ENHANCING FAIRNESS IN IN-CONTEXT LEARNING: PRIORITIZING MINORITY SAMPLES IN DEMONSTRATIONS

Jingyu Hu
University of Bristol
ym21669@bristol.ac.uk

Mengnan Du
New Jersey Institute of Technology
mengnan.du@njit.edu

ABSTRACT

Recent studies highlight the effectiveness of using in-context learning to steer large language models (LLMs) in processing tabular data, a challenging task given the structured nature of such data. Despite advancements, the fairness implications of this approach remain underexplored. This study delves into how varying demonstrations impact LLM fairness, particularly by examining the distribution of selected samples in prompts. We find that deliberately including minority samples in prompts can significantly enhance fairness awareness in LLMs, without compromising their predictive performance.

1 INTRODUCTION

Large Language Models (LLMs), known for their versatility in various natural language processing tasks, have garnered significant attention for their potential for further enhancement. One common optimization technique is in-context leanring, where input prompts are designed to guide LLMs in generating more accurate outputs without modifying the pre-trained model (Liu et al. (2023a)). Specifically, few-shot learning add few examples for in-context learning, and demonstrates an augmentation to model performance (Brown et al. (2020), Schick & Schütze (2020)). Recent studies have shown effective performance using in-context learning to process tabular data, as shown in studies utilizing datasets from UCI machine learning repository Hegselmann et al. (2023). Liu et al. (2023b) examines fairness on GPT-3.5 in comparison to traditional ML models. However, there is limited work on assessing how LLM fairness varies with different prompts in processing tabular data. Our study aims to fill this gap by investigating the impact of different demonstration strategies. To better understand the impact of few-shot learning on fairness, our proposed demonstration strategy considers the distribution of both demographic groups and target labels. Our investigation includes evaluating three advanced LLMs OpenAI (2023), i.e., Text-davinci-003, GPT-3.5-turbo, and GPT-4-turbo, across two fairness-focused tabular datasets: Credit and Adult. Our experimental findings suggest that giving priority to underrepresented samples and conscientiously including minority demographic groups and target labels during few-shot learning can significantly improve the fairness performance in LLM output.

2 EXPERIMENT DESIGN

The experiment tests and evaluates potential fairness implications of demonstration strategies with different priority given to LLMs.

Models and Datasets We use three LLMs: Text-davinci-003(Davinci), GPT-3.5-turbo, and GPT-4-turbo. Temperature is set to zero to ensure consistent responses. We consider it in a binary classification prediction scenario and select two tabular-based fairness datasets: Default of credit card clients Dataset (Credit, Yeh (2016)) and adult income (Adult, Becker & Kohavi (1996)). Credit dataset covers information on credit card clients in Taiwan, including demographics, bills, payment history, etc. Its target is to predict whether there will be an overdue payment next month. Adult dataset is to predict whether an individual's annual income exceeds 50K based on features like education level, demographics, occupation, etc.

Demonstrations Design We set zero-shot as baseline and proposed three few-shot prompt strategies. Our shots number are set to eight and extracted from training dataset with the following three shots selection strategies: (1) S1: Balanced Samples with Balanced Labels; (2) S2: Prioritize Minority Samples with Balanced Labels; (3) S3: Prioritize Minority Samples with Unbalanced Labels. We categorize subgroups by gender, with the minority group labeled as female. The detailed templates are explained in Appendix.

Evaluation Metrics Two confusion matrix-based metrics—accuracy and F1-score—are used for evaluating prediction performance. The differences and ratios between the majority and minority groups in Demographic Parity (DP) and Equalized Odds (Eodds) are used for fairness measurement.

3 Results

The results in Credit dataset (Figure 6.4.2) show that all few-shot strategies have generally improved fairness compared to zero-shot learning without lowering predictions. Fairness performance is better when prioritizing samples from minority groups (S2 and S3) compared with balanced samples selection (S1). The abnormal results related in Davin3 baseline are further discussed in the appendix 6.4.1.

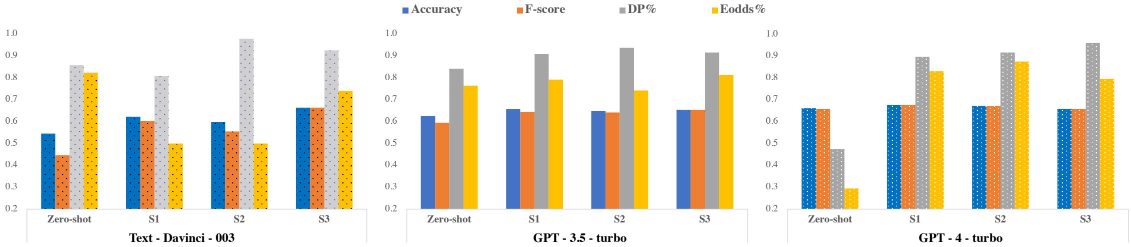


Figure 1: Prediction and Fairness Performance Comparison across LLMs with Credit Dataset

Similar findings have been found in the Adult dataset (Table 3) that prioritizing minorities with undesirable labels (i.e., females with low income) (S2) leads the most effective way to enhance fairness. In contrast, balanced prompts (S1) have much worse fairness performance. We discussed results in the appendix 6.4.2 with different K demonstrations (4 and 16) and found similar patterns.

Table 1: Performance on Adult dataset (GPT-3.5-turbo, F:Female; M:Male; H: $> 50K$; L: $\leq 50K$)

	Zero-shot (Baseline)	4F4M_4L4H (S1)	8F_4L4H (S2)	8F_8L (S3)	8F_8H (S3)
F-score	0.6669	0.7332	**0.7479**	0.7301	0.6911
DP%	0.4063	0.6475	0.7254	**0.7273**	0.5806
Eodds%	0.1111	0.3333	0.4390	**0.5517**	0.1053

4 Conclusion

This paper discusses the impact of fairness on LLMs across few-shot demonstrations with different distributions. Our experiment suggests that prioritizing underrepresented samples and considering minority demographic groups and target labels during few-shot generation could improve the fairness of LLMs. We consider extending it to multi-class scenarios and apply other sampling strategies in our future work.

5 URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR, 2023.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023a.

Yanchen Liu, Srishti Gautam, Jiaqi Ma, and Himabindu Lakkaraju. Investigating the fairness of large language models for predictions on tabular data. *arXiv preprint arXiv:2310.14607*, 2023b.

OpenAI. gpt-4-turbo, gpt-3.5-turbo, text-davinci-003, 2023. URL https://platform.openai.com/docs/models/model-endpoint-compatibility.

Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020.

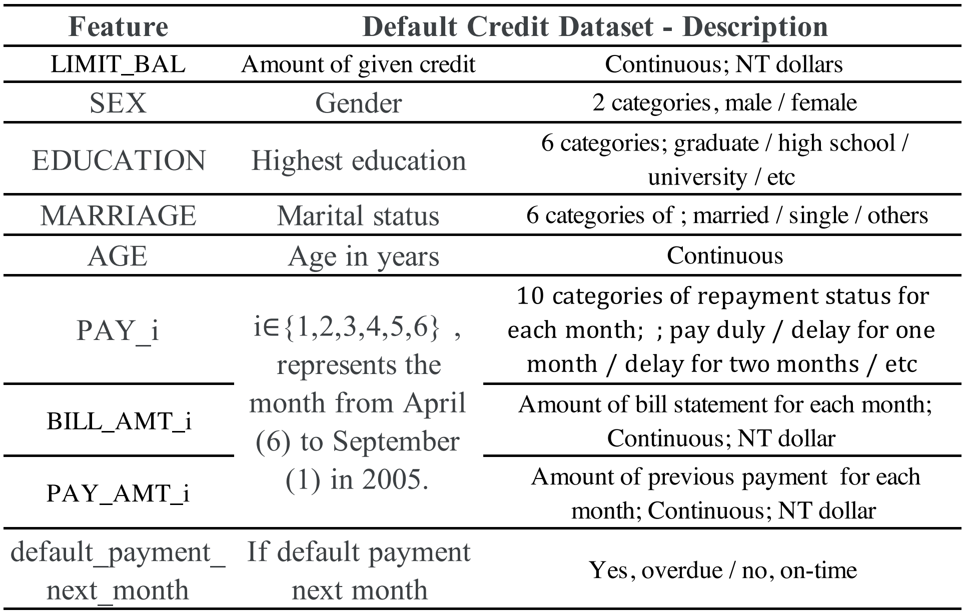
I-Cheng Yeh. default of credit card clients. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C55S3H.

6 Appendix

6.1 Dataset

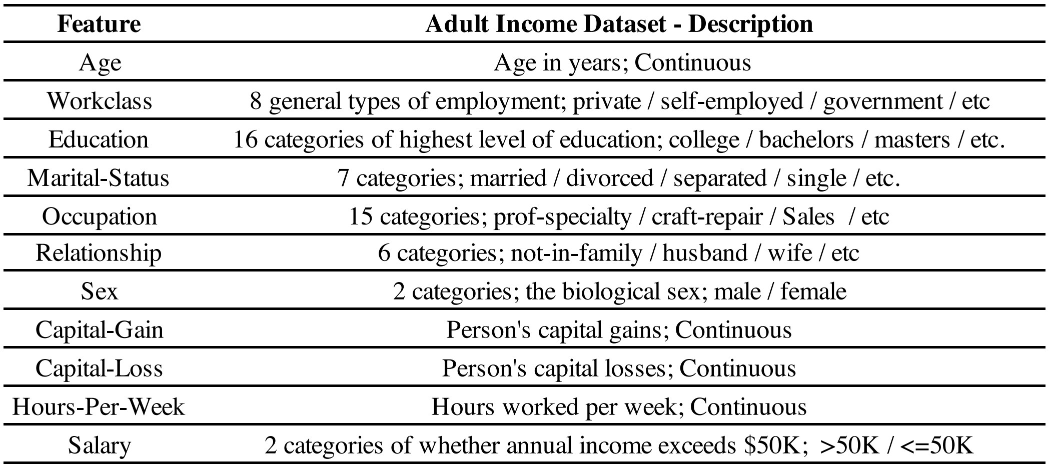
In Default Credit dataset, we calculate the mean values of PAY_AMT_i and BILL_AMT_i, and merge them into Avg_PAY_AMT and Avg_BILL_AMT separately. The raw Adult dataset contains 14 features, excluding education-num, fnlwgt, race, and native-country for this experiment. '$> 50K$' and '$\leq 50K$' are mapped to 'greater than 50K' and 'less than or equal to 50K' respectively, for better alignment with the language model. We consider the balanced group scenario and sample 512 instances with random_seed=42 in two datasets.

Table 2: Default Credit Dataset Description



Feature	Default Credit Dataset - Description	
LIMIT_BAL	Amount of given credit	Continuous; NT dollars
SEX	Gender	2 categories, male / female
EDUCATION	Highest education	6 categories; graduate / high school / university / etc
MARRIAGE	Marital status	6 categories of ; married / single / others
AGE	Age in years	Continuous
PAY_i	i∈{1,2,3,4,5,6} , represents the month from April (6) to September (1) in 2005.	10 categories of repayment status for each month; ; pay duly / delay for one month / delay for two months / etc
BILL_AMT_i		Amount of bill statement for each month; Continuous; NT dollar
PAY_AMT_i		Amount of previous payment for each month; Continuous; NT dollar
default_payment_next_month	If default payment next month	Yes, overdue / no, on-time

Table 3: Adult Income Dataset Description



Feature	Adult Income Dataset - Description
Age	Age in years; Continuous
Workclass	8 general types of employment; private / self-employed / government / etc
Education	16 categories of highest level of education; college / bachelors / masters / etc.
Marital-Status	7 categories; married / divorced / separated / single / etc.
Occupation	15 categories; prof-specialty / craft-repair / Sales / etc
Relationship	6 categories; not-in-family / husband / wife / etc
Sex	2 categories; the biological sex; male / female
Capital-Gain	Person's capital gains; Continuous
Capital-Loss	Person's capital losses; Continuous
Hours-Per-Week	Hours worked per week; Continuous
Salary	2 categories of whether annual income exceeds $50K; >50K / <=50K

6.2 Prompt Architecture

We consider both zero-shot learning and few-shot learning. Zero-shot strategy combines background and question as its prompt content without providing examples. Few-shot strategy includes three roles, and the in-context content is generated based on selected samples using different strategies (Table 5). Table 4 provides templates for our few-shot learning.

Table 4: Few-shot Learning Templates

Roles	Prompting Templates for Default Credit Dataset
Task Description	Predict if the following data will default payment next month. Answer with one of the following only: No | Yes
In-context Examples	Example 1: Amount of given credit is 490000, and SEX is male, and EDUCATION is graduate school, and MARRIAGE is married, and AGE is 45, and PAY_0 is pay duly,......, and default payment next month is No, on-time; Example 2: ......
Question	Amount of given credit is 90000, and SEX is female, and EDUCATION is university, and MARRIAGE is married, and AGE is 49, and PAY_0 is delay for one month,......, and predict whether default payment:

Roles	Prompting Templates for Adult Income Dataset
Task Description	Predict if income exceeds $50K per year. Answer with one of the following: greater than 50K | less than or equal to 50K
In-context Examples	Example 1:age is 40, and workclass is Private, and education is HS-grad, and marital-status is Married-civ-spouse, and occupation is Sales, and relationship is Husband, and sex is Male, and capital-gain is 0, and capital-loss is 0, and hours-per-week is 60, and income is <=50K; Example 2 ......
Question	age is 19, and workclass is Private, and education is Some-college, and marital-status is Never-married, and occupation is Other-service, and relationship is Own-child, and sex is Female, and capital-gain is 0, and capital-loss is 0, and hours-per-week is 15, please answer the income:

For the annotation, F/M represents gender group, and L/H represents binary target labels group. The preceding number indicates the count of samples in that group. For instance, 8F_8L means all 8 shots are low-income females.

Table 5: Few-shot Sample Selection Strategies (Adult dataset: F: Female, M: Male, H: High income ($> 50K$), L: Low income ($\leq 50K$). Default Credit dataset: F: Female, M: Male, H: Overdue/Default credit, L: On-time credit.)

Annotation	Strategy	Adult Income Dataset	Default Credit Dataset
4F4M_4L4H	S1	A balanced ratio of female-low-income : female-high-income : male-low-income : male-high-income = 2:2:2:2	A balanced ratio of female-overdue : female-on-time : male-overdue : male-on-time = 2:2:2:2
8F_4L4H	S2	prioritize minority with balanced targets of female-low-income : female-high-income = 4:4	prioritize minority with balanced targets. female-overdue: female-on-time = 4:4
8F_8L	S3	prioritize minority with imbalanced targets of 8 female-low-income samples	prioritize minority with imbalanced targets of 8 female-on-time samples

6.3 Evaluation Metrics Definition

Four confusion matrix (Table 6.3)-based metrics accuracy (Equation 1), precision(Equation 2), recall (Equation 3) and F-score (Equation 4) are used for prediction performance measurement.

Table 6: Confusion Matrix

	Predict Positive	Predict Negative
Real Positive	TP	FN
Real Negative	FP	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

Demographic Parity (DP) and Equalized Odds (Eodds) are two fairness criteria used in this experiment. The dataset is grouped by sensitive attribute gender. DP requires the prediction is similarly across different values of the sensitive attribute.

$$P(\hat{Y} \mid A = female/male) = P(\hat{Y}) \tag{5}$$

Eodds focuses on equalizing the false positive rate (FPR) and true positive rate (TPR) across different values of the sensitive attribute A.

$$P(\hat{Y} = 1 \mid Y = 1, A = female) = P(\hat{Y} = 1 \mid Y = 1, A = male) \tag{6}$$

$$P(\hat{Y} = 0 \mid Y = 0, A = female) = P(\hat{Y} = 0 \mid Y = 0, A = male) \tag{7}$$

The ratios (%) and Δ of DP and Eodds represent the corresponding proportions and differences between the sensitive groups (female and male).

6.4 RESULTS

6.4.1 THE PERFORMANCE ON DEFAULT CREDIT DATASET

As discussed above, focusing on minority samples (S2, S3) enhanced fairness more effectively in both GPT-4 and GPT-3.5 compared to S1 strategy. Notably, the fairness performance after few-shot is even lower than the baseline in Text-Davinci-3. This decline is attributed to the poor performance Davinci-3 in tabular classification task under zero-shot setting. The baseline accuracy is 54.49%, which nearly equivalent to random guessing. This indicates the Davinci-3 under zero-shot is under-fitting, so the fairness metric values at the baseline have lost their reference significance. We found all few-shot strategies boosted performance of Davinci-3, demonstrating their effectiveness in improving model fitting. Therefore, our fairness comparison in Davin-3 is across three few-shot strategy. The result shows prioritising the minority group (S2, S3) surpasses the balanced strategy (S1), aligning with our findings.

Table 7: LLMs performance on Default Credit dataset (F:Female; M:Male; H: overdue/default; L: on-time)

Default Credit Dataset	**GPT 3.5 - turbo**				**GPT 4 - turbo**				**Text - Davinci - 3**			
	Baseline	S1	S2	S3	Baseline	S1	S2	S3	Baseline	S1	S2	S3
	Zero-shot	**4F4M_4L4H**	**8F_4L4H**	**8F_8L**	**Zero-shot**	**4F4M_4L4H**	**8F_4L4H**	**8F_8L**	**Zero-shot**	**4F4M_4L4H**	**8F_4L4H**	**8F_8L**
Accuracy	0.6250	0.6562	0.6484	0.6543	0.6602	0.6758	0.6719	0.6582	0.5449	0.6230	0.5996	0.6641
F-score	0.5947	0.6453	0.6413	0.6543	0.6579	0.6758	0.6716	0.6578	0.4453	0.6030	0.5545	0.6641
	Zero-shot	**4F4M_4L4H**	**8F_4L4H**	**8F_8L**	**Zero-shot**	**4F4M_4L4H**	**8F_4L4H**	**8F_8L**	**Zero-shot**	**4F4M_4L4H**	**8F_4L4H**	**8F_8L**
Δ DP	0.0391	0.0313	**0.0234**	0.0430	0.2969	0.0547	0.0469	0.0195	0.0117	0.0586	0.0039	0.0391
DP%	0.8413	0.9080	**0.9368**	0.9160	0.4759	0.8955	0.9155	**0.9590**	0.8571	0.8077	**0.9787**	0.9254
Δ Eodds	**0.0938**	0.1016	0.1094	0.1328	0.3125	0.1250	**0.0938**	0.1094	**0.0234**	0.1016	0.0547	0.1016
Eodds%	0.7647	0.7917	0.7419	**0.8132**	0.2941	0.8298	**0.8750**	0.7955	**0.8235**	0.5000	0.5000	0.7400

6.4.2 THE PERFORMANCE ON ADULT DATASET

In adult income dataset, we also explored the consequences of prioritizing majority shots strategy(S4) in demonstration (Table 6.4.2). Results show a negative impact on the model's prediction performance when all samples are high-income males, with accuracy dropping from 68.55% to 63.28%. We also tested the scenario where all shots are high-income females and observed an improvement in performance. Overall, in-context shots with different distributions of groups and labels can impact LLM performance.

Table 8: GPT3.5-turbo performance on Adult Income dataset (F:Female; M:Male; H:$> 50K$; L: $\leq 50K$)

Adult Income	Baseline	S1	S2	S3	S3	S4
	Zero-shot	4F4M_4L4H	8F_4L4H	8F_8L	8F_8H	8M_8H
Acc	0.6855	0.7363	0.7480	0.7324	0.7051	0.6328
F-score	0.6669	0.7332	0.7479	0.7301	0.6911	0.5873
	Zero-shot	4F4M_4L4H	8F_4L4H	8F_8L	8F_8H	8M_8H
Δ DP	0.2227	0.1680	0.1523	0.1289	0.1523	0.1250
DP%	0.4063	0.6475	0.7254	0.7273	0.5806	0.4576
Δ Eodds	0.3203	0.1797	0.1797	0.1563	0.1719	0.1953
Eodds%	0.1111	0.3333	0.4390	0.5517	0.1053	0.1250

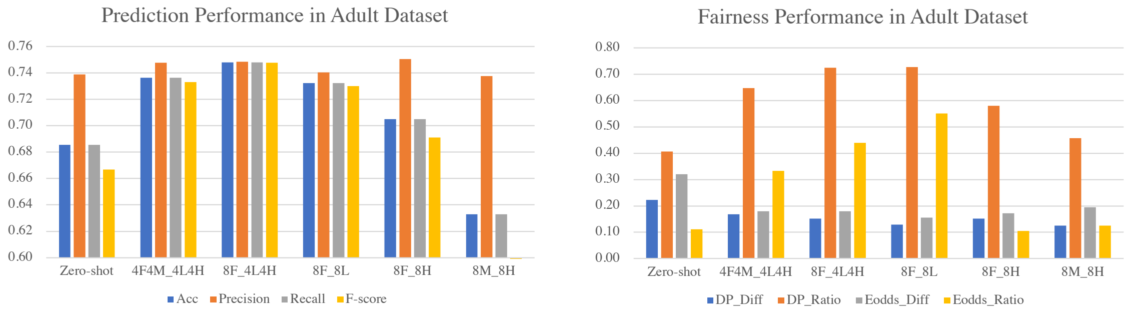


Figure 2: GPT3.5-turbo prediction and fairness performance on Adult Income dataset (F:Female; M:Male; H:$> 50K$; L: $\leq 50K$)

Table 6.4.2 shows the impact of changing the number of K-demonstrations on fairness performance. The results align with the 8-shot scenario: giving preference to minority samples (S2, S3) significantly enhances fairness. Furthermore, the influence of K quantity on model performance relies on the selected strategy. Increasing K has a limited effect in the balanced strategy (S1), and adding demonstrations even worsens performance when favoring the majority group (S4). However, increasing K-shot effectively affects both model prediction and fairness when prioritizing the minority group (S2, S3).

Table 9: GPT3.5-turbo performance on Adult Income dataset with different K (F:Female; M:Male; H:$> 50K$; L: $\leq 50K$)

	Zero-shot	**Four - shot**					**Sixteen - shot**				
	Baseline	**S1**	**S2**	**S3**	**S3**	**S4**	**S1**	**S2**	**S3**	**S3**	**S4**
	Zero-shot	2F2M_2L2H	4F_2L2H	4F_4L	4F_4H	4M_4H	8F8M_8L8H	16F_8L8H	16F_16L	16F_16H	16M_16H
Acc	0.6855	0.7383	0.7305	0.7012	0.7363	0.6309	0.7305	0.7656	0.7324	0.7090	0.6523
F-score	0.6669	0.7351	0.7256	0.6943	0.7312	0.5845	0.7293	0.7656	0.7309	0.6952	0.6168
	Zero-shot	2F2M_2L2H	4F_2L2H	4F_4L	4F_4H	4M_4H	8F8M_8L8H	16F_8L8H	16F_16L	16F_16H	16M_16H
Δ DP	0.2227	0.1719	0.1406	0.1523	0.1680	0.1680	0.1641	0.1172	0.0898	0.1367	0.1641
DP%	0.4063	0.6393	**0.6786**	<u>0.6422</u>	0.6228	0.3281	0.6818	<u>0.7857</u>	**0.8083**	0.6154	0.4085
Δ Eodds	0.3203	0.2109	0.1641	0.2266	0.1953	0.2813	0.1719	0.1641	0.1016	0.1641	0.2422
Eodds%	0.1111	0.3929	<u>0.4000</u>	**0.5833**	0.2800	0.1250	0.4444	<u>0.4615</u>	**0.5806**	0.1765	0.0000