Assignment: Ethics, Fairness and Explanation in AI

1. Exploratory Data Analysis

1.1.

(i) There are a much higher proportion of the females present who survived relative to the men. Survival was also most highly correlated with the fare passengers paid, meaning the more expensive the ticket, the more likely they survived. The travelling class was also the feature most negatively correlated with survival. This initially seems a bit counter intuitive because of the reverse ordinality of the class system (i.e. first class, though a lower number = higher wealth/privilege), however as the value of this class feature increases (to second / third class), the survival feature decreases. Additionally, there is a strong correlation between fare and class, supporting the observation that socioeconomic status, reflected through the travel class and fare paid, significantly influenced survival outcomes.

(ii) The most impactful features are identified by considering factors such as feature distribution, correlation with the target variable, and domain relevance. As such it is likely that the features highlighted above - pclass, fare and sex - would all be most important. Evidently then the features less correlated with survival - age, sibsp, and parch - would be less relevant. I can be certain because features that show a strong correlation with the target variable are likely to be important.

Understanding the historical context and domain specifics can also guide the importance of features. 'Sex' is a historically significant factor because women and children were given priority in lifeboat allocation. 'fare' and 'pclass' also show substantial correlations (positive or negative) with survival, and they reflect socioeconomic status, which could influence access to lifeboats.

(iii) One approach could be to use feature importance analysis, which can be performed using models like Random Forest or Gradient Boosting. By training such a model and examining the feature importances it assigns, one can gain insights into which features are most predictive of survival, providing a quantitative measure to support or refine the hypotheses formed from the initial visual exploratory analysis. Investigating how combinations of features, such as class and fare or age and sex, impact survival chances might reveal complex interactions that simple correlation analysis might miss.

2. Feature Attribution Explanations

2.1. Implemented the SHAP method in the provided Jupyter notebook (i-iv)

2.2. SHAP, Shapley Value Sampling and DeepLIFT comparison of feature attribution scores

(i) For the SHAP attributions, it seems that the last column feature (-9.0514, -7.9318, etc., across different instances) consistently has the largest and mainly negative impact on the model's output, suggesting it might be one of the most important features in determining the outcome. In the SHAP approach the first feature also shows significant weight and variability, suggesting it could be important in certain cases.

For the Shapley Value Sampling (SVS) and DeepLIFT (Shrikumar et al. [1]) attributions more broadly, both methods highlight different features as important, and the SVS approach honestly doesn't have that strong a cohesion with the findings of the other 2 approaches. SVS attributes significant importance to the fifth, sixth and seventh features (positive, negative, and negative respectively), while DeepLIFT's attributions are far far more subdued, with slight negative importance attributed to the final and fifth features, and a small contribution from the first and second features.

(ii) The main difference you see in the attribution methods is with the scal of their outputs. For example the SHAP attributions show a wide range of values with relatively large magnitudes across all the features, indicating a strong influence of certain features on the model's prediction. The SVS attributions also provide some insights but through a narrower view, with only a few atributes really showing much influence. DeepLIFT attributions are much much smaller in magnitude, suggesting subtler interpretations of feature impacts. These differences may also stem from the methodologies each approach uses. SVS uses a sampling-based approximation for SHAP values, leading to potential variance and a focus on more influential coalitions. DeepLIFT compares the difference between activations with a reference input, focusing on the gradient of changes in this way might not capture the full combinatorial impact in the same way as the other attribution approaches.

(iii) From task 1(a)(ii), I did suggest that pclass, fare and sex (particularly being female) would all likely be the most important/impactful features

to survival. Assuming that female sex is the final column, this indeed does align very strongly with out previous user expectations that this would be one of the most influential features on survival likelihood. It was highlighted as one of the strongst contributing features across all 3 methods. As for pclass fare, the first and fifth columns respectively, SHAP attributed a lot of importance to this first feature and DeepLift attributed a smaller amount of significance to this first feature; as for the fifth feature SVS recognised this as one of it's relevant features, and Deeplift again attributed a smaller but still present significance to this feature.

Discrepancies between expected and computed explanations can arise due to complex interactions between features being more intuitively captured by attribution methods relative to a human-centred/ intuitive expectations.

(iv) For SHAP, in considering all possible feature coalitions, it provides potentially the most comprehensive insights, with a detailed understanding of feature interactions and importance. Similar to a grid search, this makes the process very computationally intensive (particularly for large datasets or models with many features, e.g. our bean dataset). This potentially limits its applicability depending on the distribution / qualities of the data you're working with.

Shapley Value Sampling (SVS) offers a more computationally tractable approximation of SHAP values by randomly sampling coalitions, this arguably makes it more suitable for larger datasets. However, this approximation component introduces variance which could lead to inconsistencies in explainations. Additionally, the explanations here might not capture the full detail of feature contributions in such a way that exact SHAP calculations do, potentially making it comparatively less robust.

Finally, looking at DeepLIFT, potential advantages with this method are that it is efficient for gradient-based models, and can therefore provide quick insights into how input changes affect outputs, particularly usefull for use-cases like deep neural networks. A disadvantage though, is that once again it may not fully capture the combinatorial impact of features as effectively as SHAP, or also specific complex feature interactions, potentially resulting in less detailed explanations.

2.3. Quantitative evaluation of the different attribution methods through Infidelity (table on next page)

Since infidelity measures the discrepancy between the predicted change in a model's output due to feature perturbations relative to the actual change observed, the scores I've calculated here indicate a close alignment between the attribution explanation and the model's behavior, as they're generally very low.

It's first easier to evaluate how the infidelity scores vary with changes in the noise scale. This is to be expected, as increasing levels of perturbation can affect how well the attributions predict the impact on model output. SHAP's infidelity seems to increase the most significantly as the noise scale increases from 0.1 to 0.9, indicating that its explanations become less reliable under larger perturbations. The decrease in reliability that can be seen for the other two approaches is more gradual, but still echoes the same trend as noise/pertubation increases.

Looking at varying the categorical resampling probability (cat resample probability), this doesn't seem to have affected the infidelity scores for ay of the methods at all, as evidenced by consistent scores across different probabilities when the noise scale is fixed at 0.2... I don't know if this is a mistake, but otherwise it would suggests that the explanations provided by these methods are relatively robust to changes in how categorical features are perturbed.

Overall the infidelity scores indicate that SHAP method is pretty sensitive to the scale of input perturbations, with reliability decreasing the most as perturbations increase. In contrast, DeepLIFT shows more consistent performance across different perturbation scales, so it may provide more stable explanations under varying conditions. This relative stability of DeepLIFT performance across different perturbations suggests its explanations are robust, particularly in terms of handling categorical feature perturbations. However, under lower perturbation levels SHAP offers more precise explanations, as we can see in their lower infidelity scores under those conditions.

2.4. Evaluate the computational efficiency of the different methods

(i - ii) Preproccessed the Dry Bean Dataset in notebook - table somewhere on next page.

(iii) From the Titanic data runtimes, DeepLift appears to be the most computationally efficient method, followed closely by SVS, while SHAP is the least efficient by a significant margin. This suggests that for rapid analysis or when working with datasets where computational resources are limited, DeepLift or SVS might be preferable. However, it's important to consider the trade-offs in terms of the detail and interpretability of the attribution scores provided by each method. As we have previously discussed. Unironically this computational inefficiency ultimately proved prohibitive from the bean dataset. Despite working on debugging this component of the coursework for genuinely a good 10-15 hours, by the

Table 1. Mean Infidelity for Different Noise Scales and Categorical Resample Probabilities

Noise Scale	Cat Resample Proba	Mean Infidelity SHAP	Mean Infidelity SVS	Mean Infidelity DL
0.1	0.2	0.113456	0.120675	0.164945
0.5	0.2	0.210063	0.262067	0.248968
0.9	0.2	0.277250	0.290630	0.260649
0.2	0.1	0.181616	0.207093	0.215419
0.2	0.5	0.181616	0.207093	0.215419
0.2	0.9	0.181616	0.207093	0.215419

Metric	Value
Loss	91.6895
Accuracy	90.50%
Precision	92.02%
Recall	91.07%
F1 Score	91.19%

Table 2. Key performance metrics for the additional neural model after training.

submission deadline I have still had to hand in the .ipynb without these values calculated.

Method	Titanic Dataset Runtime (seconds)
SHAP	33.990
SVS	0.098
DeepLift	0.020

Table 3. Runtimes required to produce the attribution scores for different methods on the Titanic dataset.

3. Counterfactual Explanations

3.1. Consider the same Titanic dataset and the corresponding PyTorch model used for the previous task. Design a suitable distance metric to use for measuring the proximity between points. Detailed instructions are in the Jupyter notebook.

(i) Without normalisation or weighting, the standard L1 distance treats all features equally, but this can be misleading because features have different scales. For instance, age and fare values can vary significantly more than binary variables like sex (encoded as 0 or 1). Therefore, the variables with a larger-scale might dominate distance calculations, skewing the results. As you would then assume, if we use the normalised L1 distance by the range of each feature (max - min), the scale issue will be somewhat addressed, making the contribution of each feature to the distance more equitable. However, this approach still wouldn't account for intrinsic differences in the importance or relevance of each feature to the explanation / whatever task at hand.

(ii) Since we have a mixture of categorical and continuous variables, to ensure each feature is treated equally in the preprocessed dataset, we would need to normalise the continuous variables (i.e. all features other than sex and the survival label), and use one-hot encoding or a similar encoding method to convert categorical variables a numerical format. The normalisation step ensures that no single feature dominates due to scale differences, the encoding ensures that each category has equal weight in the distance calculation and then we would give equal weighting to each feature to ensures no feature is considered more important than another in the distance calculation.

(iii) Implemented

3.2. Implemented a Nearest-Neighbour Counterfactual Explainer (NNCE) & the method by Wachter et al. [2]

3.3. Generate counterfactual explanations for 20 randomly selected test instances of the Titanic dataset

Method	Validity	Cost	Plausibility
NNCE	0.81 ± 0.097	1.076 ± 0.393	0.122 ± 0.019
WAC	1.0 ± 0.0	0.827 ± 0.319	0.296 ± 0.083

Table 4. Comparison of counterfactuals performance between NNCE and WAC methods.

(i-ii)

(iii) NNCE achieves a validity score of 0.81 ± 0.097, indicating that about 81

WAC achieves perfect validity (1.0 ± 0.0), as expected from its gradient-based optimization that directly aims to change the model's prediction by fine-tuning input features. With an average cost of 0.827 ± 0.319, WAC's gradient-based method appears to identify counterfactuals closer to the original instances, showcasing its ability to adjust features to minimize distance more precisely. WAC significantly outperforms NNCE in plausibility (0.296 ± 0.083), suggesting its optimization process likely respects the underlying data distribution and real-world constraints better, leading to more realistic changes.

Comparing these two approaches highlights the advantages and trade-offs that must be considered between different counterfactual generation methods.

References

[1] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences, 2019. 1

[2] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr, 2018. 3