Distributed Data Analysis

CS5811

2021-2022

Shivangi Dubey
2039934
Brunel University London

Data Description

Column name	Description	Data Type
id	Unique Identifier	Char
gender	"Male", "Female" or "Other"	Char
age	age of the patient	Num
hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension	Factor
heart Disease	"No" or "Yes"	Factor
ever_married	"No" or "Yes"	Factor
work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"	Factor
Residence_type	"Rural" or "Urban"	Factor
avg_glucose_level	average glucose level in blood	Num
bmi	body mass index	Num
smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"*	Factor
Stroke	1 if the patient had a stroke or 0 if not	Factor

Research Question

To identify factors from everyday life and habits that may impact heart health and determine the likelihood of a stroke in patients of different genders and age groups using the same.

Data Preparation and Cleaning

Steps of Data Preparation are enumerated below.

- Checked the data-types of different columns.
- Changed the data-types of columns to suitable types.
- Checked the percentage of missing and unknown variables.

- Imputed the missing bmi values with medians of male and female persons.

(The ‘bmi’ values of males and females have a considerable difference as evidenced by the medians of both and hence we impute them separately)

- Converted the ‘Other’ gender type to ‘Female’, since it’s an outlier.

(It would have been a valid value in any other case however here there is only one row in almost 5000 rows, and hence is taken as ‘Female’ instead as it is the most frequent gender type in the data).

- We change the ‘Unknown’ smoking status to ‘never smoked’ as ‘never smoked’ is the most frequently occurring value and it makes sense not to delete the otherwise sufficient rows.

Exploratory Data Analysis

The Implementation of the EDA has been divided into two parts

1. For the numerical columns i.e:
   A. Average glucose level
   B. BMI
   C. Age
      - Univariate Exploratory Data Analysis
      1. The statistic summary of all the numerical variables.
      2. The outliers are identified for the above from the box plots.
      3. Histograms of all three to analyse the distribution of each individually.
      - Bivariate Exploratory Data Analysis
      1. The numerical columns are plotted in the form of box plots against the target variable i.e. stroke to see the distribution of the numerical columns amongst people who had had a stroke and people who hadn’t.
      2. A correlation plot is plotted to see correlation amongst the numerical columns
      3. Scatterplots are plotted for all the numerical columns at the same time to study any kinds of trends that might emerge.
2. For the categorical columns as enumerated in the list below:
   A. Gender

B. Hypertension
C. Heart_Disease
D. Ever_married
E. work_type
F. Residence_type
G. smoking_status

- Univariate Exploratory Data Analysis

1. The dataset is filtered to include only rows with people who have experienced a stroke. Then a bar plot is plotted to show the distribution of different levels in each of the categorical columns for those people with stroke history.

- Bivariate Exploratory Data Analysis

1. Colour-coded stacked bar charts are plotted to see the distribution of different levels of each categorical column amongst people with or without a stroke history.
2. Chi Square tests are done to see if there is a considerable difference between the expected and the observed and if it's due to chance or a relationship between the categorical columns and the target variable.

Machine Learning prediction

Implementation of Neural Network

Data Preparation The steps for Data Preparation for building a fully-connected artificial neural network have been enumerated below:

1. A minimax function to normalise all the numerical columns to bring the range of each column to [0,1].
2. The categorical columns are one-hot encoded to feed them to the neural network training phase.

Data Splitting

1. The dataset is then split into training and test dataset using random sampling and stratified sampling.

Training the Fully Connected Artificial Neural Network:

3. A fully connected neural network is trained for both types of sampling (random and stratified).

High Performance Computational Implementation

The technique used is pyspark and Hadoop's File Distribution System.

- The data cleaned file was put on a directory named 'input' on HDFS.
- Pyspark was initiated to open Jupyter Notebook.
- The code in Pyspark was written using pandas to read file from HDFS.
- A Random Forest Classifier was built which resulted in the accuracy of 94.8%
- The precision and recall value were 1 and 0.
- The Area under the Roc Curve was 85.5% for test data.
- Consequently a GBT (Gradient Boosted Tree Model) was built and a AUC score 82.8% was observed for test data.
- Cross-validation was attempted multiple times but due to the size of the data and limited processor capabilities, it did not converge.
- Finally the model was saved and connection to spark was disconnected.
- This was followed by disconnection with HDFS.
- The above steps were in accordance with files provided during lab sessions.

Performance Evaluation and Comparison of Methods

Results of both kinds of sampling (random and stratified) are compared using the following metrics

- Confusion Matrix
- Accuracy
- ROC Curve

Model Implemented	Specificity	Precision	Sensitivity	Recall	Area Under the curve	Accuracy	F1 Score
Neural Network (With random sampling)	0.97	0.7017	0.909	0.909	0.535	0.9764	0.7920
Neural Network (with stratified sampling)	0.96	0.55	0.89	0.89	0.4399	0.96	0.6839
Random Forest	1	1	0	0	0.855	0.954	0
Gradient Boosted Tree	0.991	4.0	0.043	4.0	0.828	0.948	4.0

- Specificity is great for all the models, but best for Random Forest.
- Sensitivity is good for the Neural Networks (better for random sampled data) but 0 for Random Forest and Gradient Boosted Tree.
- Accuracy is good for all models but best for Neural Network trained and tested on random sampled data.
- Neural Networks (trained and tested on randomly sampled data) perform the best even with great class imbalance.
- Stratified sampling is guaranteed to perform better in most cases, but here it is inferior to random sampling and the reason could be class imbalance. Stratified oversampling could be employed to make performance better.

Discussion of Findings

1. The data-types of numerical and categorical variables were adjusted.
2. The missing values of BMI and smoking_status were examined. Following approaches were adopted by the group members
   - The genders were Male,Female and Other. But the frequency of Other was 1, so it was replaced by the most frequently appearing gender (Female).
   - The median of BMI for males and females were found different but the significance was low(using ANOVA).
   - The median of BMI for people with different relationship history was found significantly different hence this status was used to impute missing values.
   - The smoking status 'Unknown' values were discussed and replaced with the frequently occurring status. (never_smoked).
3. The chi square tests were conducted between all possible combinations with smoking_status. Three were found to have a significant effect on smoking status:
   1. Hypertension
   2. work_type
   3. ever_married
4. Mostly all of the categories had a smoking_status as 'never-smoked' hence it seemed obvious to replace the 'Unknown' with it.
5. Useful insights from visualisation
   1. There are more males in the study as compared to females.
   2. The distribution of BMI is slightly skewed.
   3. The histogram of Age suggests a very weak normal distribution.
   4. Box plots give us an idea of the outliers however they are retained as getting rid of them might result in major loss of information. Hence they are classified as interesting outliers. (Aguinis et. al.)

5. Less people with hypertension and heart disease are susceptible to a stroke than others. It could be due to precautions taken by their sufferers.
6. More married people are seen to have a stroke than who have never been married.
7. Most people are employed privately, amongst those who have had a stroke.
8. More people who reside in urban areas have had a stroke history.
9. Most people have never smoked.

6. Chi-square test between categorical explanatory variables and target stroke variable have the conclusions that
   - hypertension shows a strong relation with stroke variable
   - Heart disease is another one that shows promising impact on the same.
7. Conclusion from box plots
   1. Not much of a difference in bmi between people with and without stroke history.
   2. The range of average glucose levels for the 3rd quartile for people with a stroke history is more than the people without it.
   3. Median for average glucose levels is not significantly different.
   4. Age of people with stroke history if higher than of people without it
   5. Logistic Regression is done to confirm the significance of the difference between the ages of people with and without stroke history. (p-value found is very low)
8. Correlation plot does not show strong correlation between numerical attributes.
9. Trends are not identified in the scatter plot between numerical attributes.
10. Principal component Analysis can't be performed on the data as the dataset has categorical independent variables along with numerical continuous independent variables since it does not perform well with a combination of both. (Feizi, S., & Tse, D. (2017))
11. The neural network, random forest and gradient boosted models are implemented.

12. The neural network did not converge for a threshold of 0.1, and took a long while and sometimes bizarre results for a threshold value of 0.2 and 0.3. Hence a value of 0.4 was opted for proper convergence and results with two labels , henceforth avoiding garbage values in label predictions.
13. The precision, recall, specificity and sensitivity values of the models:
    - Random Forest
    - Gradient Boosted Tree

    are absurd due to a class imbalance in the dataset. It could be resolved by adopted a number of techniques like:
    - Oversampling
    - Undersampling
    - Feature Selection

    (The above has been enumerated in the paper by Kotsiantis et.al.)

Note:

- The group members have only shared the selective model evaluation information for the implemented models, which I don't think is sufficient for the analysis of the model. Nonetheless, the following results were shared:

Name of the team member	Machine learning model implemented	Accuracy	Recall/ Sensitivity if provided	Specificity if provided
Eamonn Nemeh	Random Forest	89%	12%	-
Scott Pearson	Decision Trees	95%	100%	0
Sanket Tembekar	Decision Tree Logistic Regression	94%	-	-
Shivangi Dubey	Neural Network	97.64%	90.9%	97%

- The only model that can be compared is the Decision Trees implemented by Scott and that does not perform well in terms of specificity which is significant when it comes to medical

data. The absurd specificity could be due to class imbalance in the data.

- PCA was done as a unsupervised learning technique by one of the group members (Eamonn Nemeh), code was shared too, however no particular insights were obtained from the implementation.

References

- Kotsiantis, Sotiris & Kanellopoulos, D. & Pintelas, P.. (2005). Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering. 30. 25-36.
- https://libguides.library.kent.edu/spss/onewayanova
- Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. In Organizational Research Methods (Vol. 16, Issue 2, pp. 270–301). SAGE Publications Inc. https://doi.org/10.1177/1094428112470848
- https://www.sheffield.ac.uk/polopoly_fs/1.575550!/file/What_test_terminology.pdf
- Muhammed Kürşad Uçar, Majid Nour, Hatem Sindi, Kemal Polat, "The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets", Mathematical Problems in Engineering, vol. 2020, Article ID 2836236, 17 pages, 2020. https://doi.org/10.1155/2020/2836236
- Feizi, S., & Tse, D. (2017). Maximally Correlated Principal Component Analysis. http://arxiv.org/abs/1702.05471
- Kaggle.com. 2022. Stroke Prediction Dataset. [online] Available at: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset> [Accessed 10 February 2022].

- Rdocumentation.org. 2022. Home - RDocumentation. [online] Available at: <https://www.rdocumentation.org/> [Accessed 25 April 2022].

Data Management Plan for Research Students

1. Overview

Researcher: Eamonn Nemeh, Lakshmi Nivas Talluri, Sanket Tembekar ,Scott Pearson, Shivangi Dubey,
Project title: Prediction of stroke using lifestyle and health history.
Project duration: 3 months
Project context: Distributed Data Analysis using Machine Learning and HPCI techniques.

2. Defining your data/research sources

2.1 Where will your data/research sources come from?

It is obtained from an online data repository called "Kaggle".

The initial data is a .csv file and has 5110 observations and 12 attributes. The data consisted of textual i.e., numerical and categorical values.

2.2 How often will you get new data?

The data was only retrieved once and was not updated throughout the analysis.

How many experiments per week?

(Irrelevant.)

How will this change over time?

(Irrelevant.)

2.3 How much data/information will you generate?

Try to state this in kB/MB/GB

The original data is 69kB.

How much have you got so far?

Each group member is working with same sizes of data.

Try to estimate how this will grow for the rest of the project

They remain same throughout the process.

2.4 What file formats will you use?

What software is required to access the data? Are free/open alternatives available?

The data can be accessed online at Kaggle:

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

It's public and freely available to anyone.

What type of data does each format hold?

The data is a single .csv file.

3. Organising your data

3.1 How will you structure and name your folders and files? *Are there any set or recommended standards in your discipline?* The modified datasets were also stored and used in .csv format.
2. What additional information is required to understand each data file? *What would you need to know to reproduce the results from this data?* It is a generalised dataset with data about life and health history. Not much domain knowledge is expected.
3.3 What different versions of each data file or source will your create? How will you differentiate between different versions, for example do you plan to use files names to denote different versions, e.g. V1, V1.1, V2 etc? Each member used different names

4. Looking after your data

4.1 Where will you store your data? Laptop OneDrive.
4.2 How will your data be backed up? *How many copies?* One duplicated copy of original data. *Where are they stored?* *Laptop*. OneDrive. *How often are copies updated?* The original data is not updated so we can skip back to it any time. The data is adjusted for analysis and is updated after the processing on it is completed. These data files are stored in the project's folder on the laptop.
3. **How will you test whether you can restore from your backups?** The backups are not moved and frequently checked to see if they are not damaged or corrupted.

5. Sharing your data

5.1 Who owns the data you generate? *Is it you? Your supervisor? The University? An external partner?* The added columns in the dataset and particularly associated with person who performed the analysis and is owned by him i.e., each member of the group owns the data that is generated as a result of their analysis,
5.2 Who else has a right to see or use this data? *Your supervisor, collaborators, group members?* The supervisor, board of examiners and group members.
5.3 Who else should reasonably have access to this data when you share it? *Readers of your published work? The General Public?* The general public.
5.4 What should/shouldn't be shared and why? *Consider any ethical, legal or commercial restrictions that may affect what you share, how you share it and who you share it with?* The original dataset is public so there shouldn't be any ethical, legal or commercial restrictions on the data generated through analysis on the original one.

6. Archiving your data

1. What should be archived beyond the end of your project?

Everything? Just what you used for your thesis?

The project will be uploaded on WiseFlow by 25th of April. There isn't any need to further store the data after this deadline. Each member can store the project

6.2 For how long should it be stored?

EPSRC guidelines say "10 years from the date of last access"

None of the group members is planning to publish their analysis performed using this data, that's why there is no need to store it longer than the project's duration.

6.3 When will files be moved into the data archive/repository?

April 25th , 2022

6.4 Where will the data be stored?

Wiseflow, Github

6.5 Who is responsible for moving data to the data archive and maintaining it?

Students individually are responsible for the same.

6.6 Who should have access and under what conditions?

It is an open dataset.

7. Executing your plan

1. Who is responsible for making sure this plan is followed? *You may wish to discuss and agree this with your supervisor* The group members.
7.2 How often will this plan be reviewed and updated? *You may wish to discuss and agree this with your supervisor* The plan was discussed between the group before the analysis was initiated and was reviewed by the end of the project.
7.3 What actions have you identified from the rest of this plan? *List them here with timescales* Data Search and finalizing: Feburary10th , 2022 Data knowledge discovery: Feburary 25th , 2022 EDA result sharing: April 1st, 2022 Machine Learning and HPCI result sharing: April 15th, 2022 Discussing DMP: April 20th, 2022
7.4 What further information do you need to carry out these actions? *Where can you find this information?* Meetings online and offline. *Who might you be able to ask?* *Professors and group members.*

Notes on completing this form

- Type as much (or as little) as you feel you need to into each box: it will expand to accommodate what you write;
- You can leave or remove the prompts in grey once you're done;
- For help with completing this DMP, please contact researchdata@brunel.ac.uk

Authorship Contribution:

The group members collectively shared different datasets and among them, one was chosen for this assignment.

• Dataset Selection was done by Eamonn, Nivas, Sanket, Scott, Shivangi.

Data Cleaning was done by Eamonn, Sanket, Scott,

Exploratory Data Analysis was done by Scott, Nivas, Shivangi and Eamonn.

Model implementation is summarised below:

• Decision Tree : done by Scott and Sanket.

• Logistic Regression done by Sanket.

• Neural Networks done by Shivangi.

• Random Forests done by Eamonn.