Project #2

Robert Richter

Part 1: Training a Decision Tree Classifier

1. Prepare Data
2. Train Model
3. Merge Explanations

Preparing Data – Enums

- *DTCs require numbers only: How to transform Strings to Numbers?*
- Step 1: All fields with enums can be replaced by numbers

Enum Enum Enum Enum Enum

FailingMethod	Answer.option	Answer.explanation	Worker.profession	Worker.gender	Worker.whereLearnedToCode	Worker.country	Worker.programmingLanguage
HIT01_8	NO	hoursOffset accepts negative num	Undergraduate_Student	Male	High School	United States	Java; C++; C#
HIT01_8	NO	the conditional clause is correct fo	Undergraduate_Student	Female	High School;University;Web	United States	c#
HIT01_8	NO	The argument -2 is within the rang	Professional_Developer	Male	High School;University;Web	United States	C++;Java;PHP
HIT01_8	NO	The exception is in no way related	Professional_Developer	Male	High School;University;Other On the job	USA	C#
HIT01_8	NO	The issue cannot be with hoursOff	Undergraduate_Student	Male	High School;University;Web	United States	C++; Java
HIT01_8	NO	The failure speaks of Minutes out c	Graduate_Student	Male	University;Web	Colombia	C++
HIT01_8	NO	The usage and declaration of hour	Undergraduate_Student	Male	University;Web	USA	Java; PHP; HTML
HIT01_8	NO	lack of a proper source viewer edit	Hobbyist	Male	University;Web;Other self study	USA	Euphoria
HIT01_8	IDK	I dont know	Professional_Developer	Male	University	India	HTML
HIT01_8	IDK	it is tough	Graduate_Student	Male	University;Web	USA	java
HIT01_8	NO	The issue has nothing to do with th	Other	Male	High School;Other Training classes	United States	Do not currently use
HIT01_8	NO	its an integer value from -59 to 59.	Other	Female	University	US	Python
HIT01_8	NO	Hour offset is a valid input but it is	Professional_Developer	Male	Web	India	C#
HIT01_8	YES	The time should not be negative va	Graduate_Student	Male	University;Web	India	Java

Preparing Data – Partial Enums and Explanation Complexity

- *Step 2, Partial Enums (i.e., unstructured enums): Scoring (e.g., 1 point per programming language)*
- *Step 3, Explanation Complexity: various metrics to test (here: word count)*

FailingMethod	Answer.option	Answer.explanation	Worker.profession	Worker.gender	Worker.whereLearnedToCode	Worker.country	Worker.programmingLanguage
0	0	hoursOffset accepts negative num	0	0	High School	0	Java; C++; C#
0	0	the conditional clause is correct fo	1	1	High School;University;Web	0	c#
0	0	The argument -2 is within the rang	0	0	High School;University;Web	0	C++;Java;PHP
0	0	The exception is in no way related	0	0	High School;University;Other On the job	1	C#
0	0	The issue cannot be with hoursOff	0	0	High School;University;Web	0	C++; Java
0	0	The failure speaks of Minutes out c	0	0	University;Web	2	C++
0	0	The usage and declaration of hour	0	0	University;Web	1	Java; PHP; HTML
0	0	lack of a proper source viewer edit	0	0	University;Web;Other self study	1	Euphoria
0	1	I dont know	0	0	University	3	HTML
0	1	it is tough	0	0	University;Web	1	java
0	0	The issue has nothing to do with th	0	0	High School;Other Training classes	0	Do not currently use
0	0	its an integer value from -59 to 59.	1	1	University	4	Python
0	0	Hour offset is a valid input but it is	0	0	Web	3	C#
0	2	The time should not be negative va	0	0	University;Web	3	Java

Preparing Data – Partial Enums and Explanation Complexity

Enum	Enum	Enum	Enum	Score	Enum	Score	Word Count
FailingMethod	Answer.option	Worker.profession	Worker.gender	Worker.whereLearnedToCode	Worker.country	Worker.programmingLanguage	Answer.explanationComplexity
0	0	0	0	100	0	4	16
0	0	1	1	1150	0	2	13
0	0	0	0	1150	0	4	29
0	0	0	0	1100	1	2	11
0	0	0	0	1150	0	3	17
0	0	0	0	1050	2	2	28
0	0	0	0	1050	1	3	19
0	0	0	0	1050	1	1	60
0	1	0	0	1000	3	1	3
0	1	0	0	1050	1	1	3
0	0	0	0	100	0	2	30
0	0	1	1	1000	4	1	12
0	0	0	0	50	3	2	14
0	2	0	0	1050	3	1	8

Performance of this?

Parameters:

- ccp_alpha: 0.04
- criterion: entropy
- Test size: 20 Samples
- Explanation Complexity: Word Count

	Importance
FailingMethod	0.33
Code.LOC	0.31
Answer.order	0.11
Code.complexity	0.11
Answer.option	0.09
Answer.difficulty	0.02
Worker.yearsOfExperience	0.02
Worker.gender	0
Worker.programmingLanguage	0
Worker.country	0
Worker.whereLearnedToCode	0
Worker.score	0
Worker.age	0
Worker.profession	0
Answer.duration	0
Answer.confidence	0
Answer.explanationComplexity	0

	precision	recall	f1-score	support
0	0.78	1.00	0.88	14
1	1.00	0.33	0.50	6
accuracy			0.80	20
macro avg	0.89	0.67	0.69	20
weighted avg	0.84	0.80	0.76	20

Untuned Model even better?

Parameters:

- ccp_alpha: 0
- criterion: gini
- Test size: 20 Samples
- Explanation Complexity: Word Count

- This often holds recall & precision of 1 …

	Importance
Code.LOC	0.17
Answer.duration	0.12
FailingMethod	0.12
Answer.explanationComplexity	0.11
Code.complexity	0.09
Worker.age	0.07
Worker.yearsOfExperience	0.05
Answer.order	0.05
Answer.option	0.04
Answer.confidence	0.04
Worker.whereLearnedToCode	0.03
Worker.country	0.03
Worker.programmingLanguage	0.03
Answer.difficulty	0.02
Worker.score	0.02
Worker.profession	0.01
Worker.gender	0.01

	precision	recall	f1-score	support
0	0.93	1.00	0.96	13
1	1.00	0.86	0.92	7
accuracy			0.95	20
macro avg	0.96	0.93	0.94	20
weighted avg	0.95	0.95	0.95	20

What about the Halstead Metric?

- Parameters:
  - ccp_alpha: 0.04
  - criterion: gini
  - Test size: 20 Samples
  - Explanation Complexity: Word Count

- Halstead makes it worse
- Should we *not* look at the code?

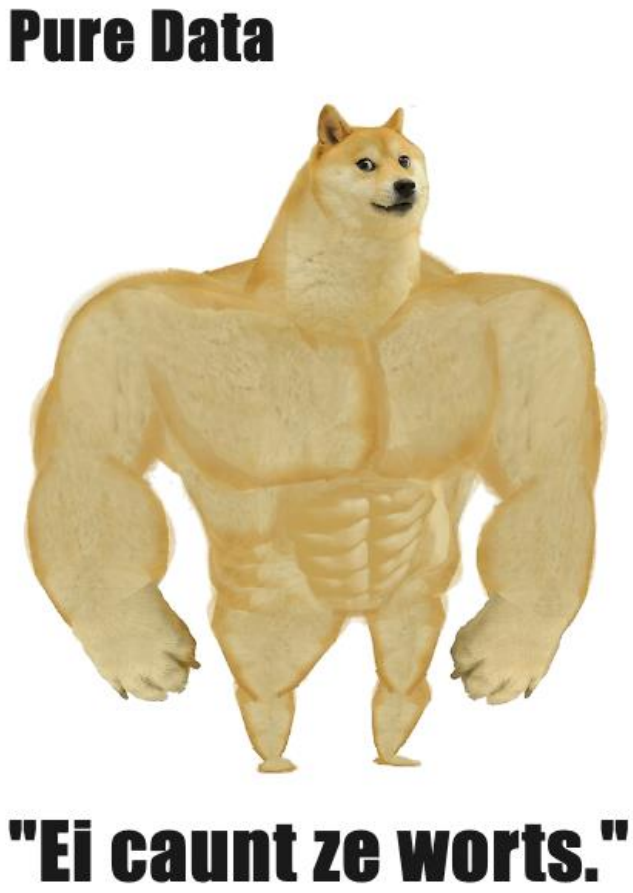
	Importance
Code.LOC	0.19
Answer.duration	0.12
Answer.explanationComplexity	0.09
Answer.order	0.08
Worker.age	0.07
Code.complexity	0.07
FailingMethod	0.06
Worker.yearsOfExperience	0.06
Answer.option	0.04
Answer.difficulty	0.04
Worker.whereLearnedToCode	0.04
Worker.score	0.04
Worker.country	0.04
Worker.programmingLanguage	0.03
Answer.confidence	0.02
Worker.profession	0.01
Worker.gender	0

	precision	recall	f1-score	support
0	0.73	0.85	0.79	13
1	0.60	0.43	0.50	7
accuracy			0.70	20
macro avg	0.67	0.64	0.64	20
weighted avg	0.69	0.70	0.69	20

Part 1: Conclusion

Running the model purely on the prepared data without tuning the DTC gave the best results.

- *What I did not mention: Adding QuestionIDs and AnswerIDs yielded perfect results (ask for it!)*



Part 2: Merge Explanations

- Model: GPT4o (via OpenAI API)
- Approach #1: Only use Test Set (answers predicted to be corrected)
  - Answers like “I cannot give a conclusive answer without looking at the code” (but that would be too easy, wouldn’t it?)
  - Also, we don’t know whether we predicted correctly
  - Improvements: Adding more context, target groups, handing over code files

Part 2: Merge Explanations

- Approach #2: Use the whole dataset with the ground truth
  - GPT-4o: 0.05 $, 56.51s
  - GPT-4o-mini: 0.003 $, 88.43s

  - But results were really nice!

The issue arises from a check on line 279 of the code which incorrectly restricts the `minutesOffset` parameter to values between 0 and 59. This restriction is inconsistent with the comments and documentation that indicate `minutesOffset` can be negative, up to -59, starting from version 2.3. The problem occurs because the check (`if (minutesOffset < 0 || minutesOffset > 59)`) throws an `IllegalArgumentException` for negative values such as -15, as set by `DateTimeZone.forOffsetHoursMinutes(-2, -15)`. To fix the issue, the check on line 279 should be updated to allow negative values by changing the condition to `if (minutesOffset < -59 || minutesOffset > 59)`. This aligns with the documented behavior and allows the code to progress correctly without throwing an exception for valid negative minute offsets.',

Appendix

Appendix 1: Adding QuestionID and AnswerID

	Importance
Answer.ID	0.45
Question.ID	0.31
Code.LOC	0.11
Answer.order	0.04
Answer.option	0.04
Code.complexity	0.02
Worker.yearsOfExperience	0.01
Answer.duration	0.01
Answer.explanationComplexity	0
FailingMethod	0
Worker.whereLearnedToCode	0
Worker.profession	0
Worker.country	0
Worker.age	0
Answer.difficulty	0
Answer.confidence	0
Worker.score	0
Worker.gender	0
Worker.programmingLanguage	0

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13
1	1.00	1.00	1.00	7
accuracy			1.00	20
macro avg	1.00	1.00	1.00	20
weighted avg	1.00	1.00	1.00	20

Number of Words
(untuned model)

	Importance
Answer.ID	0.53
Question.ID	0.21
Code.LOC	0.11
FailingMethod	0.05
Code.complexity	0.04
Answer.order	0.04
Answer.option	0.03
Answer.difficulty	0
Answer.confidence	0
Answer.duration	0
Worker.score	0
Worker.profession	0
Worker.yearsOfExperience	0
Worker.age	0
Worker.gender	0
Worker.whereLearnedToCode	0
Worker.country	0
Worker.programmingLanguage	0
Answer.explanationComplexity	0

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13
1	1.00	1.00	1.00	7
accuracy			1.00	20
macro avg	1.00	1.00	1.00	20
weighted avg	1.00	1.00	1.00	20

Number of Words
(tuned model)

	Importance
Answer.ID	0.39
Question.ID	0.26
Code.LOC	0.12
Answer.explanationComplexity	0.11
Code.complexity	0.04
Answer.option	0.04
Answer.order	0.03
Answer.difficulty	0
Worker.whereLearnedToCode	0
Answer.duration	0
Worker.score	0
Answer.confidence	0
FailingMethod	0
Worker.profession	0
Worker.yearsOfExperience	0
Worker.age	0
Worker.gender	0
Worker.country	0
Worker.programmingLanguage	0

	precision	recall	f1-score	support
0	1.00	1.00	1.00	13
1	1.00	1.00	1.00	7
accuracy			1.00	20
macro avg	1.00	1.00	1.00	20
weighted avg	1.00	1.00	1.00	20

Halstead Metric
(tuned model)