Predicting Mental Health using the BRFSS – CRISP-DM Analysis

Report Portfolio Assignment
Anja Tolpekina (5537584)

1. Business Understanding

1.1. Business Objectives

The Centres for Disease Control and Prevention (CDC) is a federal organization in the United States, tasked with the role of national public health agency. The organisation was founded in 1946 and since then its main goal has been to protect public health and safety.[1] To achieve this objective, the CDC endeavours to control and prevent disease, injury, and disability. Their focus is on infectious disease, food borne pathogens, and environmental health. The organisation studies non-infectious diseases, are as well, such as chronic diseases caused by obesity, smoking, sedentary lifestyles, and other risk factors.[2] In more recent years, the CDC has expanded their scope of health to include mental health in addition to physical health. With good reason, as depression is the leading cause of disability for individuals between the ages of 15 and 44 in the United States[3] and more than 50% of people will be diagnosed with a mental illness at some point in their life[4].

The CDC has always been a science-based organization and with the emergence of modern technologies they are becoming more data-driven as well. They have established, maintained, and shared several large datasets and survey systems. The Behaviours Risk Factor Surveillance System (BRFSS) that is used in this project, is the largest ongoing telephone health-survey in the world [5.] Findings from the BRFSS have been used for targeting and promoting healthy behaviour and it has proven to be a powerful tool.[6] Thus far, it has mostly been utilized in the prevention and understanding of physical disease, but it could also be used to further the understanding of mental health. Some interesting results have been published already. Women caregivers have been found to have significantly worse mental health than their male counterparts[7] and another study found that respondents with serious psychological distress were ten times as likely to receive treatment for mental health problems[8].

However, no comprehensive study has been published using the BRFSS to examine the risk factors for poor mental health in general as of this date. The objective of this project is to gain insight into the factors that are associated with mental health disorders by studying this dataset. Most studies that have been published on this subject, have had a priori hypotheses about risk factors or subpopulations at risk. While these uses certainly have their merits, the BRFSS data gives an opportunity to create a better understanding of which behavioural risk factors, socio-demographic factors and other possible factors most strongly influence mental health and which do not. Therefore, in this project, no assumptions will be made beforehand, to get an unbiased overview of the relationship between mental health and other information available in the BRFSS. The findings could be used for several purposes. For example, the government can allocate funding and focus interventions to target populations that are at higher risk for developing mental health illnesses. In addition, greater understanding of the causes of poor mental health, can improve treatment.

1.2. Situation Assessment

The inventory of resources for this project consists of the data that is downloaded from Kaggle[9]. This contains individual datafiles for the BRFSS between the years of 2011 and 2015. In addition, a text-file containing information on each variable in the files. The original datasets are available in SAS format on the site of the CDC[10]. These have been converted into CSV-files and posted on Kaggle. While it is generally good practice to use the original dataset, due to time constraints of this project, the converted datafiles used instead of the original ones. To clean and analyse the data, Jupyter Notebook 6.4.8 is used.

The largest constraint for this project is time, as there are 84 hours allocated to it. Therefore, not all steps of the CRISP-DM process can be carried out to the full extent of how they are conceptualized[11]. However, this is sufficient time to go through the main steps and extract some useful information from the data. As for other potential constraints, the data used for this project is entirely open source and so are the programmes used for analysis.

1.3. Data Mining Goals

In this project, the aim is to build a comprehensive model to help understand the risk factors for poor mental health and to provide possible targets to help improve and prevent mental health problems among the population. In order to achieve this goal, the relevant risk factors in the BRFSS will be utilized to build a model that predicts the respondent's mental health status. The outcome variable that will be used to quantify mental health status, is whether an individual is receiving medicine or treatment from a health professional for an emotional problem. Such a model has the potential to give us insight into which factors contribute strongly to mental health and which do not. Ideally, to achieve this goal, a model that can be easily interpreted should be chosen.

Previous studies that used data mining to predict mental health status on different data sets, have found models with accuracy scores ranging between 72-82%[12–14]. The goal of this project is to find a model that can perform at least as well. Therefore, the first success criterium is to create a model that can predict whether an individual is receiving mental health treatment with an accuracy of at least 75%. The second success criterium is to identify the factors that have the strongest relationship with the need for mental health treatment.

1.4. Project Plan

To achieve the determined data mining goals, the following steps are taken. First, the dataset is extracted from the source and loaded into Jupyter Notebook. Following, the data is checked for possible quality issues and cleaned. Then, the data is prepared for modelling. For this project, two different classification models are made, and their predictive performance is compared to ascertain which one is the best fit. Both a decision tree and random forest model are built on a part of the data and afterwards tested on a different part to ensure reproducibility and prevent an overfitted model. The model outcomes are evaluated using different accuracy measures and the best one is selected to be interpreted and the outcome discussed. Lastly, the project is deployed by publishing a GitHub repository with the code and the project report.

2. Data Understanding

2.1. Data Collection

The data used in this project was downloaded from Kaggle in CSV-format and loaded into JupyterLab. The background of the data is discussed in Section 1. There are annual BRFSS datasets available on the site of the CDC as recent as 2020. Unfortunately, these were published in SAS format. While it would have been preferable to use the most recent dataset to avoid building a model on outdated findings, no efficient way could be found to convert these files into a suitable format for Jupyter Notebook. The packages Pyreadstat[15], and Pandas[16] can be used to open SAS files, but manually checking whether the 330 variables have been converted correctly would be too time consuming and beyond the scope of this project. No other problems were encountered in data acquisition.

2.2. Data Description

The dataset is in CSV-format and consists of 330 columns and 441,456 entries. Every entry represents one respondent and 375,059 (85%) of them have completed the interview. The majority of the columns has integer values and the remaining seven are character values. Most of the integer columns represent coded data and are therefore either ordinal or nominal data, but there are also a few interval and ratio variables such as weight, height, and number of children. Four of the variables that have character values, hold information about the date of the interview and could easily be transformed into integer values. The remaining three character variables have ordinal values.
For further description of the dataset, the codebook published by the CDC can be consulted[17]. This file is included in the repository.

2.3. Data Exploration

Some initial data exploration was done in order to make a data selection, but due to the size of the dataset and the time constraints, data exploration was mostly performed after selection (which is describe below in section 3.1 Data Selection). This was done by visually inspecting the data, creating figures such as histograms and scatter plots and using functions to examine the individual values of variables. During the initial data exploration, it became clear that there was a large percentage of missing values. While this seemed at odds with the fact that 85% of respondents had completed their interview, this is because the BRFSS interview consists of both core modules, which are included in every interview across different states, in addition to more specific modules that are included in certain states but not in others. After the data selection, every variable was inspected to ensure good data quality. This is described further in section 3.2 Data Cleaning.

3. Data Preparation

3.1. Data Selection

The first step in the data preparation was to select the data that would be used to build the model. Decisions within this process were made based on the relevance of the variables to the research question and on the data quality. From the 330 variables and 441.456 records a subset was made consisting of 64 variables and 5957 records. In the following section, this process will be described in more detail.

Firstly, the BRFSS interview is not the same in every state. As is mentioned in the previous section, there are certain core modules, which are included in every interview, in addition to more specific modules that are included in certain states but not in others. One of these modules is the 'Anxiety and Depression' module which consists of questions such as: 'Over the last 2 weeks, how many days have you had little interest or pleasure in doing things?'. As the goal of this project is to get a better understanding of the risk factors for poor mental health, the variables corresponding to this module were all included in the data selection. This particular module was only included in the interviews conducted in the state of West Virginia. In total, there were 5,957 records obtained in this state and all records from other states were excluded as they did not contain the 'Anxiety and Depression' module. This decision prompted the removal of the 'State' variable as well, as this became uninformative.

Following, there were 12 variables which only contained the value 'HIDDEN'. For different reasons the data within these columns was not displayed in the published dataset. These variables where all excluded, in addition to variables that were judged to hold little meaningful information regarding the data mining goal. Examples are variables regarding the date of the interview and questions such as 'Is this the correct phone number?'. Variables with a high number of missing values were excluded for practical reasons. These variables were generally questions that belonged to specialized modules that were not assessed in West Virginia or follow-up questions such as 'Are you currently taking medicine for your high blood pressure?' following the question 'Have you ever been told that you have high blood pressure?'. Furthermore, there was a large number of calculated variables which held practically the same information as other variables in slightly different forms ('Reported weight in pounds' and 'Reported weight in kilograms'). In such cases, the variable that was the most informative or practical was chosen and the others were excluded to avoid multicollinearity. These steps collectively resulted in a large decrease in the data volume. While, this is often undesirable, in this case it had the benefit of less technical constraints, as computing a model on the original dataset would have been very time consuming.

3.2. Data Cleaning

To clean the dataset, the following factors were considered: missing data, duplicate data, outliers, structure issues. Firstly, all records with missing data were excluded. Instances containing values which code for 'Refused' or 'Don't know' were excluded as well, because they were not informative in relation to the research question. This decision decreased the 5,957 records to 1766. Instances of missing data were checked whether they could be reasonably imputed, however, upon inspection this was not expected to produce valid results. As the number of records was still sufficient for modelling, the decision was made to exclude all instances containing missing values and to not attempt imputing any. In addition, the data was checked for duplicate instances but there appeared not to be any.

Furthermore, the data was checked for outliers and values that did not belong in the dataset. A large part of the selected variables was either categorical or ordinal. These were visually inspected for incorrect values or strange distributions with histograms and by printing all the unique instances for every variable. To examine the data in the interval variables, boxplots were made and inspected. A few variables had some visible outliers but upon closer inspection, these were assessed to be not out of the scope of realistic situations (such as being 209 cm tall or eating seven pieces of fruit a day). Therefore, they were kept in the dataset for analysis.

Lastly, there were two structural issues that had to be fixed. In twelve columns that appeared to be interval scale, the number 88 was used to code for zero. For example, 'How many children less than 18 years of age live in your household?' would have the possible values 1 – 87: number of children, 88: None. In these cases, the value 88 was replaced by 0 to enable valid values for modelling. In addition, 'Yes/No' variables were coded as Yes: 1, No: 2. This was changed to No: 0, Yes: 1 for the same reason.

3.3. Data Construction

There was little data construction necessary as the dataset was very comprehensive and already provided some calculated variables that were helpful, such as BMI and whether the interviewed individual met the criteria to be a binge drinker. The only data construction performed, was One-Hot Encoding for the categorical and ordinal variables as this is required by most modelling techniques to yield accurate results. Dummy variables were created for the categorical and ordinal variables that had more than two possible values (as these were already encoded as 0/1).

After performing all the steps described above, the result was a clean dataset formatted in a Pandas dataframe and ready for modelling. The variables of the selected subset of data are shown in the Appendix and more details about these variables and all the other ones of the original dataset, can be found in the codebook pdf-file.

4. Modelling

4.1. Modelling Assumptions

The goal of the project is to gain more understanding of the risk factors for poor mental health. This will be achieved by building a model that can predict whether a person is receiving treatment for mental health conditions based on other factors. Furthermore, this model can give us insight into which risk factors are associated with the need for treatment. A classification model is suited for this kind of problem, as the goal is to predict class assignment. From the different kinds of classification algorithms, two were selected to use for modelling. Both the decision tree and random forest have different advantages and disadvantages in relation to this specific case. Decision trees are well equipped to handle both numerical and categorical data, do not require normally distributed data, and are very intuitive and easy to interpret. A random forest classifier fits multiple decision trees on different sub-samples of the dataset and uses the results thereof to build a more accurate model with less over-fitting. However, random forest can be biased when dealing with categorical variables and those make up a large part the dataset. For these reasons, the decision was made to build both models and compare their accuracy to determine which one is most suited to this particular case.

4.2. Test Design

The data will be split into a training, validation, and testing set to avoid overfitting (respectively 80%, 10%, and 10%). Furthermore, the training sample will be resampled because the target variable is not entirely balanced, as more people do not receive treatment for mental health problems than those who do. After fitting the initial decision tree on the resampled training data, the tree will be pruned. First, cost complexity pruning will be used to get the effective alpha values of subtrees. Then cross validation will be used to find the best value for alpha. This alpha value will then be used to create the definitive decision tree and the accuracy of this tree will be tested with the validation subset.
In addition, a random forest classifier will be trained on the training sample and evaluated with the validation subset. After tuning the parameters of both models and selecting the best one, the testing set will be used to evaluate the performance of the chosen model.

4.3. Model Description

To build the models the following parameter settings were used. Firstly, the maximum value for alpha was excluded from the list of potential alpha values that were generated by cost complexity pruning, as this would result in a decision tree with no leaves and only a root. Furthermore, when selecting the optimal value for alpha, the values below 0.01 were excluded as well. While these lower alpha values gave very high within training sample accuracy, they were overfitted and did not perform well on the testing data. For the cross validation, five folds were used because increasing the number of folds did not improve the results and slowed down the processing time. Lastly, for the random forest the value for n_estimators (number of trees to build) was set to 300, as a higher value did not yield a significantly more accurate model.

4.4. Model Assessment

To assess the models, confusion matrices were made, and different performance values were calculated as can be seen in Table 1. While the decision tree yields a much better sensitivity and balanced accuracy, the random forest classifier resulted in a higher specificity and accuracy. Both models performed well and met the data mining goal. However, for this particular project the decision tree is more suitable as false negatives have a higher cost in real life than false positives. This will be expanded upon further in Section 5.

Table 1

Performance measures of the decision tree and random forest models.

Performance measure	Decision Tree	Random Forest
Sensitivity	0.981	0.750
Specificity	0.736	0.888
Accuracy	0.808	0.847
Balanced Accuracy	0.858	0.819

The decision tree decided upon the class label based on the following three factors: 1. Whether the individual had ever been diagnosed with depression. 2. Whether the individual had ever been diagnosed with anxiety. 3. How many days over the last 2 weeks the individual has felt down, depressed, or hopeless.

5. Evaluation

The goal of this project was to build a model to help understand the risk factors for poor mental health and to provide possible targets to help improve and prevent mental health problems among the population. More specifically, the success criteria for this goal were to create a model that can predict whether an individual is receiving mental health treatment with an accuracy of at least 75% and to identify the factors that have the strongest relationship with the need for mental health treatment. To achieve this result, the BRFSS dataset was extracted, examined, and cleaned. Following which, a decision tree and random forest model were created.

Both models had an accuracy of 80% or higher, thus the first success criterium has been met. The decision tree was selected as the best model for this project, as it had a better balanced accuracy and sensitivity, and this outweighed the higher specificity of the random forest. It is more important to identify all the different risk groups, because the goal of this project is to enable prevention. The cost of missing a group outweighs the cost of including too many, as prevention measures do not have negative consequences for those who are not in need of them.

The performance on the second success criterium is less straightforward. While the decision tree did provide the variables which were used to predict whether an individual received mental health treatment, these are not very insightful upon closer inspection. The variables with the biggest influence on the decision tree where whether an individual had ever been diagnosed with depression or anxiety. It is not surprising that these individuals are receiving medication or other treatment for emotional problems. Unfortunately, this does not help identify potential target groups for mental health improvement, however, it does affirm the validity of the model. Another issue with the model is that it tests whether an individual is receiving treatment and not whether they are in need of it. There could be different reasons for someone in need of treatment, to be unable or unwilling to receive it. This is somewhat accounted for in the model as access to insurance is incorporated, however, it would have been better to use a different target value.

6. Deployment

To deploy the project, GitHub will be used. This is a popular platform used for software development and version control. The data script, codebook, and this report will be posted so they can be found and used by others. The GitHub repository will also include a README file with prerequisites and instructions on where to download the data. As this project was part of an assignment, there is no plan for maintenance and upkeep. If another GitHub user runs into a problem, however, the platform can be used to leave comments and flag issues so they can be resolved.

References

1. Centers for Disease Control and Prevention | HHS.gov. Accessed May 27, 2022. https://web.archive.org/web/20200410150453/https://www.hhs.gov/about/budget/fy-2020-cdc-contingency-staffing-plan/index.html

2. Mission, Role and Pledge | About | CDC. Accessed May 27, 2022. https://www.cdc.gov/about/organization/mission.htm

3. Mental Health Awareness|Diseases|Resources|Genomics|CDC. Accessed July 2, 2022. https://www.cdc.gov/genomics/resources/diseases/mental.htm

4. KESSLER RC, ANGERMEYER M, ANTHONY JC, et al. Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry*. 2007;6(3):168. Accessed July 2, 2022. /pmc/articles/PMC2174588/

5. CDC - BRFSS. Accessed May 27, 2022. https://www.cdc.gov/BRFSS/

6. CDC - About BRFSS. Accessed May 27, 2022. https://www.cdc.gov/brfss/about/index.htm

7. Edwards VJ, Anderson LA, Thompson WW, Deokar AJ. Mental health differences between men and women caregivers, BRFSS 2009. *http://dx.doi.org/101080/0895284120161223916*. 2016;29(5):385-391. doi:10.1080/08952841.2016.1223916

8. Dhingra SS, Zack MM, Strine TW, Druss BG, Berry JT, Balluz LS. Psychological distress severity of adults reporting receipt of treatment for mental health problems in the BRFSS. *Psychiatric Services*. 2011;62(4):396-403. doi:10.1176/PS.62.4.PSS6204_0396/ASSET/IMAGES/LARGE/PSS6204_0396_FIG004.JPEG

9. Behavioral Risk Factor Surveillance System | Kaggle. Accessed May 27, 2022. https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system?resource=download

10. CDC - BRFSS Annual Survey Data. Accessed May 27, 2022. https://www.cdc.gov/brfss/annual_data/annual_data.htm

11. Wirth R, Hipp J. CRISP-DM: Towards a Standard Process Model for Data Mining.

12. Laijawala V, Aachaliya A, Jatta H, Pinjarkar V. Classification Algorithms based Mental Health Prediction using Data Mining. Published online July 10, 2020:1174-1178. doi:10.1109/ICCES48766.2020.9137856

13. Tate AE, McCabe RC, Larsson H, Lundström S, Lichtenstein P, Kuja-Halkola R. Predicting mental health problems in adolescence using machine learning techniques. *PLOS ONE*. 2020;15(4):e0230389. doi:10.1371/JOURNAL.PONE.0230389

14. Hao B, Li L, Li A, Zhu T. Predicting Mental Health Status on Social Media A Preliminary Study on Microblog. *LNCS*. 2013;8024:101-110. Accessed July 3, 2022. http://ccpl.psych.ac.cn:10002

15. Roche/pyreadstat: Python package to read sas, spss and stata files into pandas data frames. It is a wrapper for the C library readstat. Accessed July 3, 2022. https://github.com/Roche/pyreadstat

16. pandas - Python Data Analysis Library. Accessed July 3, 2022. https://pandas.pydata.org/

17. CDC - 2015 BRFSS Survey Data and Documentation. Accessed June 20, 2022. https://www.cdc.gov/brfss/annual_data/annual_2015.html

Appendix

Table 2

The column names that were included in the selected dataset on which the model was build and their description.

Column Name	Description
GENHLTH	General Health
PHYSHLTH	Number of Days Physical Health Not Good
MENTHLTH	Number of Days Mental Health Not Good
POORHLTH	Poor Physical or Mental Health
HLTHPLN1	Have any health care coverage
PERSDOC2	Multiple Health Care Professionals
MEDCOST	Could Not See Doctor Because of Cost
CHECKUP1	Length of time since last routine checkup
BPHIGH4	Ever Told Blood Pressure High
TOLDHI2	Ever Told Blood Cholesterol High
CVDINFR4	Ever Diagnosed with Heart Attack
CVDCRHD4	Ever Diagnosed with Angina or Coronary Heart Disease
CVDSTRK3	Ever Diagnosed with a Stroke
CHCSCNCR	Ever told you had skin cancer?
CHCOCNCR	Ever told you had any other types of cancer?
CHCCOPD1	Ever told you have chronic obstructive pulmonary disease, emphysema or chronic bronchitis?
HAVARTH3	Told Have Arthritis
ADDEPEV2	Ever told you had a depressive disorder
CHCKIDNY	Ever told) you have kidney disease?
DIABETE3	Ever told) you have diabetes
SEX	Respondents Sex
MARITAL	Marital Status
EDUCA	Education Level
RENTHOM1	Own or Rent Home
VETERAN3	Are You A Veteran
EMPLOY1	Employment Status
CHILDREN	Number of Children in Household
INCOME2	Income Level
INTERNET	Internet use in the past 30 days?
QLACTLM2	Activity Limitation Due to Health Problems
USEEQUIP	Health Problems Requiring Special Equipment
BLIND	Blind or Difficulty seeing
DECIDE	Difficulty Concentrating or Remembering
DIFFWALK	Difficulty Walking or Climbing Stairs
DIFFDRES	Difficulty Dressing or Bathing
DIFFALON	Difficulty Doing Errands Alone
HIVTST6	Ever tested HIV
SXORIENT	Sexual orientation or gender identity
TRNSGNDR	Do you consider yourself to be transgender?
ADPLEASR	Days had little pleasure doing things
ADDOWN	Days felt down, depressed or hopeless
ADSLEEP	Days had trouble with sleep
ADENERGY	Days were tired or had little energy
ADEAT1	Days ate too little or too much

ADFAIL	Days felt like failure or let family down
ADTHINK	Days had trouble concentrating
ADMOVE	Days talked to move slower or faster than usual
MISTMNT	Receiving medicine or treatment from health pro for emotional problem
ADANXEV	Ever told you had an anxiety disorder
_ASTHMS1	Computed Asthma Status
_DRDXAR1	Respondents diagnosed with arthritis
_ RACE	Computed Race-Ethnicity grouping
_ AGE80	Imputed Age value collapsed above 80
HTM4	Computed Height in Meters
WTKG3	Computed Weight in Kilograms
_BMI5	Computed body mass index
_SMOKER3	Computed Smoking Status
_RFBING5	Binge Drinking Calculated Variable
_DRNKWEK	Computed number of drinks of alcohol beverages per week
_FRUTSUM	Total fruits consumed per day
_VEGESUM	Total vegetables consumed per day
STRFREQ_	Strength Activity Frequency per Week
_PACAT1	Physical Activity Categories
_PAINDX1	Physical Activity Index

Note. For more details about these variables, see the codebook pdf-file.