Assignment 1

Leondis Evans
username@gatech.edu

Abstract—

1 DATASETS

1.1 Mortgage Loan Data

This dataset is a collection of common data points used to determine if a loan request should be approved or denied.

What makes this an interesting dataset is the challenge to determine the proper weight of majority features without making minor features non relevant. A perfect example of this would be credit history. Credit history is a majority factor that is predictive of how a person is likely to handle current and future financial responsibilities.

1.1.1 *Features*

Name	Description
Loan_ID	Id for the loan request
Gender	Male/Female
Married	Marital status : yes/no
Dependents	0 – 3+
Education	Education level
Self Employed	Yes/No
ApplicantIncome	Applicant Income
CoapplicantIncome	CoApplicant Income
LoanAmount	Requested Loan amount

Loan_Amount_Term	Loan repayment terms
Credit_History	Credit history: bad : 0 , good:1
Property_Area	Property location

1.1.2 *Data Cleaning*

The strategy used to account for missing features was to remove the entire data row from the dataset. The justification for this is it is reasonable to assume that we can enforce the requirement of none of the required feature values to be missing. While it is a valid criticism to mention Credit history could be a missing value in a real-world example the possibility does not exist in this dataset.

1.1.3 *Data Transformation*

The loan dataset initially was a unbalanced dataset with the majority of the data for approved loans. To transform the dataset to a balanced dataset Oversampling was applied to nonapproved loans.

1.2 Stroke Data

This dataset is a collection of a common health condition to predate the likelihood of a stroke occurring.

Why this data is interesting is because it has a feature which when combined with great accuracy can predict a stroke but also has anomaly conditions outside of those ranges that also point to a high probability of a stroke.

The challenge with this dataset is how do you design a solution that favors the typical features ranges, without eliminating the anomaly conditions. Another challenge is how do you obtain a efficient number anomaly samples in order to train a model?

1.3 Features

Name	Description
id	Record Id

Gender	Male, Female, Other
Age	Age
Hypertension	Yes/No
Heart_disease	Yes/No
Ever_married	Yes/No
Work_type,	Work Type
Residence_type	Residence Type
Avg_glucose_level	Average Glucose Level
BMI	Body Mass Index
Smoking_status,	Yes/No

1.3.1 *Data Cleaning*

The strategy used to account for missing features was to remove the entire data row from the dataset.

1.3.2 *Data Transformation*

The stroke dataset initially was a unbalanced dataset with the majority of the data for no stroke. To transform the dataset to a balanced dataset Oversampling was applied to the data indicating a possibility of a stroke.

2 EXPERIMENTS

2.1 Metrics

2.1.1 *Learning Curve*

To determine the learning curve for each algorithm I plotted the results using Scikit Learning Curve model. The learning curve was implemented using k-fold cross validation with a k value of 5 and using accuracy as the scoring metric.

2.1.2 *Validation Curve*

To determine the training and test score from turning the algorithm hyper parameters I used Scikit Validation Curve model. The validation curve was implemented using k-fold cross validation with a k value of 5 and using accuracy as the scoring metric.

2.2 Decision Tree

2.3 Figures

Figures should always be centered on the page, although they may also take up the entire width and height of the text block. Figures should always be referenced in the text, and they should include a descriptive caption. Figures may also be equations, diagrams, or other kinds of content.

If your figure includes a white background (e.g. an interface design or graph), it may aid legibility to add a ¼ point black border.

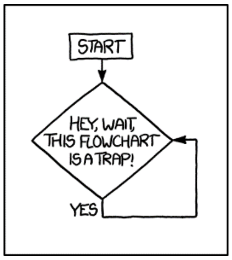

Figure 1—Make sure your flowcharts are more useful than this one. Source: XKCD.

Figure captions should be placed beneath the corresponding figure, indented 1″ on the left and right sides. The label for the figure, e.g. "Figure 1," should be set in bold italics followed by an em dash, and the entire caption should be 8.5 points with 14 points of line spacing. The *Figure Caption* paragraph style in Word will number your figures automatically. If need be, you may have one caption corresponding to multiple consecutive figures and use either locational descriptors (e.g. "top left," "middle") or labels (e.g. "A", "B") to map parts of the caption to parts of the figure. Make sure that caption falls on the same page as the corresponding figure or table; you may need to rearrange text to make this work.

In Microsoft Word, you may need to either change the image's text wrap settings to "Top and Bottom" or change the line spacing of the image to 1.0.

2.4 Tables

You have freedom to format tables in the way that works best for your data. Generally, text should be left-aligned and numbers should be right-aligned or aligned at the decimal – you can do this in Word using a decimal tab stop. The default table style (below) reduces the text size to be equal to the caption text.

Table captions should be formatted the same way as figure captions, but they should be placed above the table. The popular mnemonic for this is: figures at the foot, tables at the top. The *Table Caption* paragraph style will number your tables automatically. Like figures, tables should not exceed the margins and should be centered on the page.

2.5 Additional elements

There are additional elements you may want to include in your paper, such as in-line or block quotes, lists, and more. For other content types not covered here, you have reasonable flexibility determining how it should be used in this format.

Table 1 — Mathematical constants. Notice how the approximations align at the decimal.

Name	**Symbol**	**Approximation**	**Description**
Golden ratio	φ	1.618	Number such that the ratio of 1 to the number is equal to the ratio of its reciprocal to 1
Euler's number	e	2.71828	Exponential growth constant
Archimedes' constant	π	3.14	The ratio between circumference and diameter of a circle
One hundred	A⁺	100.00	The grade we hope you'll all earn in this class

2.5.1 *Quotes*

If you would like to quote an outside source, you may do so in quotation marks followed by a citation. If a quote is fewer than three lines, you may write it inline. It is acceptable to replace pronouns with their target in brackets for clarity. For example, "Heavy use of peer grading would compromise [the school's] reputation" (Joyner, 2016). If a quote exceeds three lines, you should set it as its own paragraph with 0.5″ side margins, using the *Blockquote* paragraph style.

> "Whether or not the grades generated by peers are reliably similar to grades generated by experts is only one factor worth considering, however. Student perception is also an important factor. […] Reliance on peer grading is one of the top drivers of high MOOC dropout rates. This problem may be addressed by reintroducing some expert grading where possible." (Joyner, 2016)

2.5.2 *Lists*

Bulleted and numbered lists are indented 0.25″ from the left margin, with the bullet or number hanging by 0.25″ (i.e., flush with the left margin).

- Like this
- And this
- And also this

3 PROCEDURAL ELEMENTS

3.1 In-line citations

Articles or sources to which you refer should be cited in-line with the authors' names and the year of publication.[1] The citation should be placed close in the text to the actual claim, not merely at the end of the paragraph. For example: students in the OMSCS program are older and more likely to be employed than students in the on-campus program (Joyner, 2017). In the event of multiple authors, list them. For example: research finds sentiment analysis of the text of OMSCS reviews corresponds to student-assigned ratings of the course (Newman & Joyner, 2018). You may also cite multiple studies together. For example: several studies have found students in the online version of an undergraduate CS1 class performed equally with students in a traditional version (Joyner, 2018a; Joyner, 2018b). If you would like to refer to an author in text, you may also do so by including the year (in parentheses) after the author's name in the text. If a publication has more than 4 authors, you may list the first author followed by 'et al.' For example: Joyner et al. (2016) claim that a round of peer review prior to grading may improve graders' efficiency and the quality of feedback given. This applies to parenthetical citations as well, e.g. (Joyner et al., 2016).

3.2 Reference lists

References should be placed at the end of the paper in a dedicated section. Reference lists should be numbered and organized alphabetically by first author's last name. If multiple papers have the same author(s) and year, you may append a letter to the end of the year to allow differentiated in-line text (e.g. Joyner, 2018a and Joyner, 2018b in the section above). If multiple papers have the same author(s), list them in chronological order starting with the older paper. Only works that are cited in-line should be included in the reference list. The reference list does not count against the length requirements.

[1] In-line citations are preferred over footnotes, and we favor APA citation format for both in-line citations and reference lists. Refer to the Purdue Online Writing Lab, or follow the above examples. You should use the *Footnote* paragraph style, with 8.5 point text and 14 point line spacing.

4 REFERENCES

1. Joyner, D. A., Ashby, W., Irish, L., Lam, Y., Langston, J., Lupiani, I., Lustig, M., Pettoruto, P., Sheahen, D., Smiley, A., Bruckman, A., & Goel, A. (2016). Graders as Meta-Reviewers: Simultaneously Scaling and Improving Expert Evaluation for Large Online Classrooms. In *Proceedings of the Third Annual ACM Conference on Learning at Scale*. Edinburgh, Scotland.
2. Joyner, D. A. (2017). Scaling Expert Feedback: Two Case Studies. In *Proceedings of the Fourth Annual ACM Conference on Learning at Scale*. Cambridge, Massachusetts.
3. Joyner, D. A. (2018a). Intelligent Evaluation and Feedback in Support of a Credit-Bearing MOOC. In *Proceedings of the 19th International Conference on Artificial Intelligence in Education*. London, United Kingdom. Springer.
4. Joyner, D. A. (2018b). Toward CS1 at Scale: Building and Testing a MOOC-for-Credit Candidate. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. London, United Kingdom. ACM Press.
5. Newman, H. & Joyner, D. A. (2018). Sentiment Analysis of Student Evaluations of Teaching. In *Proceedings of the 19th International Conference on Artificial Intelligence in Education*. London, United Kingdom. Springer.

5 APPENDICES

You may optionally move certain information to appendices at the end of your paper, after the reference list. If you have multiple appendices, you should create a section with a *Heading 1* of "Appendices." Each appendix should begin with a descriptive *Heading 2;* appendices can thus be referenced in the body text using their heading number and description, e.g. "Appendix 5.1: Survey responses." If you have only one appendix, you can label it with the word "Appendix" followed by a descriptive title, e.g., "Appendix: Survey responses."

These appendices do not count against the page limit, but they should not contain any information *required* to answer the question in full. The body text should be sufficient to answer the question, and the appendices should be included only for you to reference or to give additional context. If you decide to move content to an appendix, be sure to summarize the content and note it in relevant place in the body text, e.g., "The raw data can be viewed in *Appendix 5.1: Survey responses*."