Alexia Marcous
Springboard Data Science Career Track
August 5th2019 Cohort

Capstone Project 1 – Predicting Hypertension

Data Wrangling

Data Source

The National Health Interview Survey (NHIS) ranks as one of the largest surveys conducted annually by the U.S. government. On average, the survey covers about 100,000 people in about 42,000 households each year, sampling the civilian, non-institutionalized population (this excludes anyone who is incarcerated, residents of nursing homes and members of the armed forces living in barracks).

The Integrated Public Use Microdata Series (IPUMS) is part of the Institute for Social Research and Data Innovation at the University of Minnesota and houses the world's largest collection of individual-level census and survey data. IPUMS collects, preserves and harmonizes data from the NHIS and provides easy access to this data with enhanced documentation. They make their data freely available at ipums.org.

Feature and Sample Selection

80 features from the NHIS were selected on the basis of 1) availability since 1997, when these features were consistently available, and 2) relevance to the project topic, predicting hypertension. 22 samples representing yearly surveys taken from 1997 to

2018 were selected. These extract parameters resulted in a 449.2mb csv file containing 2,061,980 rows.

Reviewing the dataset revealed that an additional 15 features were “auto-selected” in the extract. They contained data such as a unique identifier for each row, an ID number that could be used to link to the original NHIS and sample weights. Since this information was not needed for this project, these 15 features were dropped from data extract parameters, leaving the originally-selected 80 features.

Initial Review of Missing/Unusable Values

A search for missing data in the resulting dataset revealed that one feature was missing from 1997 but was available since 1998. Another feature was actually only collected since 2013. These 2 features were dropped from data extract parameters, leaving 78 features.

As important as considering missing values was the presence of positive responses, which would contribute to the ability to predict hypertension. The reasoning here was that if there were very few positive responses to a particular question, little could be gleaned from its impact on whether a person had hypertension or not. For example, the “Activity limitation from: Alcohol/drug problem” column had no missing values, but only 346 of the responses in the 2,061,980 rows indicated that there was any activity limitation. Many features were found to have very limited numbers of positive responses and 20 features with less than 10,000 usable responses were eliminated, leaving 58 features.

Pandas Profiling, 1st Iteration

This dataset with 58 features and 2,061,980 rows was then evaluated using Pandas Profiling – a library created for exploratory data analysis that can produce a report of individual feature evaluation, summary statistics, correlations and warnings regarding anomalies and missing data. The report finished in 30 hours and 21 minutes, but produced a large amount of very helpful information, which would have taken longer to program and produce from scratch!

Of particular interest was the clarification of the distribution of survey responses to yes/no questions over a greater possible number of responses then yes/no. For example, the target feature, hypertension, had 6 possible values, with this distribution:

Data from Pandas Profiling for Target Variable on dataset with 2 million rows:

	Count	Frequency
0 – Not in Universe	1390284	67.4%
1 - No	470288	22.8%
2 – Yes	200354	9.7%
7 - Unknown-refused	406	< 0.1%
8 - Unknown-not ascertained	30	< 0.1%
9 - Unknown-don't know	618	< 0.1%

Updating data based on target feature – an invalid approach

With this information, and the focus on predicting a yes/no value for hypertension, an attempt was made to narrow the data down to 2 possible outcomes, yes or no. Values of 0 were combined with “No” and values of 7, 8 and 9 were dropped, resulting in this distribution of values and these results from a logistic regression model:

Target variable value counts:
0 1861626
1 200354

Confusion matrix:
[[363008 9013]
[28420 11955]]

Classification report:

	precision	recall	f1-score	support
0	0.93	0.98	0.95	372021
1	0.57	0.30	0.39	40375

accuracy			0.91	412396
macro avg	0.75	0.64	0.67	412396
weighted avg	0.89	0.91	0.90	412396

While these results were helpful in terms of being inspiring, they were very unhelpful in terms of validity. Further research into what "NIU" actually meant revealed that *"blanks in the original NHIS public use data files are converted to numeric values (usually beginning with a 0 or a 9, to indicate 'not in universe' cases) in IPUMS NHIS. The universe is the population at risk of having a response for the variable in question. In most cases, these are the households or persons to whom the survey question was asked, as reflected on the survey questionnaire. For example, employment variables do not include children, since the NHIS does not ask children about employment."* See Appendix A for the universe for the target variable, hypertension.

This meant that considering the 1,390,284 rows with a value of 0 for hypertension was completely invalid – because in these 1,390,284 cases, the question had not even been asked!

Filtering data based on target feature – a better approach

To better identify what rows actually were relevant, rows where the value of the target column was 0 (not in universe) or 8 (not ascertained) in this column were dropped, resulting in a dataset with 671,666 rows. This also reframed the approach from predicting a binary outcome to one of 4 possible outcomes, but with the vast majority falling into the yes/no categories:

1 470288 - no
2 200354 - yes

As expected, the F1 score dropped, even with an increase in the number of iterations in the logistic regression model, although there were improvements in the confusion matrix and the precision for a positive response, which was inspiring:

Confusion matrix:
[[86649 7165]
[26349 13966]]

Classification report:

	precision	recall	f1-score	support
1	0.77	0.92	0.84	93814
2	0.66	0.35	0.45	40315
accuracy			0.75	134129
macro avg	0.71	0.64	0.65	134129
weighted avg	0.73	0.75	0.72	134129

Pandas Profiling, 2nd Iteration

Given the reduction of rows to 33% of the original dataset and the impact on the prediction results, the Pandas Profiling report was produced again to better understand this new dataset.

To make the profiling report more understandable, a mapping for more descriptive names was created and the columns in the dataset were renamed.

Also, 30 was maximum readable number of features in the correlation heat maps, after which the axis labels overlapped and could not be deciphered. To make the heat maps readable, profiling was updated to run on batches of columns of a variable number, adding in the target column to each batch so it would be included in the correlations.

Impact on features

The new profiling report on the hypertension-focused dataset revealed interesting changes in prevalence for some features. For example, “ever told had angina pectoris” went from 67.4% NIU values in the 2m dataset to 0% in hypertension dataset. “Needed but couldn't afford dental care, past 12 months” went from 56.3% NIU to also having no zeros in the hypertension-filtered data. Others features changed in the opposite direction, for example “Ever told had a learning disability” went from 86% to 100% NIU.

Removing more features

Having learned the importance of truly understanding the data, careful examination of code book was conducted based on the distribution of values in the profiling report.

This proved to be extremely important, because while most features were coded as 0 for NIU, some features used different values to indicate that the question was not asked. For example, “Felt everything an effort, past 30 days” showed 23.7% zeros, but in looking at the common values on the report, 67.4% were coded as 6. The codebook revealed that 0 meant “none of the time” and 6 represented NIU, for a total of 91.1% unusable responses. Another example was “Days had 5+ drinks, past year” – which showed 12.8% zeros but most responses were coded as 996, which also represented NIU, and a total of 93% unusable responses. The same went for “Number cigarettes per day (current smokers)”, with no zeros but tallying 93.6% NIU with other coded values.

This thorough review resulted in dropping 11 additional columns due to a low percentage of useable data, resulting in 39 columns in the dataset. Appendix B contains this list of columns.

Data Types

Reading in the data as categorical then changing the few numeric types seemed like a more efficient way of assigning data types. However this resulted in resulted in only the Phik correlation being produced, or other correlations calculated for one feature or a subset of features. The following error was reported for each batch:

```
/Users/alexia/anaconda3/lib/python3.7/site-packages/pandas_profiling/model/correlations.py
:124: UserWarning: There was an attempt to calculate the cramers correlation, but this
failed.
To hide this warning, disable the calculation
(using `df.profile_report(correlations={"cramers": False}`)
If this is problematic for your use case, please report this as an issue:
https://github.com/pandas-profiling/pandas-profiling/issues
(include the error message: 'The internally computed table of expected frequencies has a
zero element at (0, 0).')
  correlation_name=correlation_name, error=error
```

Researching calculation of the Spearman correlation matrix revealed that the dataset.corr() call would ignore columns with non-numeric datatypes. Individual testing of individual columns proved otherwise, but reading in the dataset as categorical (or as str) produced this error:

ValueError: zero-size array to reduction operation minimum which has no identity

Research on this error resulted in references to clustering and leafs… and feeling out of my depth!

Datatyping was tested on one column, “Ever had chickenpox”, with 14.8% zeros in the full dataset. A dataset of 1,000 rows was created, with the column having all zero values. Calling dataset.corr() on this subset when the column was categorical revealed that it was not in the resulting matrix. When it was numerical, it was in the resulting matrix. This lead to the belief that a column will be excluded from analysis if it is categorical and has all zeros.

Research on the specific error message and from Pandas Profiling Github, revealed that another user reported the same issue, and that it was corrected by reverting back to 1.4.2 version of Pandas Profiling. This work was done using Pandas Profile version: 2.3.0. Reverting back that many versions did not seem to be wise.

Not assigning a datatype on creation of the dataframe resolves the issue, however Pandas Profiling shows that 20 of the 39 columns are interpreted as numeric rather than categorical.

There was no change in the F1 score from the logistic regression model whether the data was read as categorical or numeric, so for the time being and on this particular dataset, it is being decided that the datatyping does not matter.

To be continued

Continued examination of the dataset will occur throughout this project.

Appendix A

Universe for "Ever told had hypertension"

- 1974: Sample persons age 17+.
- 1976: All persons.
- 1982: Persons age 17+.
- 1983; 1985; 1988: Sample persons age 18+.
- 1984: Sample persons age 55+.
- 1989: Persons age 18+ who have ever been told they had diabetes (other than during pregnancy).
- 1990-1991: Sample persons age 18+.
- 1993: Half of sample persons age 18+ in quarters 3 and 4 (excluded from AIDS supplement).
- 1994: Half of sample persons age 18+ (excluded from AIDS supplement).
- 1997-2018: Sample adults age 18+.

Appendix B

Columns in dataset

ANGIPECEV	Ever told had angina pectoris
ASTHMAEV	Ever told had asthma
CANCEREV	Ever told had cancer
CIGSDAY	Number cigarettes per day (current smokers)
CPOXEV	Ever had chickenpox
DELAYCOST	Medical care delayed due to cost, past 12 months
DIABETICEV	Ever told had diabetes
EARNINGS	Person's total earnings, previous calendar year
EMPHYSEMEV	Ever told had emphysema
FREMEMYN	Any family member limited by difficulty remembering
FWALKYN	Any family members have difficulty walking without special equipment
FWKLIMYN	Any family member with work limitation due to health problem
HEARING	Quality of hearing without hearing aid
HEARTATTEV	Ever told had heart attack
HEARTCONEV	Ever told had heart condition/disease
HISPETH	Hispanic ethnicity
HYPERTENEV	Ever told had hypertension
INCFAM97ON2	Total combined family income
LANY	Has any activity limitation
LIVINGQTR	Type of living quarters
MARSTAT	Legal marital status
MOD10FWK	Frequency of moderate activity 10+ minutes: Times per week

POORYN	Above or below poverty threshold
RACEA	Main Racial Background
REGION	Region of residence
SEX	Sex
SMOKAGEREG	Age first smoked fairly regularly
SMOKESTATUS2	Cigarette smoking
SMOKEV	Ever smoked 100 cigarettes in life
STROKEV	Ever told had a stroke
STRONGFWK	Frequency of strengthening activity: Times per week
TYPPLSICK	Kind of usual place for medical care
ULCEREV	Ever told had ulcer
USUALPL	Has usual place for medical care
VIG10FWK	Frequency of vigorous activity 10+ minutes: Times per week
VISIONPROB	Has trouble seeing
YBARCARE	Needed but couldn't afford medical care, past 12 months
YBARDENTAL	Needed but couldn't afford dental care, past 12 months
YBARMEDS	Needed but couldn't afford prescription medicines, past 12 months
YBARMENTAL	Needed but couldn't afford mental health care, past 12 months
YEAR	Survey year