4.0
4	OCD150	0.020318	OCD150 - Type of work done last week	5.0
5	LBDTCSI	0.020247	LBDTCSI - Total Cholesterol (mmol/L)	6.0
6	LBDSIRSI	0.018674	LBDSIRSI - Iron, refrigerated serum (umol/L)	7.0
7	BPXSY2	0.018398	BPXSY2 - Systolic: Blood pres (2nd rdg) mm Hg	8.0
8	LBXSCK	0.018380	LBXSCK - Creatine Phosphokinase (CPK) (IU/L)	9.0
9	LBXSAPSI	0.018037	LBXSAPSI - Alkaline Phosphatase (ALP) (IU/L)	10.0
10	LBDSGLSI	0.017755	LBDSGLSI - Glucose, refrigerated serum (mmol/L)	11.0
11	LBDSTRSI	0.017117	LBDSTRSI - Triglycerides, refrig serum (mmol/L)	12.0
12	LBDSUASI	0.016689	LBDSUASI - Uric acid (umol/L)	13.0
13	BPXSY3	0.016686	BPXSY3 - Systolic: Blood pres (3rd rdg) mm Hg	14.0
14	LBDSBUSI	0.016679	LBDSBUSI - Blood Urea Nitrogen (mmol/L)	15.0
15	LBXSOSSI	0.016464	LBXSOSSI - Osmolality (mmol/Kg)	16.0
16	LBXSGTSI	0.016271	LBXSGTSI - Gamma Glutamyl Transferase (GGT) (I...	17.0
17	LBXSATSI	0.016128	LBXSATSI - Alanine Aminotransferase (ALT) (IU/L)	18.0
18	LBXSASSI	0.016035	LBXSASSI - Aspartate Aminotransferase (AST) (I...	19.0
19	LBXSKSI	0.015926	LBXSKSI - Potassium (mmol/L)	20.0
20	LBDSPHSI	0.015747	LBDSPHSI - Phosphorus (mmol/L)	21.0
21	PFQ090	0.015581	PFQ090 - Require special healthcare equipment	22.0
22	LBDHDD	0.015561	LBDHDD - Direct HDL-Cholesterol (mg/dL)	23.0
23	LBDHDDSI	0.015554	LBDHDDSI - Direct HDL-Cholesterol (mmol/L)	24.0
24	BPXDI2	0.015385	BPXDI2 - Diastolic: Blood pres (2nd rdg) mm Hg	25.0
25	LBDSTPSI	0.015357	LBDSTPSI - Total Protein (g/L)	26.0
26	BPXPLS	0.014904	BPXPLS - 60 sec. pulse (30 sec. pulse * 2)	28.0
27	INQ030	0.014642	INQ030 - Income from Social Security or RR	29.0
28	LBDSCASI	0.014468	LBDSCASI - Total Calcium (mmol/L)	30.0
29	LBXSGB	0.013877	LBXSGB - Globulin (g/dL)	31.0
30	LBXSAL	0.013743	LBXSAL - Albumin, refrigerated serum (g/dL)	32.0
31	BPXDI3	0.013732	BPXDI3 - Diastolic: Blood pres (3rd rdg) mm Hg	33.0
32	PFQ054	0.013299	PFQ054 - Need special equipment to walk	34.0
33	PAD680	0.013267	PAD680 - Minutes sedentary activity	35.0
34	LBXSCLSI	0.013045	LBXSCLSI - Chloride (mmol/L)	36.0
35	BPQ020	0.013035	BPQ020 - Ever told you had high blood pressure	37.0
36	LBDSTBSI	0.012629	LBDSTBSI - Total Bilirubin (umol/L)	38.0
37	IND235	0.012521	IND235 - Monthly family income	39.0
38	INDHHIN2	0.012381	INDHHIN2 - Annual household income	40.0
39	LBXSNASI	0.012250	LBXSNASI - Sodium (mmol/L)	41.0
40	INDFMIN2	0.012245	INDFMIN2 - Annual family income	42.0
41	BPXML1	0.012068	BPXML1 - MIL: maximum inflation levels (mm Hg)	43.0
42	PFQ057	0.011605	PFQ057 - Experience confusion/memory problems	44.0
43	LBXSC3SI	0.011255	LBXSC3SI - Bicarbonate (mmol/L)	45.0
44	PFQ049	0.010469	PFQ049 - Limitations keeping you from working	46.0
45	DLQ050	0.009314	DLQ050 - Have serious difficulty walking?	48.0



46	INQ020	0.009209	INQ020 - Income from wages/salaries	49.0
47	DMDHHSZE	0.008346	DMDHHSZE - # of adults 60 years or older in HH	50.0
48	DMDHHSIZ	0.008168	DMDHHSIZ - Total number of people in the House...	51.0
49	RIDRETH3	0.007112	RIDRETH3 - Race/Hispanic origin w/ NH Asian	52.0
50	DMDFMSIZ	0.007062	DMDFMSIZ - Total number of people in the Family	53.0
51	DMDEDUC2	0.006804	DMDEDUC2 - Education level - Adults 20+	54.0
52	MCQ092	0.006350	MCQ092 - Ever receive blood transfusion	55.0
53	DMDMARTL	0.006141	DMDMARTL - Marital status	57.0
54	OCD390G	0.006119	OCD390G - Kind of work you have done the longest	58.0
55	DLQ080	0.006097	DLQ080 - Have difficulty doing errands alone?	59.0
56	DLQ040	0.006078	DLQ040 - Have serious difficulty concentrating?	60.0
57	BPQ080	0.005868	BPQ080 - Doctor told you - high cholesterol level	61.0
58	BPACSZ	0.005792	BPACSZ - Coded cuff size	62.0
59	RIDRETH1	0.005755	RIDRETH1 - Race/Hispanic origin	63.0
60	DIQ010	0.005377	DIQ010 - Doctor told you have diabetes	64.0
61	INQ080	0.005086	INQ080 - Income from retirement/survivor pension	65.0
62	SMD460	0.004900	SMD460 - # of people who live here smoke tobacco?	66.0
63	INDFMMPC	0.004877	INDFMMPC - Family monthly poverty level category	67.0
64	BPAEN2	0.004184	BPAEN2 - Enhancement used second reading	72.0
65	BPAARM	0.003973	BPAARM - Arm selected	73.0
66	BPAEN3	0.003723	BPAEN3 - Enhancement used third reading	74.0
67	DLQ010	0.003578	DLQ010 - Have serious difficulty hearing?	75.0
68	INQ060	0.003415	INQ060 - Income from other disability pension	77.0
69	SMQ020	0.003335	SMQ020 - Smoked at least 100 cigarettes in life	79.0
70	DLQ020	0.003305	DLQ020 - Have serious difficulty seeing?	80.0
71	DMDBORN4	0.003111	DMDBORN4 - Country of birth	81.0
72	MCQ010	0.003111	MCQ010 - Ever been told you have asthma	82.0
73	DLQ060	0.003104	DLQ060 - Have difficulty dressing or bathing?	83.0
74	DMDHHSZB	0.003014	DMDHHSZB - # of children 6-17 years old in HH	84.0
75	BPXPTY	0.002878	BPXPTY - Pulse type	85.0
76	INQ012	0.002795	INQ012 - Income from self employment	86.0
77	MCQ220	0.002795	MCQ220 - Ever told you had cancer or malignancy	87.0
78	PAQ620	0.002628	PAQ620 - Moderate work activity	88.0
79	INQ090	0.002625	INQ090 - Income from Supplemental Security Income	89.0
80	PAQ665	0.002602	PAQ665 - Moderate recreational activities	91.0
81	MCQ080	0.002473	MCQ080 - Doctor ever said you were overweight	93.0
82	RIAGENDR	0.002404	RIAGENDR - Gender	94.0
83	INQ140	0.002384	INQ140 - Income from interest/dividends or rental	97.0
84	INQ150	0.002307	INQ150 - Income from other sources	98.0
85	DIQ050	0.002189	DIQ050 - Taking insulin now	99.0
86	PAQ635	0.002108	PAQ635 - Walk or bicycle	100.0
87	PAQ605	0.002103	PAQ605 - Vigorous work activity	101.0
88	INQ132	0.002067	INQ132 - Income from state/county cash assistance	102.0
89	DMQMILIZ	0.001857	DMQMILIZ - Served active duty in US Armed Forces	103.0
90	MCQ203	0.001848	MCQ203 - Ever been told you have jaundice?	104.0
91	MCQ053	0.001765	MCQ053 - Taking treatment for anemia/past 3 mos	105.0
92	DMDHHSZA	0.001465	DMDHHSZA - # of children 5 years or younger in HH	108.0
93	PAQ650	0.001432	PAQ650 - Vigorous recreational activities	109.0
94	DMDCITZN	0.001299	DMDCITZN - Citizenship status	110.0